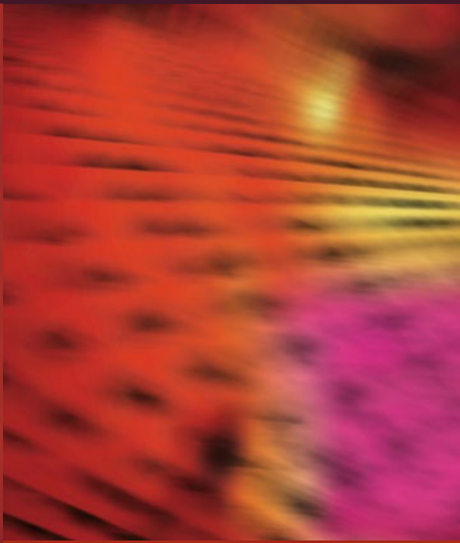
THE FRONTIERS COLLECTION



Shyam Wuppuluri
Francisco Antonio Doria

THE MAP AND THE TERRITORY

Exploring the Foundations of Science, Thought and Reality

Foreword by Sir Roger Penrose
and Afterword by Dagfinn Føllesdal

EXTRAS ONLINE

Springer

THE FRONTIERS COLLECTION

THE FRONTIERS COLLECTION

The books in this collection are devoted to challenging and open problems at the forefront of modern science, including related philosophical debates. In contrast to typical research monographs, however, they strive to present their topics in a manner accessible also to scientifically literate non-specialists wishing to gain insight into the deeper implications and fascinating questions involved. Taken as a whole, the series reflects the need for a fundamental and interdisciplinary approach to modern science. Furthermore, it is intended to encourage active scientists in all areas to ponder over important and perhaps controversial issues beyond their own speciality. Extending from quantum physics and relativity to entropy, consciousness and complex systems—the Frontiers Collection will inspire readers to push back the frontiers of their own knowledge.

More information about this series at http://www.springer.com/series/5342

For a full list of published titles, please see back of book or springer.com/series/5342

Shyam Wuppuluri · Francisco Antonio Doria
Editors

The Map and the Territory

Exploring the Foundations of Science, Thought and Reality

Foreword by Sir Roger Penrose and
Afterword by Dagfinn Føllesdal

Springer

Editors

Shyam Wuppuluri
R. N. Podar School
Mumbai
India

Francisco Antonio Doria
Advanced Studies Research Group
Universidade Federal do Rio de Janeiro
Rio de Janeiro
Brazil

Additional material to this book can be downloaded from http://extras.springer.com.

ISSN 1612-3018 ISSN 2197-6619 (electronic)
THE FRONTIERS COLLECTION
ISBN 978-3-319-72477-5 ISBN 978-3-319-72478-2 (eBook)
https://doi.org/10.1007/978-3-319-72478-2

Library of Congress Control Number: 2017961750

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

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

Foreword

This volume provides a wonderful collection of essays by very distinguished scientists, mathematicians and philosophers. We find here their numerous and very different deep and individual conceptions of the relationship between the actual world we live in and how we perceive and understand that world. The variety presented here is stunning in its breadth and diversity of outlook.

In accordance with such diversity, it is difficult for me to comment without interposing something of my own personal viewpoints which have come about from a lifetime's study of mathematics and the foundations of physical theory. It is indeed hard for me not to be hugely influenced by both the extraordinary subtlety and power of the mathematical structures that have been developed over many centuries, where not only is the precision inherent in these mathematical edifices breathtaking when the theory works well, but also in such theories there revealed a supreme beauty in the coherence and frequent unexpected applicability that one finds in these physical laws when they are at their most successful.

We now have, for example, clocks that are so precise that had they been started at the time of the Big Bang they would still remain true to within a second. But what do we mean by true? This refers to an internal consistency between theory and observational facts whenever it becomes possible to bring the two together. Much of this precision comes about from those two great revolutions of twentieth-century physics, namely general relativity and quantum mechanics, both of which theoretical constructions are confirmed in observation to an extraordinary degree. The clocks just referred to, for example, depend upon a deep relevance of the two most important formulae of twentieth-century physics, namely Albert Einstein's central formula of relativity theory $E = mc^2$ and Max Planck's foundation stone of quantum mechanics $E = h\upsilon$. The first states the equivalence of energy with mass and the second, the equivalence of energy with frequency, and put together we get the equivalence of mass with frequency, whence stable massive particles must themselves possess oscillatory frequencies of incredible precision. Yet, these two great theories do not sit comfortably together. Indeed, in a deep sense, Einstein's general relativity is technically inconsistent with the foundational tenets of quantum theory.

Should one take the view that they are just our best way of understanding the world in its largest scales and in its smaller scales, where there might be no reason to expect that some overarching and internally consistent mathematical scheme might be waiting in the wings, someday to be discovered to encompass both as limiting approximations? My own view is certainly that there must be something of this nature, and that ultimately we may be fortunate enough to come across such an overall mathematical framework which will override our current disparate attempts to account for the actions of the knowable universe—in principle at least.

As our current theories stand, there is a particular issue about quantum mechanics that is not shared by general relativity. In the latter, there appears to be a well-established ontology whereby the mathematical models that we try to construct consistently with the equations of the theory can present us with allowable pictures of what we may be able to refer to as candidates for inhabitants of 'the real world'. In quantum mechanics, what we are presented with is something very different where there is little agreement between different proponents of what the theory might mean. Is the wavefunction real? If so, does it satisfy the unitary equation of Schrödinger? If so, how does this address the issue of 'Schrödinger's cat' whose 'real' existence would be described as being in a superposition of death and life? Or is the very wavefunction a mere mental construction providing us with just a way of calculating probabilities of something which then becomes real—or what? In my view, there are strong reasons for taking the standpoint that there must be some form of reality in the wavefunction but that this does not always satisfy Schrödinger's actual equation, and something different then comes about in 'reality' from time to time? Perhaps, this 'really' happens only as soon the space–time curvatures of Einstein's gravitational theory begin to impinge on the structure of quantum mechanics. Might such a scheme be needed before an overall ontological consistency can be provided for quantum theory?

What about determinism? Current quantum mechanics, in the way that it is used, is not a deterministic scheme, and probabilistic behaviour is taken to be an essential feature of its workings. Some would contend that such indeterminism is here to stay, whereas others argue that there must be underlying 'hidden variables' which may someday restore a fully deterministic underlying ontology.

Personally, I do not insist on taking a stand on this issue, but I do not think it likely that pure randomness can be the answer. I feel that there must be something more subtle underlying it all. What view we take about the ontology of the world seems to be intimately tied up with what equations, or other mathematical constrictions our theories define for us. It is my view that many of the puzzles that people have in relating the formalism of quantum mechanics to the behaviour of the physical world come about from a committed belief in the universal correctness of the quantum formalism as it stands today. To me, there is a profound question about this widely held belief among established physicists that one should not monkey with this formalism and take what it says to be an unquestioned truth. It is this that, in my view, leads to many of the difficulties that people have with providing a fully consistent ontology for quantum mechanics.

In this volume, you will find many alternative positions on this and many other issues that arise in relation to the whole concept of ‘ontology’ and what it may actually mean. Moreover, the tests that are applied to physical theories in order to see whether they are consistent with nature are often extraordinarily refined. Much of our experience of the world itself is in circumstances where behaviour can be consistent without expectations mainly because we have seen such things frequently before. We are unlikely to test the behaviour of a spilled glass of orange juice by delving into the detailed equations that physics has presented us with. Instead, we tend to have a faith that if such a situation were studied in detail using all the equations that we believe to be relevant, then there would be consistency with what we observe. Is this faith justified? Probably in the case of a glass of orange juice, this is so. But how about situations when it comes to the behaviour of biological systems and their growth patterns? In the case of animals, and how they might behave in the face of different external circumstances, do we fully trust our equations? How about the behaviour of a human brain? Do we have the same faith that those laws that serve us so well with inanimate entities will serve equally in the case of human behaviour? Might there be something different when it comes to consciously controlled actions? Might we need to extend our physical pictures to something beyond the kind of mathematical theory that has worked so well for us so far?

Clearly, there are many questions about what reality might be and whether or not our physical theories are close to providing a universal picture of how the world operates. These theories are—or at least have been so far—mathematical theories with reasonably sound underpinnings of consistency, despite some puzzling issues of their ontological status. If the mathematics ever comes to fit the behaviour of the world in a way which appears to be absolutely precise, would we choose to identify actual reality with well-prescribed terms in this mathematical formalism? Could we live with a picture where we and all our surroundings are simply parts of the Platonic world of purely mathematical abstractions? A view is not uncommonly put forward that the world is, in some sense, simply a computational ‘simulation’, like the running of a computer program. This is a viewpoint that I find hard to relate to. If the operation of our universe is merely a simulation, then what is the ‘thing’ that it actually simulates? Our current technology, which depends so strongly upon the actions of computers, seems to render such a picture plausible. But that is not my own picture of how our universe can operate. Mathematics, yes, it must deeply underlie the workings of the world, but that does not imply that the world operates in an entirely computational manner. There is far more to mathematics than that.

Oxford
August 2017

Roger Penrose

This page is intentionally left blank[1]

[1] But it is not blank! It has some text! The map is not the territory.

Preface

After we came out of the church, we stood talking for some time together of Bishop Berkeley's ingenious sophistry to prove the nonexistence of matter, and that everything in the universe is merely ideal. I observed that though we are satisfied his doctrine is not true, it is impossible to refute it. I never shall forget the alacrity with which Johnson answered, striking his foot with mighty force against a large stone, till he rebounded from it—I refute it thus.

—Life of Samuel Johnson, James Boswell

The irony of life is that it is lived forward but understood backward

—Soren Kierkegaard

Ob nicht natur zuletzt doch ergründe?

—Wolfgang Von Goethe

In this volume, some of the world's leading thinkers come together to expound upon the topic of the map/territory distinction in the foundations of science, the process of thought and even reality itself, whatever that may be. Science longs for simplicity. As Einstein once remarked, 'everything should be made as simple as possible, but not simpler'. One of the chief goals of science is to find a minimalistic set of equations that can describe all the happenings in the universe, so short that a person sitting at a cafe, sipping *caffè macchiato*, *in angello cum libello*, can scribble it down on the back of his coffee bill. These bite-sized equations hold within themselves a myriad of complex interrelationships between various areas of knowledge and therefore also with the real world. Knowledge and ignorance, as ever, share a *ménage-à-trois* relationship with thought. The more we know, the more we realise that there is to know, and the more we realise how much we do not know.

To think is to represent, whatever the nature of such representation. There is undoubtedly a deep connection between the name and that which is named, phonetics and script, a picture of a person and the person it shows, thought and the object of thought, a map of Vienna and Vienna itself, a finger pointing at the moon and the moon, etc. We all grew up reading those classic stories of romance, in which a troubled princess trying to escape from the kingdom stares endlessly at the picture of an imaginary prince, and lo and behold, the prince materialises from the picture and saves her! Too good for a fiction plot and too bad for science.

Representations are handy and tempting, and they come so naturally to us that we quite often end up committing the category error of over-marrying the representation with what is represented, so much so that the distinction between the former and the latter is lost. This is a form of intellectual *harakiri* that prevents us from understanding the subject. ‘If all we have is a hammer, everything looks like a nail’, as the saying goes. Similarly, if all we have is a map, everything looks like a territory. Sometimes, there may be no territory corresponding to our map, in which case our map is just a convenient representational tool, like a mnemonic, but a plethora of paradoxes and inconsistencies surface when we consider the most successful abstractions (maps) to be a part (or an attribute) of the real world.

Therefore, it is imperative for a student or a researcher of science to differentiate between the computational tool and what it computes, to distinguish the map from the territory it represents. ‘The map is not the territory’, remarked Alfred Korzybski. There are multitudes of maps that we use to ‘represent’ the reality out there. They differ both in form and substance. The scientist in this sense resembles a cartographer. Only a cartographer knows how hard is it to represent a map of the earth on a sheet of paper. Every step towards perfecting the map involves a sacrifice—adding some feature to the map that does not have any intuitive or direct correspondence with the territory or ignoring many complexities of the territory.

For instance, consider an apple. One can apply a name and a price tag to it and study the economics and geography of the commodity. Or an apple may just be a collection of sensory perceptions like taste, colour, touch, etc., that lead us to the basic idea of an apple. Or one can describe it as a biological system and apply genetics and the other formalisms of biology to study it. Or model it as a point-like particle and apply Newtonian mechanics to it. Or see it as a point in 4D space–time that instantiates an event and apply the principles of relativity to it. Or see it as a vast collection of sub-atomic particles obeying the laws of quantum mechanics, quantum field theory, string theory and so on.

Which one of these is the **apple that’s out there**? Or is there an apple out there, apart from these maps (notions)? Here, we are concerned with the epistemology/ontology distinction. Can we transform one map into the other? Or is there a global map that can simulate every other map under some constraint? Do all of these maps co-exist? In the same vein, to what extent are our scientific maps accurate in portraying their corresponding territory? What about the things like numbers, sets, classes and functions? What about space, time, fields and operators? Are they a part of our map (computational/visualizational tools) or are they part of the territory (reality)? If two maps cannot be integrated, is this a limitation of our scientific cartography or is it the nature of the underlying territory itself that prevents us from such an attempt? Foundational questions of this sort play an important role in science, especially in modern physics (grand unified theories). It is safer to let the gaps remain as gaps while we let our maps remain as maps, rather than giving in to the seemingly seductive approach of trading in our understanding and intermingling maps with territory to fill in the conceptual gaps—however, much this may comfort us and appeal to our tastes!

The eminent philosopher W. V. O. Quine quotes Otto Neurath in his magnum opus ‘Word and object’, *‘Neurath has likened science to a boat which, if we are to rebuild it, we must rebuild plank by plank while staying afloat in it. The philosopher and the scientist are in the same boat…’* We can further imagine this particular vessel to be the ‘*Ship of Theseus*’, which at every point has to maintain consistency with the established truths and in some ‘sense’ preserve its structure. But is this really the case? Modern science, with its numerous interconnections between so many different fields, is reminiscent of the interconnections between the neurons in our brain. There are also meta-maps—so to speak—which serve as mortar between the different maps. It is almost impossible to speak of any subject or map in isolation, or establish a hierarchy of fields to show what arises from what. Everything co-exists. It is the whole that gives meaning to its parts, and the parts that give meaning to the whole.

Beneath all the richness of these maps is our consciousness, which colours them and in turn gets coloured by them. Our thoughts are so densely connected with each other that it is impossible for us to step twice into the same metaphorical river of thoughts. As Sartre says, in every attempt to enter consciousness, we are seized by a whirlwind and thrown back outside. We then turn to language, our only hope, which also plays an important role in the mapping process. All the categorization our cognition exercises bears an intricate relationship with language. For instance, how is it that the patterns of tilings we see become a tiling of patterns? Consider also the statement: ‘There are three red balls in the urn’. Is it that the property of ballness is substantiated thrice? Or is it the property of threeness that is substantiated by a set of three red balls? Or is it that the property of redness is thrice substantiated by ballness? Or conversely? Which attribute is a part of reality and which one is not? Is this a situation where our language (*façon de parler*), which is playing Wittgensteinian games here, would put an end to these a priori/a posteriori disputes?

Above all, who are Homo sapiens but a bunch of evolved apes, selected by the Darwinian selection process and nurtured by nature over thousands of generations? Evolution has definitely contributed to our understanding of the world, by giving us brains and language, in a direct or indirect manner. How far does nature qua evolution control the very modalities that we use to picture it? For instance, we cannot see the third dimension in a straightforward manner, in the way we see two dimensions. Neither can we fly in the air like birds. We cannot drink and talk at the same time. Neither can our skin harness light energy from the sun, as plants do, to provide fuel for our everyday lives. Nature has blessed each of its species with their own modalities, allowing them to establish their own relationship with the reality they perceive and interact with. While the above limitations are physical in nature, we assume our brain is free to ponder anything, and that no one can imprison our imagination. To some extent, we have overcome these physical limitations and taken several steps ahead with a sense of victory, seeing the third and even fourth dimensions using technology, and even getting hold of infinite dimensions with the help of induction and advanced mathematics. We have also discovered that we do not always need wings to fly, just as we do not necessarily need legs to walk. This notion of abstraction, abstracting walking from legs and flight from wings, has given us cars and aeroplanes. But are there things we would have thought otherwise had we been granted different sensory systems? We definitely do not perceive the world the way (say) a goldfish perceives it. Are there truths that a goldfish alone knows and that perhaps we can never know? As Wittgenstein once said, 'If a lion could speak, we wouldn't be able to understand it'. So evolution definitely fences in the very way we think and reveals to us only those aspects beneficial for our survival. But the question is, to what extent? Is there a way out of the metaphorical Platonic cave erected around us by the nature?

Amidst this pessimistic and chaotic mass of questions, is there any chance of finding clarity and order?

It is hoped that the articles in this collection will be of some help here, authored by intellectual giants who can provide us with deep insights into the nature of maps and territories. When this volume was planned, it seemed natural to organise the articles into sections to facilitate understanding, and in the hope that a global meaning will emerge from these contextual viewpoints when we finally come to join the dots. We have thus divided the volume according to field, namely philosophy, physics, mathematics/computer science, biology/cognitive science, and a miscellaneous section which includes literature and geography. Every article in each field deals with the underlying issue of the map/territory distinction and addresses the problem from its own point of view, in the context of that particular field. The authors have invested considerable time and energy to make the articles accessible both to researchers and to those with only a rudimentary knowledge of the subject.

Is the map the territory? Are we trying to answer a question or question the answer? Join us on this journey if you would like new perspectives on questions like these.

Juhu, Mumbai, India Shyam Wuppuluri

Acknowledgements

> *We are but dwarfs mounted on the shoulders of giants, so that we can see more and further than they; yet not by virtue of the keenness of our eyesight, nor through the tallness of our stature, but because we are raised and borne aloft upon that giant mass.*
>
> —Bernard of Chartres

The giants who have contributed to this volume have helped us to see further than we could ever see. I would like to return a part of their favour by thanking them for going out of their way, investing their valuable time and making this piece of work possible. Assembling a volume with the length and breadth of this is no cakewalk. Let me begin thanking everyone who have helped me do so. I shall do this in a chronological way.

While the overall structure and contents were in place, with 40 authors on board but still moored in the harbour, I sought a co-editor for feedback. I am eternally grateful to my co-editor, an erudite scholar, Prof. Francisco Antonio Doria and his lifelong collaborator and genius Prof. Newton Da Costa, for the excitement, support, suggestions, pertinent feedback and kindness they showered me with, at various stages, since the volume's inception. Such debts as this cannot be cleared just by thanking. I hope their friendship and collaboration will continue to inspire many others, while they continue to spread goodwill to those around them. Dr. Angela Lahee, editor and founder of the Frontiers Collection, has been exceptionally kind right through the production process, and played the role of a stake, maintaining the structure of this intellectual tent against the stormy desert night. Her timely editorial support, well-aimed criticism, support, finesse, and immense expertise are warmly acknowledged herewith.

I would next like to thank Sir Roger Penrose for his generosity in contributing a foreword to the volume. He definitely needs no introduction and I am sure his huge contribution to mathematics and various sciences will continue to inspire generations to come. I also would like to thank Mrs. Petrona Winton, PA and ATI administrator at Oxford, for her kind cooperation and help. When I first approached Prof. Dagfinn Follesdal, student of W. V. O. Quine, he not only responded positively, but also went to the trouble of going through the other articles and leaving us with a beautiful afterword. I extend my most heartfelt thanks to him for this warm gesture.

I feel a little guilty that I cannot mention each author here, but I would like to thank them all wholeheartedly for their efforts, and especially those who generously and patiently incorporated feedback and customised their article, despite their pressing academic commitments. I thank Prof. Gregory J. Chaitin and his wife Prof. Virginia Chaitin for their kindest support, their contribution and more than anything for the unconditional love they have offered. I hope that the new phase of their life brings them great joy! I would like to then thank Prof. Ian Stewart, Prof. Barry Dainton for their efforts and support. I would also like to thank Prof. John R. Searle and Prof. Tyler Burge for their time, patience, interaction and support. I also thank Prof. Joseph Kouneiher for his contribution. I acknowledge his humility, intellect and his kindness herewith.

I must not forget the support and encouragement I received from Prof. Barry Mazur and I am exceedingly grateful for it. I would also like to thank Prof. B. J. Rao for his kindness and support, along with Prof. Hans Herlof Grelland, Prof. Donald Janelle, and Prof. Robert Tally for their willingness to work on another volume of mine, in another time and another space. Lastly, I am immensely thankful to Mr. Sooryadeepth Jayakrishnan and his team at Springer for bringing this book to fruition. His timely support during typesetting phase, seamless coordination and expertise is acknowledged herewith.

I would also like to thank those who, though not directly related to this project, have helped me at various stages during my life. I reserve my utmost gratitude for Prof. Sujatha Ramdorai, who has been a pillar of support at various stages of my life. The spark she ignited in me burns forever, regardless of the laws of thermodynamics! I thank and admire Mr. M. G. Subramanian for his Zen-themed advises and for the support he offered at various times and also Mrs. Avnita Bir and other noble colleagues at my office for their support.

I would like to thank Prof. Douglas Hofstadter and Prof. Daniel C. Dennett, who inspired me and so many others, endlessly through their books and their personality. I am also thankful to Prof. John Stachel for the continued kindness and encouragement that he has provided me with. Finally, I would like to thank all those who have shown kindness towards me, as well as those who did not, for each in their own way has contributed to my betterment. I, like Lazarus, have managed to survive so far on the breadcrumbs of love and kindness that have fallen from the table of rich—rich in heart. The boon or the curse of existence, I truly believe, lies in such love and kindness. This reminds me of something I wrote some time ago:

"Where does the fault lie? In our Stars?" asked the puzzled boy of his grandfather, a retired physicist. "Or is it in space and time itself?" he continued. "Why is it that it takes forever to see that someone is good? And only an instant to prove that someone is bad! Is time the culprit here? Is it space that prevents two people being together emotionally or physically?"

"No!" his grandfather exclaimed, patting the boy on the shoulder. "The fault is neither in our stars, nor in space and time. It's in our love, which weaves the very fabric of existence. Love, which is the alpha and the Omega. Love, which is the journey and the destination."

Juhu, Mumbai Shyam Wuppuluri

Contents

Part I Philosophy

Part II Theoretical Physics

Part III Mathematics/Computer Science

Part IV Biology/Cognitive Science

Part V MISC

About the Editors

Shyam Wuppuluri is working as a Research Associate in R. N. Podar institute. As a computer science graduate, he has a long-standing interest in various areas of philosophy, theoretical physics, mathematics and cognitive science. Prior to this, he edited the volume ‘Space, time and limits of human understanding, foreword by John Stachel, afterword by Noam Chomsky’, which was published in Springer—the frontiers collection, 2016.
Email: shyam.wuppuluri@gmail.com

Francisco Antonio Doria is a Brazilian physicist. He is a Professor Emeritus at the Federal University of Rio de Janeiro, where he currently teaches on the foundations of economic theory at the graduate School of Engineering (UFRJ COPPE). He has a B.Sc. in chemical engineering and a Ph.D. in mathematical physics (under the guidance of Leopoldo Nachbin). He has made contributions to the gauge field copy problem in quantum field theory and proved with Newton da Costa several incompleteness theorems in mathematics, physics and mathematical economics, including the undecidability of chaos theory. He is a member of the Brazilian Academy of Philosophy, a Senior Fulbright Scholar at Stanford University, 1989–1990, and a visiting researcher at the mathematics department, University of Rochester, 1979–1981. He thinks of himself as a philosopher and literary scholar with a humanist education, and has had as students Marcelo Gleiser and José Acacio de Barros, among other noted researchers. He likes to trace his interdisciplinary interests to a seventeenth-century relative, the noted Portuguese writer Father Antonio Vieira (1608–1697).
Email: fadoria63@gmail.com

Part I
Philosophy

Chapter 1
Maps and Territories in Scientific Investigation

Evandro Agazzi

Introduction

Scientific investigation has always been understood, at least in the Western tradition, as an effort to acquire *knowledge* about *reality* and, originally, this reality was understood as the *totality* of what exists. Already in the 'classical' Greek culture, however, a certain partition of such a totality was recognized: Plato and Aristotle, for example, admitted the existence of certain fundamental *sciences* (such as mathematics, physics and theology) characterized by the special nature of their subject matter. This was the first appearance of 'territories' in the broad domain of human knowledge and, at the same time, the explicit recognition that the 'ordinary' or commonsensical cognitive approach to reality (that considers it 'globally' or 'as a whole') had to be superseded by a more reliable and solid knowledge. Ordinary knowledge is almost entirely constituted by the content of sense perceptions that produce a large display of *opinions*, many of which turn out to be erroneous. The earliest philosophers have spent lot of reflection on this issue, and Parmenides had separated and opposed the realm of opinion (*doxa*) and the realm of truth (*aletheia*). Nevertheless, one must admit that there are also true opinions. Therefore, a different partition had to be proposed in which the requirement of truth were reinforced by a warrant of absolute solidity that opinion fails to offer. This new kind of knowledge was called *episteme* and was translated by *science* in modern languages; it was characterized by the fact that in it truth is accompanied by "a discourse providing the reasons" for that truth. If, using a modern terminology, we call "justification" such a discourse, and use the term "belief" to express the notion of opinion, we find already in Plato and Aristotle the characterization of science as "true belief supported by a justification" that is rather common also in contemporary epistemology.

E. Agazzi (✉)
Interdisciplinary Center for Bioethics, Panamerican University of Mexico City,
Mexico City, Mexico
e-mail: evandro.agazzi@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_1

Of course, the great problem consists now in specifying in what such a justification should consist, and already Plato and Aristotle had developed the doctrine that a ‘scientific’ knowledge, or *science* in a proper sense, must be a *deductive* discourse, in which truth is all-pervasive because it derives from the immediate truth of ‘first principles’ thanks to a rigorous logical deduction. How such first principles (endowed with universal, necessary and immediate truth) could be established was investigated with great acumen by those two authors and we are not going to recall their reasoning here. What is of interest for us here is that, as a consequence of those reflections, they recognized that the very demanding level of science could not be attained in whatever cognitive enterprise, but only regarding a few privileged domains, and according to different degrees of perfection. According to a classification already established by Aristotle (and which remained stable for many centuries within Western philosophy), a first distinction is based on the different *aims* in view of which a certain investigation is made: we must distinguish the *theoretical* (or ‘speculative’) sciences in which only the disinterested search of truth is pursued, from the *practical* sciences, that investigate the human *praxis,* that is, what is the best way of realizing a ‘good life’ as an individual and as a citizen (they are ethics and politics), and from the *poietical* sciences, that concern the *poiesis*, that is, the ‘production’ of concrete objects or results. In Greek they were called *technai,* were translated in Latin by *artes* and in modern languages by ‘arts’ (not, however, in the sense of ‘fine arts’ or activities regarding the creation of beauty, but in the more traditional sense of ‘arts and crafts’ covering a wide display of professional activities). This was, in a way, already a partition of the realm of science in ‘territories’, but additional partitions were realized within each of such territories. We have already mentioned the subdivision of the theoretical sciences into mathematics, physics and theology, and that of the practical science in ethics and politics. The great territory of the poietical sciences was subdivided in a rich variety of subterritories, such as architecture, medicine, navigation, military strategy, painting, sculpture, and a lot of minor crafts.

Modern ‘Scientific Revolution’

A new perspective appeared with the *scientific revolution* occurred in the Renaissance and inaugurated by Galileo. The first impression could be that the new science was essentially the expression of the decision to ‘restrict’ the investigation to the domain of physical entities, that is, to the territory of the traditional physics. If this view were correct, however, one could not see in what sense this natural science should be considered *new*. The novelty explicitly advocated by Galileo consisted in the proposal of a new *cognitive approach* that amounted to the creation of a new ‘territory’ not in the domain of the theoretical sciences, but rather in epistemology itself. To put it explicitly, he still considered the aim of theoretical investigation to be that of acquiring *truth*, but wanted to introduce, beside the traditional distinction of opinion and Absolute science, a third kind of knowledge, that we shall call Relative science. This was not a subterritory of the classical science (Absolute science) but actually a new territory within epistemology, whose application concretely concerned (at that moment) the study of physical bodies, but whose characteristics were independent of this particular application and were potentially open to be used in the study of many other subject matters.

The fundamental characteristics of this new model are summarized by Galileo in a passage of his third letter to Marcus Welser on the sunspots:

> In our speculating we either seek to penetrate the true and internal essence of natural substances, or content ourselves with a knowledge of some of their affections. Attempting the essence I hold to be as impossible an undertaking with regard to closest elemental substances as with more remote celestial things…. But if what we wish to fix in our minds is the apprehension of some affections of things, then it seems to me that we need not despair of our ability to acquire this respecting distant bodies just as well as those close at hand – and perhaps in some cases even more precisely in the former case than in the latter.[1]

In these lines Galileo makes a sharp distinction between the internal 'essence' and the 'affections' of the natural entities, and in addition declares that we can hope to attain some knowledge of such entities only if we confine our attention to their affections. We need simply to remember that the effort of knowing the essence had been considered the specific attitude of *philosophy* at least since Socrates. Therefore, we can conclude that Galileo's proposal was, at least in part, that of abandoning the strictly *philosophical* viewpoint in investigating nature, a viewpoint that had been elaborated in the classical ideal of science as a deductive discourse deriving truth of factual statements from the universal and necessary 'absolute' truth of the first principles. This methodological prescription of restricting the investigation to the study of certain properties (or 'accidents', according to the scholastic terminology of that time) was only the first step in the 'relativization' of scientific inquiry advocated by Galileo. A second not less decisive relativization consisted in the fact that not whatever accidents (or "affections") of the physical bodies were considered susceptible of such a new investigation but only those which could be considered as "real accidents", objectively intrinsic to a physical body and not depending on the subjective perception of the single person, as is explained in a celebrated passage of Galileo's *Assayer*:

> Now I say that whenever I conceive any material or corporeal substance, I immediately feel the need to think of it as bounded, and as having this or that shape: as being large or small in relation to other things, and in some specific place at any given time; as being in motion or at rest; as touching or not touching some other body; and as being one in number, or few, or many. From these conditions I cannot separate such a substance by any stretch of my imagination. But that it must be white or red, bitter or sweet, noisy or silent, and of sweet or foul odour, my mind does not feel compelled to bring in as necessary accompaniments. Without the senses as our guides, reason and imagination unaided would probably never arrive at qualities like these. Hence I think that tastes, odours, colours, and so on are no more than mere names so far as the object in which we place them is concerned, and that they reside only in the sensitive body so that, once the animal is removed, they are all

[1] Galileo, *Opere* V, pp. 187–188; translated in Drake (1957), pp. 123–124.

> removed and annihilated as well. But since we have imposed upon them special names, distinct from those of the other primitive and real accidents, we wish to believe that they really exist as actually different from those.[2]

It is already clear from this passage that the "real accidents" of the physical bodies are those that can be represented through geometry and arithmetic or, in short, the *mathematizable* properties of the material bodies, and this is confirmed in another famous passage of the *Assayer* where Galileo explicitly declares mathematics as the only proper language for expressing the objective features of the universe:

> Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures without which it is humanly impossible to understand a single word of it.[3]

The two restrictions that we have considered thus far regard what we could call the 'ontology' of the new science, that is, the investigation of certain selected accidents instead of the essence. This shift, however, entailed a serious consequence: according to traditional ontology, the essence of a certain reality was supposed to constitute the fountainhead from which descends the concrete behaviour of that entity in the different concrete circumstances; therefore, when one believes that it is possible to grasp the essence of certain objects is also confident that an 'absolute' knowledge of these objects, endowed with necessity, can be attained. But if the knowledge of the essence is discarded as an impossible enterprise, what remains is only a *conjectural* form of knowledge, which is relative to the force of the cognitive tools applied. This is precisely what Galileo is conscious of, though he is confident that such a knowledge is solid. The new tool that he has invented is the *experimental method*: the careful mathematical description of a certain regular process in which only certain selected measurable magnitudes are considered induces the scientist to formulate a general mathematical hypothesis. This hypothesis is then tested, by artificially producing a situation in which the intended 'accidents' of a concrete material body are carefully measured, and their value is put in the hypothetical mathematical expression. They are the 'initial conditions' of the process whose final result should coincide with the result of the calculation of the mathematical hypothesis. If such a confirmation is ascertained in a reasonable number of repetitions of the experiment, the mathematical hypothesis is considered well established and becomes what is usually called a *natural law*. Galileo has applied this method in several investigations and has also pointed out its conjectural nature in short statements like, for instance, the following passage in a letter to G. B. Baliani:

[2] Galileo (1623), *Opere* VI, pp. 347–348; translated in Drake (1957), p. 274.

[3] Galileo (1623), *Opere* VI, p. 232; translated in Drake (1957), pp. 237–238.

> I argue *ex suppositione* about motion, so that even though the consequences should not correspond to the events of the natural motion of falling heavy bodies, it would little matter to me, just as it derogates nothing from the demonstrations of Archimedes that no moveable is found in nature that moves along spiral lines. But in this I have been, as I will say, lucky: for the motion of heavy bodies and its events correspond punctually to the events demonstrated by me from the motion I defined.[4]

The Proliferation of the Sciences

The novelty introduced with the modern natural science was a significant change in the determination of the 'territories' of the different sciences. The ancient subdivision, as we have seen, relied on the specific aim of a certain science (producing the partition into theoretical, practical and poetical sciences), or on some general characteristics of the entities (or 'substances') investigated (material, immaterial, abstract). This kind of partition did not disappear in the sciences of modernity and is clearly visible in those sciences that we can call 'descriptive', like Astronomy (that studies the celestial bodies), Zoology (that studies animals), Botany (that studies plants). The intellectual work in these sciences is not negligible, but mainly consists in the elaboration of classifications, taxonomical ordering, comparative study of similarities and differences, and no application of the experimental method occurs.

In modern sciences, however, the focus is shifted from 'substances' to 'accidents' (to use the traditional ontological vocabulary), that is, to properties and relations (we shall call them *attributes*) and a science is specifically characterized by the fact of selecting few attributes and then studying reality *from the point of view* of these attributes, independently of the fact that they are present in a certain entity or in another kind of entities. For example, mechanics considers reality from the point of view of mass, displacement in space, duration in time (that define motion), and force as something producing the change of motion. Using these concepts one can study the motion of a falling stone, the orbits of certain planets, the oscillations of a pendulum, the velocity of a car, the force needed by a bird to fly in the air, and so on. Despite this great variety of substances that can be considered from the point of view of mechanics, there are also plenty of substances and processes that cannot be investigated from the point of view of this science, simply because they do not possess the attributes specifically constituting the 'point of view' of mechanics. For instance, a toothache has no mass, the change of the price of a commodity on the market has nothing to do with a displacement in space and time, the rate of increase of the bacteria population in a biological sample cannot be analyzed as a motion produced by a force, despite that in all these examples a suitable mathematical treatment can be applied. In a similar way we can consider attributes such as metabolism as being fundamental (possibly in simultaneous

[4] Galileo, *Opere*, XVIII, pp. 12–13; translated in Drake (1975), p. 156.

conjunction with other attributes) for defining the point of view of life sciences, or the capability of forming representations of things as the fundamental attribute of psyche.

What we have expressed through the colloquial phrase “point of view” can be expressed in a more rigorous way by speaking of *concepts,* which are the intellectual representation of the attributes. The concepts, in turn, are formulated through the specific ‘technical’ *predicates* of the language of a given science. Therefore, we can say that a well defined science considers reality by means of its specific concepts and speaks of reality through its specific disciplinary language. Using a more sophisticated terminology, we can say that every well defined science considers reality within its *conceptual space* and this conceptual space encompasses the whole *territory* of that science.

We have seen (in the example of mechanics given above) that many ‘things’ can be included in the territory of one single science, but also the reverse is true: one single thing can be included in the territory of several different sciences (for example, a dog can be studied from the point of view of mechanics, chemistry, biology, physiology, psychology, sociology, economics, and so on). This remark invites us to recognize the fruitfulness of the territory-approach, because it makes us aware that there are several ‘aspects’ of reality that can be captured by adopting suitable *perspectives,* and there is no cognitive advantage in trying to reduce the variety of such perspectives.

If we call *perspectivist* the approach that underscores this awareness, we can call *reductionist* the opposite approach, that considers pre-scientific and purely commonsensical the acceptance of that varieties of perspectives, whereas the genuine scientific view should consist in developing in depth the potentialities of a certain ‘fundamental’ science, such that the other domains could be ‘incorporated’ in the territory of the fundamental science.[5]

The reductionist attitude is usually the historical consequence of the outstanding performance of a particular science, both in terms of cognitive advancements and practical applications. This occurred in a spectacular way for mechanics which, in the course of the 19th century, seemed capable of understanding and explaining the phenomena perceived through the five senses and traditionally studied by separate sciences (like Acoustics, Optics, and the sciences of Electricity, Magnetism and Heath) that in such a way could be considered as ‘sub-territories’ of mechanics. A better understanding of that reduction led to the admission that it consisted in the production of ‘mechanical models’ of the phenomena investigated, e.g. in optics, electromagnetism and thermodynamics, but even this less ambitious view turned out to be untenable, even before the birth of quantum mechanics and relativity theory at the beginning of the 20th century. This ‘imperialism’ of mechanics, by the way, had manifested itself, in less radical and more generic forms, also regarding other sciences like, for example, biology: already Descartes had proposed a

[5] Two works in which perspectivism is presented with special enphasis are Agazzi (2014) and Dilworth (2008). A presentation and critical discussion of reductionism is found in Agazzi (1991).

mechanical interpretation of the human body,[6] and other authors had pushed this interpretation much further,[7] while in the 19th century a serious struggle opposed the 'vitalist' and the 'mechanist' scientists regarding the interpretation of the nature of life. In this 'dilatation', however, mechanics had lost its precise connotation as a science and had become first an umbrella covering the whole of physics and chemistry, and then a kind of general metaphysical framework for the understanding of all natural phenomena, coming back, in a way, to the level of common sense. This is attested, for example, by the fact that the word "mechanism" is currently used in many contexts that have actually nothing to do with mechanics proper (for instance, when natural selection is proposed as the 'mechanism' explaining biological evolution, or when the interaction of certain economic factors is seen as the 'mechanism' explaining the occurrence of a social crisis, or when the presence of certain personal 'propensities' or temperamental inclinations is proposed as the 'mechanism' capable to account for the behaviour of a person). It should be noted that the reduction to a 'fundamental' science does not entail that this science is fundamental in any precisely understood sense (like, e.g., when physics is taken as the fundamental science because '… after all, anything existing is made of protons, electrons and other elementary particles …'). It is sufficient that a science has attained a sufficient level of respectability so that its supporters do the job of 'reading' through the tools of their conceptual space the contents of other sciences. This has occurred, for instance, with the so-called 'sociological turn' in the philosophy of science inaugurated by Thomas Kuhn's book *The Structure of Scientific Revolutions*[8] in which the acceptance or rejection of scientific theories was not considered as dependent on logical consistency and empirical tests, but essentially on the acceptance of the scientific community at a given historical time. This view, accepted by Paul Feyerabend and developed up to its extreme consequences by other scholars of the same trend, has gradually come to the conclusion that science is simply a social product, and that what is often considered as a scientific fact or a scientific portrayal of how the world really is, is simply a social imagery, up to the point that even he difference between what is scientific and what is not is contingent upon the appreciation of society at a given historical time.[9] Just to add another example, we could mention the ingenious discourses through which certain authors have applied the concepts and theories of psychoanalysis to explain facts belonging to literature, politics and scientific research.

[6] This he did in his treatise *L'homme*, written in 1648 but published only after his death. It is reprinted in the second edition of the *Oeuvres* (vol. 14). An English translation done by Thomas Steele Hall was published by Prometheus Books in 2003.

[7] The most famous was certainly La Mettrie, whose *L'homme machine* appeared at Leyden in 1748 and is now available in a French-English edition. See La Mettrie (1748).

[8] See Kuhn (1962).

[9] Among the most representative authors of this trend let us mention Barnes (1977), Bloor (1976), Knorr Cetina (1981), Latour-Woolgar (1979).

From Territories to Mappings

We have said that the constitution of a certain territory of investigation amounts to the determination of the 'conceptual space' of a certain scientific discipline, that is, to the indication of the specific concepts that will be used to 'speak' in this discipline. It is clear, however, that concepts do not produce by themselves any speech or discourse: they must be put together, combined, related in what were traditionally called *judgments.* A judgment, as such, is a mental—or better an intellectual—entity that can be expressed in a *statement* of a given language, but also through different statements not only (obviously) of different languages, but also of the same language. This happens because the judgment is a *representation,* a global *model* or *image,* and several models or representations can be 'imagined' or constructed by using the same concepts, all of them being *meaningful.* This situation can be considered as satisfactory or unsatisfactory depending on the point of view, and the criterion can be found in considering that in the very nature of the notions of representation, model or image is inevitably included the termination '*of*', that indicates the link of those mental entities with 'something *of* which' they are images, models or representations. We shall express this fact by saying that these entities (and in the first place the *judgment* of which they are the content) have a *referential nature*. By this we mean that they 'refer' to some entity to which they are or can be applied and is, so to speak, the 'target' toward which they are oriented, or simply the *object* of which they are the representation, the model or the image. The 'something' so understood is usually called *referent* in semiotics and philosophy of language, and this is why we have spoken of the 'referential' nature of these mental entities. The fact that the concepts present in a given conceptual space can be used for constructing several 'meaningful' models or representations can be appreciated differently from the point of view of their referentiality. A positive appreciation consists in considering this fact as the advantage that a model can have a 'plurality' of applications; a negative appreciation consists in seeing this fact as an indeterminacy or 'vagueness' when our aim is to represent a precise 'intended' object.

We can call *mappings,* or *maps* for short, the intellectual entities just mentioned (i.e. representations, models and images) that can be constructed within one and the same conceptual space. They correspond to what are usually called *theories* in the empirical sciences, and, from what we have said, it appears that they must be evaluated not only from the point of view of their conceptual consistency, but also from their referential performance. This remark puts forward a really challenging task, that of providing *criteria of referentiality* by means of which we can 'check the maps'.

To give a first idea of the complexity of this task we can consider a well-known example, that is, the comparison of the corpuscular and the wave-theory of light. Light can be considered a 'thing' easily identified in ordinary experience, and the problem of its nature was approached within the territory of mechanics already in the 17th century, when a corpuscular theory of light was proposed by Newton (in which a light beam was considered as a swarm of microscopic material particles travelling along rectilinear paths in the empty space), whereas Huygens proposed a different theory, in which a light beam was considered as a wave, i.e. a perturbation propagating itself in an impalpable ‘luminiferous ether’ filling the whole cosmic space. This short presentation indicates that both theories have a ‘pictorial’ aspect that can be grasped in ordinary knowledge, nevertheless they cannot be compared with the cognitive tools of common sense. Light must become the *object* of a deeper scrutiny that took place first in the territory of mechanics, where both rival theories (in full keeping with the general mechanical science and both formulated in an impeccable mathematical form) were submitted to *experimental tests.* These tests consisted in a display of *operational manipulations* of the light beams by means of optical *instruments* that allowed for the exact measurement of several properties of the light beams, of which each rival theory was able to account for in a more or less convincing way, until the moment in which some ‘crucial experiments’ (essentially due to Young and Fresnel in the first two decades of the 19th century on the interference and diffraction of light) refuted the the corpuscular Theory We can express this event as the fact that the ‘map’ of the corpuscular theory had failed to prove adequate in the indication of a crucial point, whereas the other has succeeded in passing this test. Half a century later light became the object of a brand new territory, i.e., electromagnetism. This happened with the creation of Maxwell’s electromagnetic theory of light that removed the ‘mysterious’ nature of the luminiferous ether. replacing it with the concept of electromagnetic field. Yet Maxwell himself believed that some mechanical model of this field should be found in the future, and this was the hope of many scientists at the end of the 19th century, despite the difference of the concepts adopted and even of the experimental devices through which this field can be explored and monitored (as Hertz had shown by inventing—seven years after Maxwell’s death—his oscillator for the ascertaining of the electromagnetic waves). These hopes were the residuals of the reductionist view that had dominated much of the physical sciences in the 19th century and was rapidly dissolving. Already at the beginning of the 20th century, that is, with Einstein’s discovery of the photon in 1904 some aspects of the corpuscular view of light reappeared, but light had actually become the object of a new territory, quantum physics, and the photon was one of the elementary ‘particles’ for which both the wave and the corpuscular models are only partially adequate, that is, partial maps.

Conclusions

The substance of this discourse, that we have tried to make easily understandable by presenting an example, can by summarized by saying that scientific investigation, globally understood as the effort of knowing reality in the most reliable way, necessarily splits into *territories*, characterized by the selection of a few attributes of reality that will constitute the special point of view of each science, or the conceptual space within which the understanding of reality will be attempted. Within this conceptual space various *theories* are elaborated, each proposing a model or representation of those parts of reality that can be investigated within the special territory (or using the specific concepts) of the science concerned. These models, however, cannot refer to reality (even to those selected attributes of reality that have been chosen) unless they are equipped with *operational* tools that make possible to link the mental entities with the non-mental aspects of reality. They are *criteria of reference* thanks to which the models or images become *maps* whose components are supposed to match the *intended* features of reality that have been posited as the *objects* of the science concerned. In this sense the objects of a science are 'constructed' by the combined synergy of the conceptual elements of a theory and its operational criteria of reference, and for that are 'relative' to such choices, but not arbitrary. In fact they are necessary and (in a certain sense) sufficient for designing and performing the experiments that are destined to test the theory, but they cannot determine the *result* of the experiment. They are the conditions for formulating the question, but they do not offer the answer, they can indicate where we have to look for, but they do not predict what we will actually find.

If we now consider that the model, or image, or representation in which a theory consists must, in science, also be formulated in *statements*, we must derive from what we have said, that these statements must be *true*, not in an 'absolute' sense (that is, true in themselves and 'independently from anything else') but true *of* their intended objects, of their referents. Therefore, this 'relativity to the specific objects is the characteristic of scientific truth.

The failure of the program of finding mechanical models of all physical phenomena has been interpreted, by certain philosophers, at the end of the 19th century, as the evidence that science is unable to offer a reliable knowledge of reality. Nevertheless, this pessimistic view can be avoided by recognizing that science (including in it also the broad display of the social sciences and humanities) is constituted by a rich variety of 'territories', that is, of really distinct *disciplines* each having its own *domain of objects* which is also its specific 'conceptual space'. Within this conceptual space several *scientific theories* are proposed offering different ways of *mapping the territory.* Both moments rely upon an 'hermeneutic' activity of human reason that presides over concept formation and theory construction (through which interpretation and explanation of facts are proposed), accompanied by the presence of *operational procedures* that are essential for linking the maps with the objects approached within the territories. This justifies a *realistic conception of science*,[10] that is, a view that gives back to science its traditional aim of being a solid knowledge of *truth* which (though being different from 'certainty' which is only an epistemic feature, and even compatible with the

[10] See Agazzi (2016).

fallibility of the human cognition) is intrinsically the property of statements that correspond with the facts they describe. This truth is partial, but genuine, like the map of the underground of a city that is strictly limited to the indication of the stations and the lines connecting them and does not indicate, for example, where the cathedral or the central hospital are located. This, and other, information can be offered by other reliable maps of the city, and are mutually compatible and complementing each other. In a similar way, the complementation of different justified theories within a science, and of different sciences in the human effort of knowing and understanding reality can only increase the global amount of human knowledge, much better than any reductionist effort of artificially ‘unifying’ the sciences. Such a unity is certainly desirable, but must be pursued through a different effort, that is suggested by general system theory, according to which the different subsystem of a global system preserve their identity and specificity and at the same time are linked by a net of mutual relations that contribute to the functioning of the global system, that is endowed with properties and capable of performances that are neither similar to those of the single subsystems, nor thinkable as an addition of them.

References

E. Agazzi, *The Problem of Reductionism in Science* (editor) (Kluwer Academic Publisher, Dordrecht, Boston, London, 1991)

E. Agazzi, *Scientific Objectivity and its Contexts* (Springer, Cham, Heidelberg, New York, Dordrecht, 2014)

E. Agazzi, Scientific realism within perspectivism and perspectivism within scientific realism. Axiomathes **26**(2016), 349–365 (2016)

B. Barnes, *Interests and the Growth of Knowledge* (Routledge and Kegan Paul, London, 1977)

D. Bloor, *Knowledge and Social Imagery* (Routledge and Kegan Paul, London, 1976)

R. Descartes, *Oeuvres*, ed. by C. Adam, P. Tannery, vols. 13 (Cerf, Paris, 1897–1913), New revised edition: Vrin, Paris, 1964–1974

C. Dilworth, *Scientific Progress*, 4th edn. (Springer, Dordrecht, 2008)

S. Drake, *Discoveries and Opinions of Galileo* (Doubleday, New York, 1957)

S. Drake, Galileo’s new science of motion, in *Reason, Experiment and Mysticism*, ed. by M.L. Righini Bonelli, W. Shea (Science History Publications, New York, 1975), pp. 131–156

S. Drake, C.D. O’Malley, *The Controversy on the Comets of 1618* (Univ. of Pennsylvania Press, Philadelphia, 1966)

G. Galileo, *Opere*: *Le opere di Galileo Galilei*, Edizione Nazionale a cura di A. Favaro, Barbera, Firenze, 1929–1939. Repr. 1964–66, 20 vols

G. Galileo, *(1623) Il Saggiatore*, in *Opere*, cit., vol. VI, pp. 197–372. Partial English trans. by S. Drake in *Drake-O’Malley*: *The Assayer* (1966), pp. 151–336

G. Galileo, *(1632) Dialogo sopra i due massimi sistemi del mondo*, in *Opere*, cit., vol. VII. English trans.: *Dialogue Concerning the Two Chief World Systems*, transl. by Stillman Drake, forward by A. Eistein, 2nd edn. (University of California Press, Berkeley, Los Angeles, 1967)

G. Galileo, *(1638) Discorsi e dimostrazioni matematiche intorno a due nuove scienze*, in *Opere*, cit., vol. VIII. English trans.: *Two New Sciences*, trans. by S. Drake (University of Wisconsin Press, Madison, Wisc., 1974)

K.D. Knorr-Cetina, *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science* (Pergamon Press, Oxford, 1981)

T. Kuhn, *The Structure of Scientific Revolutions* (Chicago Univ. Press, Chicago, 1962)

J.O. La Mettrie, *(1748) L'homme machine,* Leyden. French-English edition with notes by G. Carman Bussey, *La Mettrie, L'Homme Machine, Man A Machine*, Open Court, La Salle, Ill., (1912)

B. Latour, S. Woolgar, *Laboratory Life: the Social Construction of Scientific Facts* (Sage, London, 1979)

Chapter 2
On the Ontology/Epistemology Distinction

Michele Marsonet

Ontology and Epistemology

There are good reasons for thinking that any sharp division between ontology and epistemology is untenable, because ontology is characterized by the fact that objects are standardly seen by us in terms of a conceptual apparatus that is substantially driven by mind-involving elements. In other words, we are not able to define an "an sich reality", i.e. a natural world conceived of in ways totally deprived of mind-involving concepts. The key factor here is the distinction between noumenal (=extra-phenomenal) reality, about which a great deal can be known in scientific theorizing, and an sich (="purely objective", altogether mind-independent) reality, about which precious little can be known from the point of departure of the standard conceptual scheme human beings deploy in science and everyday life alike. Any "absolutely objective" ontology is then left in the background. The so-called "objective facts" are always opaque, there is constantly a gap between a person's impressions and the world's actualities. In the end, where objective reality is at stake there is always cognitive opacity: being evident is something confined to the realm of subjectivity. Human beings are tied to their cognitive limits and to an imperfect and conceptual-based knowledge of the world.

So let's ask again: can we really draw a *precise* border line between ontology and epistemology? A positive answer to this question looks attractive at first sight, mainly because it reflects convictions deeply entrenched in our commonsense view of the world. However, anyone wishing to clarify the distinction between the ontological and the epistemological dimensions (without having recourse to unwarranted dogmas) should recognize that such a positive answer poses more problems than it is meant to solve. The separation between factual and conceptual,

M. Marsonet (✉)
Department of Philosophy, University of Genova, Genoa, Italy
e-mail: michele.marsonet@unige.it

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_2

in fact, is not sharp and clean, but rather fuzzy.[1] To this recognition another remark should be added. As long as humans are concerned, the world is characterized by a sort of "ontological opacity" which makes the construction of any absolute ontology very difficult. Our ontology is characterized by the fact that the things of nature are seen by us in terms of a *conceptual apparatus* that is substantially influenced by mind-involving elements. All this has important consequences on both the question of conceptual schemes and the realism/anti-realism debate.

Theoretically, we may admit that a clear distinction can be drawn between the natural world on the one hand, and the social-linguistic world on the other. However, it should not be difficult to understand that we began to identify ourselves and the objects that surround us only when the social-linguistic world emerged from the natural one, and this in turn means that our criteria of identification are essentially social and linguistic. Leaving aside any kind of Platonism, and recognizing in a pragmatist mood that the concept of "truth" is essentially tied to human interests, we need an intersubjective criterion giving rise to the notion of a world which is both objective and mind-independent. In other words, the distinction subject/object is not to be found *in* nature: it arises when men have such an intersubjective criterion, i.e., within a social world which is created by men themselves. It is important to note that these remarks do not entail the total identification of the aforementioned two worlds. We can admit that a border line between ontology and epistemology really exists but, as long as we are concerned, such a distinction looks less definable today than it was usually thought to be in the past.

There are two reasons which explain why things are so. On the one hand conceptualization gives us access to the world while, on the other, it is the most important feature of our *cultural* evolution (which is distinct from—although not totally alien to—biological evolution).[2] This does not mean to diminish the importance of the latter, which is specifically geared to the natural world and, after all, precedes our cultural development from the chronological viewpoint. However, it is cultural evolution that distinguishes us from all other living beings that happen to share our planet with us. While the idealistic thesis according to which the mind produces natural reality looks hardly tenable, it is reasonable to claim instead that we perceive this same reality by having recourse to the filter of a conceptual apparatus whose presence is, in turn, connected to the development of language and social organization.

This is the reason preventing the aforementioned precise distinction between ontology and epistemology. For example, it might be stated that ontology's task is to discover what kinds of entities make up the world ("what there is", in Quine's terms), while epistemology's job is to ascertain what are the principles by which we get to know reality. It is obvious, however, that if our conceptual apparatus is at

[1] The reference work in this case still is Quine's paper "Two Dogmas of Empiricism", in Quine (1980), pp. 20–46. For a more recent perspective see McDowell (1994).

[2] The distinction biological/cultural evolution is constantly present in pragmatist authors like James, Peirce, and Dewey. For a contemporary assessment see Rescher (1990).

work even when we try to pave our way towards an unconceptualized reality, our access to it entails anyhow the involvement of the mind. Resorting to a paradox, it might even be said that any unconceptualized reality turns out to be an *image* of the mind (even though, it is worth repeating it, this recognition does not force us to deny the mind-independent existence of unconceptualized reality).

At this point a crucial problem must be faced. Since the rejection of *any* scheme/content distinction looks hardly tenable,[3] the question arises whether it is more appropriate to speak of "scheme" (singular) or of "schemes" (plural). This is not a rhetorical question, as it might seem at first sight. What lies behind it is, rather, the question of ontological pluralism, which is in turn connected to the existence of possible alternative ways of conceptualizing the world.

The basic importance of such a question was understood by William James. At the beginning of the past century, in fact, he wrote that "It is possible to imagine alternative universes to the one we know, in which the most various grades and types of union should be embodied",[4] and went on saying that "The 'absolutely' true, meaning what no farther experience will ever alter, is that ideal vanishing-point towards which we imagine that all our temporary truths will some day converge […] meanwhile we have to live to-day by what truth we can get to-day, and be ready to-morrow to call it falsehood".[5] The conclusion of this line of reasoning is that the great scientific and metaphysical theories of the past were adequate for centuries but, since human experience has boiled over those limits, we now call these theories only relatively true. Those limits were in fact casual, and "might have been transcended by past theorists just as they are by present thinkers".[6]

Naturally James was not the first to note that our world-view can never be absolute, and that intelligent creatures whose experiential modes are substantially different from our own are bound to conceptualize reality in a rather diverse way. James, however, provided us with a clear picture which anticipates the contemporary debate on conceptual schemes. He claimed in this respect that "Were we lobsters, or bees, it might be that our organization would have led to our using quite different modes from these of apprehending our experiences. It might be too (we can not dogmatically deny this) that such categories, unimaginable by us to-day, would have proved on the whole as serviceable for handling our experiences mentally as those which we actually use".[7]

[3] See especially D. Davidson, "On the Very Idea of a Conceptual Scheme," in Davidson (1984), pp. 183–198. The paper was originally published in *Proceedings and Addresses of the American Philosophical Association*, 47, 1974, pp. 5–20. See also R. Rorty, "The World Well Lost", in Rorty (1982), pp. 3–18. I cannot take this problem into account here. For a criticism of Davidson's and Rorty's positions see Haack 199.

[4] James (1907), p. 156.

[5] *Ibid.*, pp. 222–223.

[6] *Ibid.*

[7] James (1907), p. 171.

Someone might object that these are only mental experiments, whose importance cannot be overvalued. However, mental experiments play a key role in both philosophy and science. No doubt they are hypothetical devices, but they also allow us to enter the dimension of *possibility*. By resorting to them we are able to imagine how the world could have been in the past, could be today, or could turn out to be in the future. This is a specific characteristic of our relationship with the world, which is strictly geared to the cultural type of evolution mentioned above. Rationality is, thus, largely a matter of *idealization*. Although our natural origins and evolutionary heritage must be duly deemed important, we must give way as well to the recognition that there is indeed something that makes us unique. Only human beings are able to take idealities into account and to somehow detach themselves from the present world. Rationality may also be seen as the expression of mankind's capacity to see not only how things actually *are*, but also how they *might have been* and how they *could turn out to be* if we were to take some courses of action rather than others: the concept of possibility plays indeed a key role. It should eventually be noted that the dimension of possibility plays quite an important function even in the scientific domain, since scientific theories concern possible rather than actual reality. Newton's theory of universal gravitation takes into account the ideal mass in ideal space, and its status of scientific theory is granted by the fact that it holds for *any* mass.

In short, possibility are a key component of our social-linguistic world, i.e., of the specifically human way of dealing with reality. Possible worlds and possible individuals are actual or potential products of our conceptual apparatus, and any strategy meant at eliminating them appears doomed for failure. The dimension of possibility, besides being strictly tied to hypothetical reasoning, plays a fundamental role in our comprehension of both the natural and social-linguistic worlds. But it should also be clear that the aforementioned dimension must anyhow make reference to some kind of agent, and the agent itself is thus an inevitable point of departure. We are compelled to adopt such a stance, because this is the only way opened to us for gaining accessibility to the world. No one denies that it would be good to transcend our conceptual machinery in order to glimpse at how the world really is, independently of any view we can hold about it. This, however, cannot be done because of the very way we happen to be made. Unlike some forms of idealism, we can recognize the presence of things that are real in the sense of being mind-independent but, on the other hand, it must be added that human beings have access to those things only through their conceptual apparatus.

Even scientific world-views continuously evolve, which means that the scientific enterprise has an essentially *historical* character. As Werner Heisenberg pointed out, science always is the result of the encounter between the natural world on the one side, and human conceptions, practical interests and needs on the other. Conceptual schemes determine our comprehensive world-view, but things will not change much if we shift to the scientific vision of the world. The appeal to mental experiments is useful not only in daily life, but in the scientific domain too, because in this case science itself makes us understand that it permits us to know the world from a particular perspective, which is in turn geared to the way we happen to be made and to the specific relationships we entertain with the environment which surrounds us.

The Jamesian point that it is possible to imagine alternative universes to the one we know, and that intelligent creatures whose experiential modes are substantially different from our own are bound to interpret reality in a diverse way, must be taken seriously. In other words, we should recognize that the natural environment in which we live (and of which we are a substantial part) has an essential bearing on conceptualization (including the scientific one). We would not conceptualize the world the way we do were we not sensitive to some physical parameters like, for instance, light or heat. Science provides reliable information on the world, but this information is always relative to a particular framework, and it is a mistake to think that the limits of our cognitive capacities only have an aprioristic character. We are mainly bound by *empirical* limits, due to the fact that we inquire into nature by means of an apparatus which answers certain stimuli, but not others. Nothing in our actual science leads us to rule out the hypothesis that, in other natural environments, the development of science might have taken quite a different course. In order to give plausibility to this hypothesis we must only admit the existence of worlds whose natural environment is substantially diverse from our own, and certainly this is not mere science fiction.

All this explains why the existence of the so-called conceptual schemes is one of the most controversial issues in philosophy today. Its importance lies in the fact that, depending upon what strategy one chooses to foster, this theme has an important bearing on many related questions, among which the problem of scientific realism, the relations between ontology and epistemology, and the role that our conceptualization of the world plays in a realist vs. idealist outlook on reality. As was hinted before, however, it would be wrong to assume that the issue is fundamental *only* for philosophy. For example, according to Niels Bohr's principle of complementarity we have, on the one side, a sort of Kantian world-in-itself which is both unknowable and indescribable, and on the other side an "us" which, unlike in Kant's picture, is not stable and determined.[8] This means that, in our inquiries about the world, different questions can all receive coherent answers, with the disquieting effect that a comprehensive and coherent image of reality cannot be achieved. It is as if, conducting different experiments, we were to change conceptual scheme: the world experienced will in any case be diverse, and there is no way to combine the world of our experience with the various, differing conceptual schemes.

[8] See Stapp (1993).

Davidson and Rorty on Conceptual Schemes

A conceptual scheme is, according to some dictionaries of philosophy, “a set of concepts and propositions that provide a framework for describing and explaining items of some subject-matter along with criteria for recognizing which phenomena are to be considered deviant and in need of explanation”,[9] or “the general system of concepts with which we organize our thoughts and perceptions. The outstanding elements of our everyday conceptual scheme include spatial and temporal relations, other persons, meaning-bearing utterances of others, and so on. To see the world as containing such things is to share this much of our conceptual scheme”.[10]

It follows from the previous general definitions that, when dealing with conceptual schemes, philosophers take into account the beliefs and assumptions formulated, for example, in science and morality. The comprehensive outlooks on the world generated by some community—when taken together—form an inclusive theory in terms of which the members of that community explain and interpret both their empirical and moral experience. The limits imposed on the term “community”, on the other hand, are determined by philosophers themselves. “Community” may mean in this context a particular society or all members of humankind.

The key point of the contemporary debate on conceptual schemes is, however, the following. Given the fact that thought (i.e. the manipulation of concepts) is not possible without the existence of language, conceptual schemes are most of the times (although not always) identified with languages or, even better, with sets of *inter translatable* languages. If this is true, learning a language means to acquire the conceptual scheme it embodies and, as a matter of fact, according to this view a conceptual schemes *is* a language. Moreover, as was mentioned previously, one can ask whether there is only one conceptual scheme or many; if a plurality of conceptual schemes is admitted the problem of *relativism* arises. To sum up, a conceptual scheme is a “frame of reference”, that is to say the view-point, or set of presuppositions or of evaluative criteria within which a person’s perception and thinking always occur, and which constrains in a selective way the course and outcome of these human activities.

It is well known that, according to Donald Davidson, men would be unable to interpret speech from a different conceptual scheme as even meaningful. He claims that, since translation proceeds according to the “principle of charity”, and since it must be possible for an omniscient translator to make sense of what we say and of how we behave, we can be assured that most of the beliefs formed within the commonsense conceptual framework are true. Davidson challenges the scheme-content dualism, and mentions both “a dualism of total scheme (or language) and uninterpreted content”, and “a dualism of conceptual scheme and

[9] H. I. Brown, “Conceptual scheme”, in Honderich (1995), pp. 146–147.

[10] “Conceptual scheme”, in Blackburn (1996), pp. 72–73.

empirical content".[11] All this may be seen as the contemporary way of dressing the Kantian distinction between the contents that the noumenal world sends to us and the forms that we place upon them thanks to the particular structure of our categorial system. What we have here is a real dichotomy between these two elements, in the sense that the (conceptual) scheme is "other than" the (non-conceptual) content that is opposed to it. It goes without saying that the characterization Davidson takes into account and criticizes is subject to the attacks that Sellars, Quine and others addressed to the so-called "Myth of the Given".

According to such a view, there are effects that the external world produces on our senses. We cannot, so to speak, defend ourselves from these effects: the fact that we are in the world means that we are inevitably affected by external reality. They are episodes experimented through our senses, and only later we are able to locate them into a conceptual framework. The scheme/content dualism is thus explained: such a dualism works in so far as the scheme is given a conceptual—and independent—character while the content is something totally different, with no mediations at all. Scheme and content are indeed opposed elements, and their ontological status is obviously diverse.

Davidson's attack to the scheme-content distinction is supported by a set of arguments purporting to reject, first of all, the thesis that totally different conceptual schemes can actually exist. To put things in a sketchy manner, he equates having a conceptual scheme with having a language, so that we face the following elements: (1) language as the organizing force; (2) what is organized, referred to as "experience", "the stream of sensory experience", "physical evidence"; and finally (3) the failure of inter translatability. It follows that "It is essential to this idea that there be something neutral and common that lies outside all schemes […] The idea is then that something is a language, and associated with a conceptual scheme, whether we can translate it or not, if it stands in a certain relation (predicting, organizing, facing, or fitting) with experience (nature, reality, sensory promptings)".[12]

If this is the situation—Davidson goes on—then we could say that conceptual schemes that are different in a radical way from each other correspond to languages that are not inter translatable. How can we, however, make sense of a total failure of inter translatability among languages? For sure "we could not be in a position to judge that others had concepts or beliefs radically different from our own".[13] Davidson's conclusion is that if one gives up the dualism of scheme and world, he will not give up the world, but will instead be able to "re-establish unmediated touch with the familiar objects whose antics make our sentences and opinions true".[14]

On his part Richard Rorty fully endorses Davidson's stance. Starting from Quine's rejection of the notion of "meaning" as anything transcending what is

[11] D. Davidson, "On the Very Idea of a Conceptual Scheme", cit., pp. 187 and 189.

[12] *Ibid.*, pp. 190–1.

[13] *Ibid.*, p. 197.

[14] *Ibid.*, p. 198.

contextually defined in predicting the behavior of other people, Rorty deems it impossible to distinguish an untranslatable language from no language: “Once we imagine different ways of carving up the world, nothing could stop us from attributing ‘untranslatable languages’ to *anything* that emits a variety of signals”.[15]

For example, we may imagine aliens who endorse a totally different view of the world, so that their language would—in principle—turn out to be untranslatable into our own. This means that they would carve up the world according to a completely different conceptual scheme. However Rorty thinks that “[…] for all we know, our contemporary world is filled with unrecognizable persons. Why should we ignore the possibility that the trees and the bats and the butterflies all have their various untranslatable languages in which they are busily expressing their beliefs and desires to one another? […] So I think that to rule the butterflies out is to rule out the Galactics and the Neanderthals, and that to allow extrapolation to the latter is to allow for the possibility that the very same beliefs and desires which our Galactic descendants will hold are being held even now by the butterflies”.[16]

It follows that any language must, to count as a language, be translatable into our own, and that—quite surprisingly—the large majority of our present beliefs must be true. In other words, “[…] ‘the world’ will just be the stars, the people, the tables, and the grass—all those things which nobody except the occasional ‘scientific realist’ philosopher thinks might not exist. The fact that the vast majority of our beliefs must be true will, on this view, guarantee the existence of the vast majority of the things we now think we are talking about”.[17] By endorsing this line of thought, we no longer need the notion of “the world” conceived of as an “independent reality”, a notion which is endorsed by those thinkers who claim that different conceptual schemes carve up the world differently.

Davidson’s and Rorty’s solution is radical, but we are bound to ask at this point what the expressions “reality” and “world” mean for them. Let us assume that they can be identified with the world of common sense which is formed by the familiar objects whose antics—as Davidson says—make our sentences and opinions true or false. These familiar objects are tables, chairs, houses, stars, etc., just as we perceive them in our daily life. One is not entitled to ignore, however, that the current discussions on the problem of scientific realism arise because there is a strong asymmetry between the commonsense view of the world and the scientific one (or the “manifest” and the “scientific” images of man-in-the-world, to put it in Wilfrid Sellars’ terms).[18] For instance, the table that we see with our eyes is not the same table that we “see” through the mediation of scientific instruments, and this fact is not trivial. It is rather easy to reach a high level of intersubjective agreement among the individuals present in a room about the color, size and weight of a table, and it can also be granted that we form our beliefs in this regard by triangulating—in a

[15] R. Rorty, “The World Well Lost”, cit., p. 6.

[16] *Ibid.*, pp. 9–10.

[17] *Ibid.*, p. 14.

[18] W. Sellars, “Philosophy and the Scientific Image of Man,” in Sellars (1963), pp. 1–40.

Davidsonian sense—with our interlocutors and the surrounding environment. Such an agreement, however, becomes problematic when we try to reconcile this vision of the world with what present science tells us about it.

So being in touch with such familiar objects as tables, chairs and stars "most of the time" (as Rorty specifies) has a fundamental bearing only on the ontology of common sense, since our actual science shows that quite a different representation of reality can actually be provided (or, even better, it shows that those objects might not exist as men perceive them). Naturally, one can always resort to an objection of the following kind: why should we deem the table viewed as a collection of subatomic particles more important than the table that our eyes see in daily life? After all, we can conduct our life well enough even ignoring what science claims (just like men did for many thousand years). This, however, looks like a serious under evaluation of the scientific enterprise.

It may be noted at this point that the contemporary approach to the problem of conceptual schemes is nothing but the reformulation of an old question, i.e., that of cognitive adequacy from the standpoint of an entirely different sort of cognitive beings. Today some authors seem to exclude this possibility from the onset, but one need not follow them on such a path. It seems to us that we should not uncritically accept Davidson's statement that "since there is at most one world, these pluralities are metaphorical or merely imagined".[19]

Davidson, as we saw before, associates conceptual schemes with languages, and then adopts linguistic inter translatability as *the* identity criterion for conceptual schemes themselves. Subsequently we are told that, in order to call something "a language," say L_0, we must be ready to accept the idea that the statements of L_0 can be translated into those of our own language (let us call it L_1). It easily follows from this line of reasoning that, if this cannot be done, L_0 is not a language at all. According to Davidson "we must conclude […] that the attempt to give a solid meaning to the idea of conceptual relativism, and hence to the idea of a conceptual scheme, fares no better when based on a partial failure of translation than when based on total failure. Given the underlying methodology of interpretation, we could not be in a position to judge that others had concepts or beliefs radically different from our own".[20] One may point out, however, that linguistic inter translatability cannot be such an absolute criterion, because in certain circumstances we are able to realize that some sort of language is used, even though we cannot translate it into our own language.

Larry Laudan has noted in this regard that there is no reason to assume the presence of different world-views only when there are no criteria of inter translatability among them. "Only with the so-called linguistic turn"—he claims—"have philosophers supposed that conceptual schemehood is to be understood in terms of non-translatability. Aristotle's cosmos and Einstein's universe represent very different world-views. With Davidson, I believe that each can be made intelligible to

[19] D. Davidson, "On the Very Idea of a Conceptual Scheme", cit., p. 187.

[20] *Ibid.*, p. 197.

adherents of the other. But only someone as wedded to the translation thesis as Davidson is would imagine that the latter fact (viz., inter translatability) constitutes grounds for denying that they represent different conceptual schemes".[21] The absolute primacy that Davidson places on translatability should thus be rejected, hence Laudan's proposal to identify conceptual schemes on ontological, axiological and methodological (and not exclusively linguistic) criteria.

The fact is that Davidson resorts to a sort of "pansemanticism" which sees linguistic behavior as the *only* behavior that really counts, while Laudan's approach is more articulated. The above mentioned pansemanticism endorsed by Davidson clearly transpires when he tells us that "[…] if all we know is what sentences a speaker holds true, and we cannot assume that his language is our own, then we cannot take even a first step towards interpretation without knowing or assuming a great deal about the speaker's beliefs. Since knowledge of beliefs comes only with the ability to interpret words, the only possibility at the start is to assume general agreement on beliefs."[22] Here we have a good reason for claiming that Davidson is far less "post-analytic" than Rorty depicts him to be. The overemphasis placed upon linguistic behavior is, in fact, a typical trait of the analytic school, which tends to forget the fact that, after all, man came first and language later. Language is a relatively recent factor in the history of our evolution, as science shows us, and many parts of our behavior are guided by non-linguistic criteria. We can avoid the aforementioned analytic overemphasis only by recognizing that language is *not* the whole of reality, but a social product created essentially for practical purposes.

Quine and Conceptual Schemes

How can we be sure, however, that our beliefs really bear on the world? If we read carefully Quine's rejection of the second dogma, it is clear that he relates the empirical significance of our statements about the world with their possibility of being subject to what he calls, in Kantian terms, "the tribunal of experience".[23] So it turns out that when we want to verify whether a belief of ours reflects what there is, the recourse to experience is always necessary. In other words, beliefs may be accepted if, and only if, the judgment on their validity ultimately rests on experience itself. However Quine obviously deems language, too, very important, and this means that language must be accommodated into the picture if the picture itself

[21] Laudan (1996), p. 13.

[22] D. Davidson, *ibid.*, p. 196.

[23] "Our statements about the external world face the tribunal of sense experience not individually but only as a corporate body" (W. V. Quine, "Two Dogmas of Empiricism", cit., p. 41).

purports to be coherent. “It is obvious”—he states—“that truth in general depends on both language and extra linguistic fact.”[24] On the one hand extra linguistic reality and language are not identified but, on the other, Quine’s thesis is that they cannot be separated with a neat border line: “Taken collectively, science has its double dependence upon language and experience; but this duality is not significantly traceable into the statements of science taken one by one”.[25] Starting from these premises, once the second dogma is rejected also the first one is: no statement can ever be free from an ultimate reference to experience. All this means that conceptual schemes or world-views, like the ones provided, say, by Newtonian mechanics or quantum theory, are the primary bearers of truth: the truth of a statement strictly depends on the particular conceptual scheme one currently adopts. And it may be noted that this is confirmed by scientific practice. It is the theory (for instance quantum theory) which a scientist endorses that instructs him in the use of the notations he currently works with.[26]

This is why in Quine’s empiricism without dogmas both language and experience play a key role. Not only is truth—which can be primarily predicated of a conceptual scheme—dependent on both language and experience, but language itself appears to be a factor which cannot be equated *totally* with experience: it manages to maintain somehow a certain independence. When dealing with the formation of beliefs, we must take into account an “external” element (experience) and an “internal” one (language). The question, of course, is to find out how this internal factor can be properly accommodated into the Quinean picture, because it is not difficult to see that “language” plays here the same role that “meaning” used to play in the analytic/synthetic dualism that Quine rejects. It follows that in Quine’s picture the dichotomy between statements true in virtue of meaning alone, and statements whose truth depends also on how the external world is, is not overcome completely: it is given, rather, a new formulation.

Within our conceptual scheme we can appeal to something outside the system, i.e. the world, so that concepts and beliefs are somehow controlled by external constraints. The story with language is however different, since language, in Quine’s view, does not seem to be a factor whose ultimate legitimacy relies on something outside the conceptual sphere, and this means in turn that we face once again a dualistic situation. Despite the many oscillations present in Quine’s writings, we obviously have a real distinction which works for whole systems: the empirical content of a conceptual system or world-view never determines, just by itself, its empirical significance, because in any case we also need the contribution of language in order to make our picture coherent. Ontology is thus “internal” to language, which means that it is internal to a conceptual scheme or world-view. We

[24] *Ibid.,* p. 36.

[25] *Ibid.,* p. 42.

[26] See Putnam (1995), pp. 59–61.

can now see that language becomes the factor that guarantees our autonomy from the natural world. It is a limited kind of autonomy, but its limited range is sufficient for somehow granting us a special status in the natural order of things. And just at this point we meet some relevant difficulties.

Quine rejects the old notion of "meaning", but at the same time he underlines the presence and the extent of man's conceptual sovereignty in the formation of conceptual schemes or world-views.[27] This amounts to saying that the so-called "empirical significance" is something more than mere empirical content: we somehow have a *creative* role in the elaboration of conceptual schemes. Quine goes on claiming that "The philosopher's task differs from the others' […] in no such drastic way as those suppose who imagine for the philosopher a vantage point outside the conceptual scheme that he takes in charge. There is no such cosmic exile. He cannot study and revise the fundamental conceptual scheme of science and common sense without having some conceptual scheme, whether the same or another no less in need of philosophical scrutiny, in which to work".[28]

The key point here is the following, although we formulate it in non-Quinean terms. The content of the world (reality) and the content of a conceptual scheme (world-view) do not coincide, and this happens because of the just mentioned conceptual sovereignty we exert. The conceptual sphere is not reducible to the natural order of things: it is the realm of rationality *and* of meaning.[29] But a natural temptation arises. In other words, we might be tempted to say that the content of conceptual schemes is a pure product of our mind, a mind that operates more or less freely and which is not controlled by external limits. Quine, of course, does not endorse such a stance, but the real nature of the conceptual sphere remains in his works somehow mysterious: his behaviorism does not explain its presence.

For Quine empirical significance is an acceptable notion because we can explain it entirely in terms of the working of receptivity. But the previous remarks make us reflect on the fact that, according to this picture, not everything can be investigated in terms of natural science: there is, here, room for a return of the a priori on the stage. Language is something whose ultimate comprehension lies outside the domain of science. It is also worth noting how strong the influence of Quine's lesson is on Davidson, despite all their proclaimed differences. In a recent article by Davidson we find the following remarks: "It would be good if we could say how

[27] "We cannot strip away the conceptual trappings sentence by sentence and leave a description of the objective world; but we can investigate the world, and man as a part of it, and thus find out what cues he could have of what goes on around him. Subtracting his cues from his world view, we get man's net contribution as the difference. This domain marks the extent of man's conceptual sovereignty—the domain within which he can revise theory while saving the data". Quine (1994), p. 5.

[28] *Ibid.,* pp. 275–276.

[29] The similarities between Quine's and Ajdukiewicz's conceptions in this regard are analyzed in (Jakubiec 1986).

language came into existence in the first place, or at least give an account of how an individual learns his first language, given that others in his environment are already linguistically accomplished. These matters are, however, beyond the bounds of reasonable philosophical speculation".[30] This is quite an important statement. It means, in the first place, that our capacity of describing correctly the surrounding environment is taken more or less for granted and, furthermore, that the basis of language cannot (neither needs to) be explained.

In Quine, however, positions like this do not fit well the rest of his speculative building. If we conceive experience as the stimulation of sensory receptors—as Quine does—we seem to rule out the possibility of rational links between experience itself and beliefs, and conceptual schemes may be viewed not just as piecemeal beliefs, but rather as sets of logically interconnected beliefs. And this, in turn, seems to be a vindication of Sellars' theses. Sellars told us that the world of concepts is essentially formed by rational relations. In his most famous essay he claimed that when we describe the "states" that lead us to knowledge we not only describe them empirically, but also locate them in a logical space which has a rational character. And only within this logical-rational space are we able to *justify* what we say.[31]

Let then ask ourselves: Does the word "language" convey the same complexity as the term "conceptual scheme"? Quine, following Davidson's criticisms, subsequently abandoned the notion of conceptual schemes in favor of languages. He wrote: "It seems that in Davidson's mind the purported third dogma is somehow bound up with a puzzling use on my part of the phrase 'conceptual scheme' […] I have meant it as an ordinary language, serving no technical function […] A triad—conceptual scheme, language, and world—is not what I envisage. I think rather, like Davidson, in terms of language and the world. I scout the *tertium quid* as a myth of a museum of labeled ideas. Where I have spoken of a conceptual scheme I could have spoken of a language".[32]

In my view these statements make things somewhat too simple. Quine's acceptance of Davidson's arguments is too fast: he surrenders the theses put forward by his former pupil. Conceptual schemes are not reducible to successions of scattered beliefs. There is a structure in conceptual schemes because, as was previously noted, they are characterized by logical relations which hold beliefs together. Davidson's picture is articulated into the triad objects-causes-beliefs, but such a picture does not seem to accommodate the "conceptual sovereignty" which plays such an important role in Quine's *Word and Object*.

[30] D. Davidson, "Three Varieties of Knowledge", in Phillips Griffiths (1991), p. 157.

[31] W. Sellars, "Empiricism and the Philosophy of Mind", in Sellars (1963), p. 169.

[32] W. V. Quine, "On the Very Idea of a Third Dogma", in Quine (1981), p. 41.

Ajdukiewicz on Conceptual Apparatuses

Strangely enough, when one gets involved in the contemporary debate on conceptual schemes he very rarely finds mention of Kazimierz Ajdukiewicz.[33] However, the theses of the Polish philosopher in this regard are quite important, and I deem his contribution to be no less original—and no less controversial—than those of Quine, Davidson and Rorty. Let us note from the onset that even for Ajdukiewicz conceptual schemes *are* languages. Furthermore, I would like to point out that Ajdukiewicz wrote his main essays on this topic in the 1930s, i.e. half a century before the debate on conceptual schemes started following the publication of Davidson's paper in 1974. This confirms, I believe, the originality of his approach. Wolenski rightly notes that Ajdukiewicz's approach is linguistic all the way down, and that he devised no evolutionary structure in language. But this, of course, is no reason for denying the importance of his theses for the contemporary debate: Ajdukiewicz's limits, after all, are the limits of analytic philosophy itself.

Jan Wolenski writes in this regard that "Ajdukiewicz treated cognitive processes as inseparably connected with language: we always think in some language, and our statements are meanings which are attributes of sentences in some language L. Hence cognition, or, to put it more rigorously, cognition as a product, can be identified with the meaning of sentences. This is the essential methodological intention of Ajdukiewicz's semantic epistemology".[34] In the Polish philosopher's works the expression "conceptual apparatus" replaces "conceptual scheme". Ajdukiewicz's notion of conceptual apparatus is strictly tied to close and connected languages, and thus has a more technical connotation than what is meant today by the expression "conceptual scheme". Wolenski tells us that "the class of meanings of a closed and connected language was termed by Ajdukiewicz the conceptual apparatus of that language. It follows from the appropriate definitions that two conceptual apparatuses are either identical or have no element in common. And a consequence […] is that if two conceptual apparatuses have at least one element in common, then they are identical. Thus conceptual apparatuses never overlap. Ajdukiewicz held that every meaning belongs to some conceptual apparatus. Hence open languages are mixtures of various conceptual apparatuses".[35]

Let me note, at this point, that Ajdukiewicz's theses are subject to the same criticisms I previously addressed to Davidson and Rorty. Laudan's remark that only a full endorsement of the linguistic turn's main tenets may explain why so many philosophers insist on equating conceptual schemehood to languagehood apply to Ajdukiewicz as well. Only now, following the rise of post-analytic philosophy and the rediscovery of pragmatism, the basic tenet according to which linguistic

[33] With the exception, of course, of such texts dealing specifically with Polish philosophy of our century such as Skolimowski (1967) and Wolenski (1989).

[34] Wolenski (1989), p. 199.

[35] *Ibid.*, pp. 204–205.

behavior is the sole behavior that really matters has openly been challenged.[36] We all know, of course, that Ajdukiewicz's conception of language is an autonomous one, since he took language to be a product which is independent of action. Wolenski reminds us that "[…] the problem is clarified immediately when we consider the fact that Ajdukiewicz was not interested in the origins of a language, but in language as a product. The thesis on the autonomy of language acquires meaning when we bear in mind the difference between actions and their products (taken over by Ajdukiewicz from Twardowski). An objective assignment of meanings to expressions is possible only when language is treated as a product".[37]

Yet, this fact does not distinguish Ajdukiewicz's ideas from the main stream of the linguistic philosophy of the past century. What makes Ajdukiewicz's thought so appealing is the fact that he cleverly anticipated, in the 1930s, many theses that are commonly discussed today. So we find out that his "radical conventionalism", despite its several shortcomings, has many precious insights too, because in a famous paper dating back to 1934 he wrote: "Of all the judgments which we accept and which accordingly constitute our entire world-picture, *none* is unambiguously determined by experimental data; every one of them depends on the conceptual apparatus we choose to use in representing experiential data. We can choose, however, one or another conceptual apparatus which will affect our whole world picture".[38]

Subsequently Quine became famous for saying more or less the same thing, while Ajdukiewicz's contributions are still ignored by most Western philosophers. Quine remarks that there are many implicit background assumptions which make all the difference to how we interpret our experiences, and how we make our final evaluation of statements. This means that we cannot simply get meaning from experience, since there are no "neutral" observations available to men. And it is precisely because our conceptual judgments meet experience as a *body* that we must allow for possible revisions at any place within that body, so that "no statement is immune to revision".[39] And, if this is right, we must even allow for the possibility of changes in our verdicts on what is experienced itself.

Compare Quine's statements with the following by Ajdukiewicz: "No articulated judgment is absolutely forced on us by the data of experience. Experiential data do indeed force us to accept certain judgments if also we are based on a particular conceptual apparatus. However, if we change this conceptual apparatus, we are freed of the necessity of accepting these judgments despite the presence of the same

[36] See for example Devitt (1991) and N. Rescher, "The Rise and Fall of Analytic Philosophy", in Rescher (1994), pp. 31–42.

[37] Wolenski (1989), p. 205.

[38] K. Ajdukiewicz, "The World-Picture and the Conceptual Apparatus", in Ajdukiewicz (1978), p. 67.

[39] V. V. Quine, "Two Dogmas of Empiricism", cit., p. 43.

experiential data".[40] It is clear that what Quine defines as our "conceptual sovereignty" plays a key role even in this context, although the words used by the two authors are not the same. In any event, the striking similarity between the two philosophers is clearly detectable, once again, in the following statements by Ajdukiewicz (written in 1935): "Even the epistemologist cannot speak without a language, cannot think without a conceptual apparatus. He will thus make his decision as to truth in a way which corresponds to his world-perspective".[41]

Not only that: even logic is, according to the Polish philosopher, "relative to" a particular conceptual apparatus, and a change in the conceptual apparatus means a change in logic, too.[42] On his part, Quine claims that "revision even of the logical law of the excluded middle has been proposed as a means of simplifying quantum mechanics; and what difference is there in principle between such a shift and the shift whereby Kepler superseded Ptolemy, or Einstein Newton, or Darwin Aristotle?".[43]

Conclusions

In the final analysis I deem it necessary to point out that conceptual schemes are neither born out of nothing nor established on aprioristic bases. Their aim is to provide us with means for thinking about—and for speaking of—a reality which includes ourselves. We can add four partial definitions of "conceptual schemes" to the ones provided previously. In a first sense they are (A) *sets of socially codified beliefs*, that is to say belief-structures that are warranted by social use. In a second sense conceptual schemes are (B) sets of *logically interconnected* beliefs, i.e. structures in which our conceptual sovereignty above nature plays an essential role. In a third sense conceptual schemes are (C) *world-views*, i.e. interpretations of the world. In a fourth sense they are (D) *operational perspectives* on the world, i.e. means by which men interact with the surrounding environment. Meanings (A), (B), (C) and (D) are related to one another. In each case, conceptual schemes are instruments devised for *practical* purposes. By stressing this fact we wish to rule out any attempt to reify conceptual schemes, to conceive of them as self-subsistent and metaphysical entities which exist independently of human subjects and social structures. In our view they are primarily tied to the dimension of human action, and must be seen as elements of the agent/environment interaction.

As a matter of fact data concerning *non*-verbal action and behavior can lead us to ascribe beliefs in quite a plausible way. No doubt translatability helps a great deal, but certainly it is *not* an a priori condition for ascribing beliefs. These remarks pave

[40] K. Ajdukiewicz, *ibid.*, p. 72.

[41] K. Ajdukiewicz, "The Scientific World-Perspective", in Ajdukiewicz (1978), p. 117.

[42] Wolenski (1989), p. 208.

[43] W. V. Quine, "Two Dogmas of Empiricism", cit., p. 43.

the way towards understanding what conceptual schemes really are. They are a sort of *practical* metaphor which is supposed to convey the outcome of our categorization of reality. One should always be careful not to ascribe to them any metaphysical or self-subsistent feature: in other words, we must produce no *reification* of conceptual schemes, because their real nature is practical and functional. In order to understand what a conceptual scheme is we must not have recourse to abstract idealizations, because the comprehension of its nature can only be achieved by looking at how it *works*. Dewey's idea that our explanatory mechanisms are themselves the products of inquiry, in turn, opens the door to another key notion: "conceptual innovation". If we look at the history of science, for example, it is easily understandable that we, men living in the twentieth century, form our conception of the sun in quite different terms from those of Aristotle, or our conception of the heart in terms very different from those of Galen. The presence of different conceptual schemes may thus be explained by the process of conceptual innovation which—at least thus far—never came to an end in human history.

We should thus challenge Davidson when he says that "we get a new out of an old scheme when the speakers of a language come to accept as true an important range of sentences they previously took to be false."[44] The point at stake is in fact different, since a change of scheme is not just a matter of saying things differently, but rather of saying different (in the sense of *new*) things.

In other words, a scheme *A* may be committed to phenomena that another scheme *B* cannot even envisage: Galenic physicians, for instance, had nothing to say about viruses because those entities lay totally beyond their conceptual dimension. This means that our classical logic based on the principle of bivalence is not much help in such a context. Some assertions that are deemed to be true in a certain scheme may have no value in another scheme, so that we need to formalize this truth-indeterminacy by having recourse, say, to a Lukasiewicz-style many-valued logical system in which, besides the classical *T* and *F*, a third (Indeterminate) value *I* is present. We have, in sum, a much more complex picture than the one contained in Davidson's paper. It is important to note, once again, some hints contained in Ajdukiewicz's works. Jerzy Giedymin writes that "If different world-pictures cannot be compared either logically […] or experimentally, are they equally good or can they not be compared in any way whatever?—They can be compared and evaluated in the process of '*human understanding*' […] or from an '*evolutionary*' point of view".[45] In other words, we can understand Galenic medicine, but a Galenic physician would lack the conceptual apparatus for understanding ours.

So, to deny that different conceptual schemes exist is a little absurd. Of course, as I said previously, the expression "conceptual schemes" is a metaphor: we cannot see or touch them as we do with physical objects. Their presence, however, is

[44] D. Davidson, "On the Very Idea Idea of a Conceptual Scheme," cit., p. 188.

[45] J. Giedymin, "Editor's Introduction", in Ajdukiewicz (1978), p. xl. Giedymin refers to Ajdukiewicz's 1935 paper "The Scientific World-Perspective", cit.

detectable from human behavior, and this means that they are tied to the dimension of human *action*. Conceptual schemes, in sum, evolve, because they are processes and not immutable structures.

One should always take into account the broader models (conceptual schemes, cultural traditions) by means of which we judge our sentences—including, for example, the mythological ones—to be true or false. They are part of the "framework of conceptual thinking"[46] and, as long as men are concerned, they can think because they are able to measure their thoughts by having recourse to standards of correctness and of relevance. The aforementioned "framework of conceptual thinking" somehow transcends the individual thought of individual thinkers. This explains why there is truth and error with respect to it, even though we may talk of entities which do not exist in the physical world. There is indeed a correct and an incorrect way to describe this framework.

As it was said previously, we must not take conceptual schemes to be independent and metaphysical entities detached from any form of life. This fact gives them a sort of opacity which makes any kind of definition unsatisfactory from a purely logical point of view. Every time we try to overcome the metaphorical level of discourse we run into trouble, and any attempt at defining precisely what a scheme is, apart from the practical and functional role it plays in our cognitive endeavors, seems doomed for failure. We seem somehow to be prisoners of the metaphor we ourselves have devised, and this is the problem faced every time we try to set up a clear distinction between ontology and epistemology.

University of Genoa

Chair of Philosophy of Science

Dean, School of Humanities

References

K. Ajdukiewicz, *The Scientific World-Perspective and Other Essays, 1931–1963* (Reidel, Dordrecht, 1978)

S. Blackburn, *Oxford Dictionary of Philosophy* (Oxford University Press, Oxford, 1996)

H.I. Brown, Conceptual scheme, Hondelrich (1995), pp. 146–147

D. Davidson, (1984), On the very idea of a conceptual scheme, in *Inquiries into Truth and Interpretation*, (Clarendon Press, Oxford, 1985)

D. Davidson, The structure and content of truth. J. Philos. **57**(1990), 279–328 (1990)

D. Davidson, (1996), Subjective, intersubjective, objective, in *Realism and Truth*, ed. by P. Coates, M. Devitt, 2nd edn. (Oxford, Blackwell, 1991)

T. Honderlich, *The Oxford Companion to Philosophy* (Oxford University Press, Oxford, 1995)

S. Haack, *Evidence and Inquiry. Towards Reconstruction in Epistemology* (Blackwell, Oxford, 1993)

[46] For a definition of this expression see W. Sellars, "Philosophy and the Scientific Image of Man", cit.

W. James, *Pragmatism* (Longmans Green & Co, London, New York, 1907)

H. Jakubiec, J. Wolenski, Ajdukiewicz and Quine. Sci. Sci. **5**(1986), 83–98 (1986)

L. Laudan, *Beyond Positivism and Relativism* (Westview Press, Boulder-San Francisco-Oxford, 1996)

J. McDowell, *Mind and World* (Harvard University Press, Cambridge (Mass.), 1994)

H. Putnam, *Pragmatism: An Open Question* (Blackwell, Oxford, 1995)

W.V.O. Quine, Two Dogmas of Empiricism, in *From a Logical Point of View* (Harvard University Press, Cambridge (Mass.), 1980), 4th printing, pp. 20–46

W.V.O. Quine, *Word and Object* (The MIT Press, Cambridge (Mass.), 1994), 20th printing

N. Rescher, *The Riddle of Existence* (University Press of America, Washington D.C., 1984)

N. Rescher, *The Strife of Systems* (University of Pittsburgh Press, Pittsburgh, 1985)

N. Rescher, *Scientific Realism* (Reidel, Dordrecht-Boston, 1987)

N. Rescher, *A Useful Inheritance: Evolutionary Aspects of the Theory of Knowledge* (Rowman & Littlefield, Savage, 1990)

N. Rescher, *A System of Pragmatic Idealism*, vol. I (Princeton University Press, Princeton, 1992)

N. Rescher, *A System of Pragmatic Idealism*, vol. II (Princeton University Press, Princeton, 1994a)

N. Rescher, *American Philosophy Today and Other Philosophical Studies* (Rowman & Littlefield, Lanham, 1994b)

R. Rorty, The world well lost, in *Consequences of Pragmatism* (University of Minnesota Press, Minneapolis, 1982), pp. 3–18

W. Sellars, Philosophy and the scientific image of man, in *Science, Perception and Reality* (Routledge & Kegan Paul, London, 1963), pp. 1–40

W. Sellars, Empiricism and the philosophy of mind. Science, Perception and Reality **1963**, 127–196 (1963)

W. Sellars, Language as thought and as communication. Philos. Phenomenol. Res. **29**(1969), 506–527 (1969)

H. Skolimowski, *Polish Analytic Philosophy* (Routledge & Kegan Paul, London, 1967)

H. Stapp, *Mind, Matter, and Quantum Mechanics* (Springer-Verlag, Berlin, Heidelberg, New York, 1993)

J. Wolenski, *Logic and Philosophy in the Lvov-Warsaw School* (Kluwer, Dordrecht, 1989)

Chapter 3
Intuition in Classical Indian Philosophy: *Laying the Foundation for a Cross-Cultural Study*

Anand Jayprakash Vaidya and Purushottama Bilimoria

Introduction

The *central question* that this paper aims to lay a foundation for is:

> CQI: What can we learn about intuition, and how can our understanding of the purported phenomenon of knowledge by intuition be enhanced through a *cross-cultural philosophical investigation* of it?

One can gain a better understanding of the relevance of CQI by contrasting it with two distinct questions:

> EQI: What can we learn about intuition, and how can our understanding of the purported phenomenon of knowledge by intuition be enhanced through an *experimental investigation of it?*
>
> AQI: What can we learn about intuition, and how can our understanding of the purported phenomenon of knowledge by intuition be enhanced by an *analytic investigation of it*?

A. J. Vaidya
Center for Comparative Philosophy, San Jose State University,
1 Washington Square, San Jose, CA, USA

P. Bilimoria (✉)
Graduate Theological Union, Berkeley, CA, USA
e-mail: pbilimoria@gtu.edu

P. Bilimoria
University of California, Berkeley, CA, USA

P. Bilimoria
University of Melbourne, Melbourne, VIC, Australia

P. Bilimoria
Center for Dharma Studies, 2400 Ridge Road, Berkeley, CA, USA

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_3

Psychology, cognitive science, and experimental philosophy have provided a lot of engaging research on EQI.[1] Analytic philosophers and phenomenologists have provided a lot of engaging insight on AQI. Our plan here is to begin work on CQI. Our work will proceed by examining theories and uses of intuition across five different schools of Indian thought. We will use the term ***intuition*** to refer to the English terms ‘intuition’ and ‘intuitive’, as well as the Sanskrit terms *prajñā, yogaja pratyakṣa, pratibhā pramāṇa, ārṣajñāna*, and *siddhadarśana*. These Sanskrit terms are often translated as being in the semantic range of at least some of the prominent uses of ‘intuition’ and ‘intuitive’ in English.[2] The core use of *intuition* we will be engaging can roughly be captured as follows: *An *intuition* is a mental state that is an information-bearing awareness that is not the consequence of an explicit conscious inference, testimony, or sensory perception of one’s immediate environment*. The main sources we will engage are:

1. The Nyāya Theory
2. The Vaiśeṣika Theory
3. A Buddhist Theory
4. The Yoga Theory
5. The Mīmāṃsā Critique

This list of sources is not exhaustive of the possible sources one could engage in a general study of *intuition*. As one might imagine when considering EQI, there are numerous theories and uses of *intuition* in:

6. Moral Philosophy
7. Philosophical Methodology
8. Philosophy of Mathematics
9. Phenomenology
10. Cognitive Science
11. Psychology

and

12. Behavioral Economics

However, aside from a brief treatment of 6 and 8 for the purposes of the present work, we shall not be dealing with 7, 9–12. The motivation for our restriction of possible sources is primarily based on the fact that the most important recent work on *intuition* is Osbeck and Held’s, O&H, (2014) *Rational Intuition*. Their excellent work brings together important work across (6)–(12). We partly conceive our work here as a complement to their work by way of adding in information about (1)–(5), and then offering some comparisons between *intuition* talk in Indian philosophy and *intuition* talk in Western philosophy, especially with respect to

[1] For example see Alexander (2012) and Kahneman (2011). For discussion see Vaidya (2010).

[2] For example, I will be leaving out uses of ‘intuition’ on which the speaker means no more than what is conveyed by ‘having hunch’ or ‘making a guess’.

moral philosophy, the philosophy of mathematics, and both Yoga and Buddhist accounts of *intuition*.

We will close this introduction by making two sets of comments. The first set of comments will be about the recent history of studies of *intuition* from a comparative point of view. We will do this by providing some preliminary comparative commentary on O&H's *Introduction to Rational Intuition*. The purpose of this preliminary commentary will be to draw into focus some work by 20th century Indian philosophers on *intuition* that engages a core point that O&H draw out in their sketch of the history of *intuition* research in the 20th century. The second set of comments will serve as an orienting guide to classical Indian *pramāṇa* theory, within which one finds discussion of *intuition*.

**Intuition* in 20th Century Western Psychology and Philosophy and Indian Philosophy*

According to O&H's *Introduction* to *Rational Intuition* there is a stark contrast in how *intuition* talk was received in scientific circles at the beginning of the 20th century in comparison to contemporary discussions. Their main point is that in the beginning of the 20th century research by scientists into *intuition* was more or less frowned upon. By contrast, research on *intuition* is now growing in a number of scientific fields, such as economics and linguistics.

On their account, John Laird's *Introspection and Intuition* voices an important view about intuition that displays the early disdain toward *intuition* research. In this work Laird comments on *intuition* with respect to those that follow Bergson's philosophy. He says of Bergson's followers that:

> [They] believe that psychology is a science touched with the palsy of the intellect, and tarred with that practical brush which can never find use for truth, while intuition pertains to any metaphysics that understands itself, and consequently is beyond the scope of scientific psychology.
>
> (O&H 2014: 1).

Laird's comments suggest, *prima facie*, that *intuition* talk is beyond the scope of science. One might further unpack the unscientific nature of *intuition* talk at the beginning of the 20th century by drawing attention to the influence of logical positivism on the growth of psychology. On some accounts of logical positivism metaphysics is non-verifiable by definition, since the propositions that are in the domain of metaphysics admit of no method of verification. One might conjecture on the basis of this position the following argument:

1. Psychology is scientific.
2. Metaphysics is non-scientific.
3. Intuition is tied to metaphysical understanding.

4. So, intuition is not tied to scientific understanding.

The argument itself is suspect in many ways. For example, it could be that *intuition* operates in different ways but is important in both science and metaphysics. At least one way to argue for this position would be to defend the view that 'intuition' does not pick out a common kind of mental state. Nevertheless, the importance of the argument lies not in its soundness, but rather in the cultural pressure, as opposed to the rational pressure, the ideology behind it places on *intuition* talk.

What is interesting from a comparative philosophical point of view is that the view of *intuition* expressed by Bergson, and criticized by Laird, is echoed in the work of at least two important early 20th century Indian philosophers: Sri Aurobindo and S. Radhakrishnan. For example, Radhakrishnan associates *intuition* with the notion of "integral experience," which according to Hawley (2006) can be understood in three ways. First, intuition is integral in the sense that it coordinates and synthesizes all other experiences. Second, it is integral in that all other experiences are integrated into a unified whole. It is integral as it forms the basis of all other experiences. All experiences are at bottom *intuitional*. Third, it is integral in that it integrates one's experience into the life of the individual for social purposes and action.[3] But most importantly, Aurobindo and Radhakrishnan share a religious/mystical conception of *intuition* in experience. Thus, if science and religion are seen as opposites, then within at least one strand of 20th century Indian philosophy *intuition* is presented as being something that is outside of the scope of scientific investigation.

However, it is important to note that this is not the whole story. K. C. Bhattacharyya, another important and influential early 20th century Indian philosopher, does not appear to share the views of Aurobindo and Radhakrishnan. Bhattacharyya's conception of *intuition* derives from his conception of metaphysics. Metaphysics for Bhattacharyya is conceived of as being non-empirical and a priori. More importantly, it is not necessarily mystical and religious, like that of Aurobindo and Radhakrishnan. On one interpretation, Bhattacharyya's metaphysics would be construed as being Husserlian in nature—concerning the science of all sciences and the essences of all entities. On the Husserlian account of metaphysics, *intuition* would be thought of as a way of gaining evidence for the nature of entities as studied in metaphysics.[4]

[3] See M. Hawley (2006) for this characterization of 'integral' in the work of Radhakrishnan.

[4] This conception of K. C. Bhattacharyya is influenced by Mohanty's (1993b) reading of Bhattacharyya as a metaphysician especially with respect to his views on reflective experience and metaphysics. Mohanty glosses his thought as follows, "Reflection is an act of distinguishing, whose objective correlate is the distinct entity *qua* distinct. Space, time or self, which are objects of metaphysical knowledge, are all given in pre-reflective experience, but only as undistinguished from, and fused with the empirical world. It is the task of metaphysics to let them emerge in their distinctness and with their full autonomy (pg. 35)".

As a conjecture it is possible that the different conceptions of the role of *intuition* in human experience coming out of Indian philosophy in the 20th century came from two distinct pressures that are a function of how Indian philosophy was to be presented as India moved out of colonial rule by the British. On the one hand, there was a desire on the part of some, such as Radhakrishnan, also the first President of India, to present Indian philosophy as being somehow unique and different from Western science (Bilimoria 1995). On this front the attraction would have been to present *intuition* as a core part of Indian philosophy and associated closely with mystical experience and religious thought as opposed to Western science, which is said to be based on reason, logic, mathematics, and empirical evidence. There is some evidence, as will be seen, for this view to have been prevalent within classical Indian philosophy; however, it is not the only view of *intuition* to be found. On the other hand, there would also have been a desire on the part of some to make Indian philosophy somehow rigorous to Western minds by associating *intuition* with something familiar from mathematics and classical metaphysics. These two opposing streams have not had the same effect on Western receptions of classical Indian uses of *intuition*. Puligandla (1970) notes the tension in the discussion of the title of his paper, *Phenomenological Reduction and Yogic Meditation*.

> The title of this paper will certainly strike some readers as strange and especially those who naively believe that *it is a far cry from the Western rational philosophies to the Eastern mystical musings*. But those who are familiar with both know that the former are no more entirely rational than the latter are entirely mystical. (1970: 19, *emphasis added*)

His discussion is focused on a comparison between *intuition* in Husserl's phenomenology and *intuition* in Patañjali's *Yoga-Sūtras*. However, the points he makes about *intuition* talk in the *Yoga-Sūtras* and Husserl's discussion of transcendental phenomenology and the epoché are but only one locus for identifying similarities between Eastern and Western discussions of *intuition* talk. It is not only the Yoga school of classical Indian philosophy or the Advaita Vedānta school that has something to say about *intuition*. Rather all of the schools have something to say about *intuition*.

As O&H note in their introduction, *intuition* talk is often hard to tie down and move forward on because there are so many uses of the term. To ameliorate this difficulty and guide future research they provide a fascinating and extremely useful table of various uses of *intuition* in Western philosophy and other disciplines. Their table is not exhaustive of all uses of *intuition*, but it provides one with a strong foothold on some of the many different uses one can run across. Form a comparative point of view, it is important to point out, in contrast to the excellent table they provide, that Mohanty (1993a) also offers a table for contrasting differences between *intuitive* and *non-intuitive* knowledge.

Intuitive knowledge	Non-intuitive knowledge
1. Immediate awareness	1. Mediate knowledge
2. The object is given	2. The object is constructed
3. The knowledge is non-conceptual	3. The knowledge is conceptual
4. The knowledge has absolute certainty	4. The knowledge may have only relative certainty
5. The knowledge is concrete	5. The knowledge is abstract
6. The knowledge is of the unique individual	6. The knowledge is of the general
7. The knowledge is knowledge by identity	7. The knowledge is knowledge by difference
8. The knowledge is disinterested	8. The knowledge is motivated
9. The knowledge is ecstatic awareness	9. The knowledge is detached cold and intellectual

Mohanty deploys his table for the purposes of discussing different uses of *intuition* *cross-culturally*. For example, he notes that while Kant's conception of intuition would accept (1) and (2) under intuitive knowledge, it need not accept (5) and (6). By contrast, for a Buddhist thinker it is quite clear that what is known through *intuition* is the unique particular, which makes acceptance of (6) central. Later we will see how the Buddhist conception of *intuition* allows for the generation of a problem concerning *intuition* in relation to its proper objects that is similar to a problem found in Western discussions of *intuition* concerning mathematical and moral truths.

Finally, one might ask: why is there an absence of cross-cultural discussions of *intuition*? Perhaps the reason is that there has been a strong separation in 20th century Western philosophy between philosophy and religion. The strong separation between the two is largely due to the influence of logical positivism on 20th century Western philosophy. However, for the purposes of many philosophical topics, such as *intuition*, the separation between philosophy and religion has hindered potential growth in theorizing in much the same way that separating philosophy from science hinders growth in both philosophical and scientific theorizing. The experience of *intuition* is a phenomenon in the human condition. As a consequence, a comprehensive understanding of it must be generated through a reflective engagement across all areas of discourse in which it is treated.

Classical Indian Pramāṇa *Theory*

In classical Indian philosophy there are six orthodox schools and three heterodox schools. An orthodox school accepts the ultimate authority (*prāmāṇya)* of the sacred texts known as the Vedas (*Śruti*), and a heterodox school rejects the ultimate authority of the Vedas (Bilimoria 2008a: 20–21, 294–6). Alongside *Śruti*, the contingent authority of sometimes 5 (plus or minus 1), means of knowing (and arriving at cognitions (*jñāna*), understandings, and beliefs, including moral judgment), are widely accepted; namely, perception (*pratyakṣa)* (direct naïve cognition), inference (inductively deductive cognition) (*anumāna*), testimony (*śabda,* of which *Śruti* is the pinnacle), analogy *(upamāna*), and *arthāpatti* (counterfactual presumption), to which cognition of absence (*abhāva*) (or 'non-perception', *anupalabdhi*) is also added. All schools of Indian philosophy discuss a particular kind of 'extraordinary' mental state,—we might class under 'anomalous cognition', or 'trans-sensory perception', which is variously called in Sanskrit terms: *yogaja* (literally, 'born of yoga', in shortened form, '*yogī-*') *pratyakṣa*, *prajñā, pratibhā, ārṣjñāna*, or *siddhadarśana* (the sight of the yogically-accomplished adept, much like the uncanny 'occulted vision' of the mystic). The core debate, in this context, is over whether *yogaja-pratyakṣa* is a *pramāṇa*, either as a stand-alone (*sui generis*) means of knowing or as an extension of one of the above *pramāṇas*. It is usually aligned with perception and inference. In other words, the core question is: Is yogic perception a means of acquiring knowledge about something, which is substantially distinct from other sources of knowledge either in the kind of things known or the way of knowing? Two schools, the Mīmāṃsā and the Cārvāka argue that it is not.[5] The remaining seven argue that it is. However, some of the seven schools disagree over exactly how the mental state is an instrument of knowledge, what its fundamental nature is, and whether *yogaja pratyakṣa*, *ārṣjñāna*, *pratibhā,* and *siddhadarśana* should be thought of as being the same. Because of the immense literature on uses of *intuition* in classical Indian philosophy, we will focus our discussion on certain schools. And even when we are discussing certain schools we will only focus on specific figures within each school. Again, this work only aims to lay a foundation for future cross-cultural studies of *intuition*.

The Nyāya Theory

Within the eminent tradition of the Nyāya ['Reasoning'] School of philosophy, stretching from its founder Akṣapāda Gautama to members of the Navya-Nyāya, the so-called New School, such as Udayana, Gaṅgeśa, Viśvanātha and Jayanta Bhaṭṭa, there has been a great deal of discussion over a kind of perception called, *extraordinary perception*, EP.[6] There are at least two different understandings of EP: the *person*-based and the *universal*-based. On the *person*-based model of EP, a perception is said to be *extraordinary* because of the kinds of things that the perception is directed at and because of the nature of the kind of person that can have such a perception. On the *universal*-based model of EP, a perception is said to be *extraordinary* because the kind of thing one is related to is itself *extraordinary* in some way. Thus, the main contrast between the two models revolves around whether it is the person or the kind of thing the person is said to be related to that is extraordinary.

[5] See Das (2002): 419.

[6] Gautama*, Nyāyasūtra* (NS) (2.1.34: 497–8).

Yogaja-pratyakṣa on some accounts can be taken to be a multi-phase sensory perception that the yogin is capable of having. A yogin is also called—depending on the state of self-realization attained and as befits her transcendental stature—variously a *yukta* (one absorbed in continual *samādhi*-state), *viyukta* or *kevalin* (in a state of emptiness or bereft of all conceptual cognitive content).[7] So, in addition to cognizing particulars (entities, qualias, events in the ordinary order of things, with their respective universals imperceptibly inhering in and qualifying the object), the yogin has cultivated the higher capacity of perceiving those very *universals* as the property of 'sameness' *(sāmānya*), and a second-order *universal* of universals, which is knowledge of an even more special kind. There may even be a separate third-order perception that embeds a distinct knowledge of the categories of samenesses or modal universals (*dravya* substantives, propertied subsistences, and timeless events), unattached to any particulars, events or even classes. The knowledge of the infinite-expansive self, omniscience, liberation (*mokṣa)*, *summum bonum* or *niḥśreya,* would be four instances of this extraordinary *super-pramāṇa*.

To elaborate on the trope of the 'sameness of universals' a little further, we shall cite a couplet from the celebrated middle-Nyāya text, *Bhāṣā-Pariccheda*:

alaukikastu vyāpārastrividhaḥ parikīrtitaḥ
sāmānyalakṣaṇo jñānalakṣaṇo yogajasthā (BP **63)**

The text here speaks of three operational modalities, *vyāpāra,* or conjuncts (i.e. of the mind with its object of awareness), in the case of 'extraordinary' perception (of the unusual type), namely, i) ones based on common features *(sāmānya),* ii) those based on knowledge (*jñānalakṣaṇa)*, and those that arise from *yoga* (concentration). A word on the use of *vyāpāra* is apposite: in Gautama's time this was simply called '*sannikarṣa*', as in *indriyasannikarṣa* (sense-organ contact), but because mind *(manas*) is only tendentiously a sense-organ (sixth sense), technically the operational feature of 'conjunct' is preferred, and its fitting object is, in the first modality, *sāmānyalakṣaṇa*—*lakṣaṇa* being the 'structure of cognition'—*sameness* as a common generic feature, *prakaratā,* structuring the cognition. (Commentators, however, continued to use the term *sannikarṣa*). This common feature or characteristic (*prakāra*) may cut across—or be a pervasive, therefore common, feature of —a number of substantive occurrences. So it is not just *sāmānya* as *jāti* (e.g. natural kind universals or real universals in Lockean-Kripkean distinction forged over nominal universals), or smokeness in seeing smoke bellowing from wood-fired stove, with which the sense-organ has direct contact, but the association of *this* smokeness to *all* instances of smoke and smoky things generically, remembered, portended, or predicted, and otherwise. This is what is said to be critical here: this hypergeneric coordinate is the conjunct in the cognition: *sāmānyalakṣaṇa*; what one cognizes is the substantive sameness of the feature of 'smoke' across a number of instantiations. An analogy may be drawn with the 'type-token' distinction, in as much as, for example, a dollar coin can be substituted for a dollar note, or if soiled,

[7] *Bhāṣā-Pariccheda* (BP) 65: *yogajo dvividhaḥ prokto yukta-yuñjānabhedaḥ.*

yet another dollar note, because what each has in common is the same *value* (hence the class or 'type') of a dollar in the nation's assigned currency. The tenderer does not have a direct sense-connection with the 'value' but this is inherently there (*samavāya)* in the transactional act.

The second modality is with respect to the decisive *knowledge* (*jñāna*) of that 'sameness' as being possessed by such-and-such a substantive (smokeness to smoke); the text seems to be keen on emphasizing that there is a separate connection that is being made between seeing smokeness (as feature) across numerous instances of smokeness and the second-order *knowledge* that these are features of 'smoke'; otherwise there would be no connection of 'sameness' (as ubiquitousness) in seeing smoke in the kitchen and seeing smoke at a distance when there is fire on the mountain; so this is a special conjunct the mind achieves with respect to something as simple as the presence of smoke in place *k* and then again in place *m*. Nyāya considers this function of the mind that knows smoke by knowing the connection based on common feature to be somehow 'extra-ordinary'. Nothing here is said about the knowledge of the *object* as such that possesses smoke, for recall the definition is directed at working through the limits of *laksaṇa,* i.e. the phenomenology of the cognition of the universals.

The third operational modality is with respect to *yoga*, where by concentration (aided by meditation), the accomplished yogi is able to have (come to possess) knowledge of every sameness, hence universal, universal of universals, and everything that could possibly or modally be connected with these universals; in fact a highly attained or enlightenment yogi is said to be always connected with the *knowledge* of substantives in all possible worlds, and thereby with everything that there is to know: *yuktasya sarvadā bhānam, chintāsahakṛto'paraḥ* (BP 66). The *yukta* doesn't have to know each and every individual thing (which is not a requirement of omniscience so understood); and furthermore, it follows that the last part of the claim here blocks the skeptic's doubt or question: how does one *know* that one knows everything? It suffices that there is virtual connection through the knowledge of universals, and the overarching universals, the *archē*, etc.

Thus, to summarize the contrast: on the *universal*-based model of EP, a perception is said to be *extraordinary* because it involves an ordinary sensory connection to something, a universal, which is extraordinary, because when one is appropriately connected to the universal by a sensory connection, they are through the nature of universals, also connected to all prior and future instances of the universal. *Sāmānyalakṣaṇapratyakṣa* means *universal*-based sensory connection. By contrast, as Das (2002) characterizes the *person*-based model we get the following:

> The Naiyāyyikas hold that the supernormal perception of an individual, i.e., a *yogin* is also as real as any other perception. They call such a perception a supernormal one, for such perceptions are beyond the range of normal perception. They can perceive the subtle objects, atoms, and minds of others, air space, time, etc. through this perception. Jayanta Bhaṭṭa[8] describes yogic perception as the perception of subtle, hidden, remote, past, and

[8] NM: 95.

> future objects and considers it to be the highest excellence of human perception. And he rejoins that *yogins* perceive all objects in all places through cognition simultaneously. The supernormal state of mind acts as the supernormal sense-object contact *(alaukika sannikarṣa*). This type of contact is known as *yogaja sannikarṣa* which causes *yogaja pratyakṣa*. (Das 2002: 419–420).

The core account of intuition under the *person*-based model of EP is that intuition is a form of supernormal perception, a kind of perception where one's normal perceptual capacities are enhanced so as to allow one to intuit the past and the future, subtle things, elusive things, imperceptible (*adṛṣṭa)* traces of entities and events receded in time, and even remote entities or events (mapped or divinized non-inferentially through portending traces). Of course, one might simply ask: how can the mind make contact with future objects or events, decidedly elusive, i.e. of the *adṛṣṭa* category, so as to have supernormal perception of them? Here it is interesting to note that the Nyāya do not hold that *intuition* requires *contact*, rather they hold that the supernormal overall state of the mind is sufficient to generate the intuition. In normal perception the sense comes into contact (*indriya-sannikarṣa*) with the objects that are thereby what is perceived by the knower. However, in *intuition*, it is because the mind is in a *supernormal* state that it can deliver *intuitions* that have elements that are (i) about the past, (ii) about the future, (iii) about entities that are remote in space, (iv) entities that are very subtle, like air, and perhaps even (v) partially occluded or hidden. The form of *contact* here is called *samyukta-samyoga-sannikarṣa*:[9] a presentification (literally, transcendental *conjunct*[10] of the *noesis* with its intentional sameness, the *noeta*, outside, as we described in the instance of the grasping of universals above).

In that sense, one could say that *yogaja-pratyakṣa* embeds a trans-sensory cognition, for the mind (*manas*) that is said to be *vibhu* or extensionally pervasive and which comes directly into contact with or has phenomenal *access* to an elusive object (such as a receded entity or a yet-to-be event, the universals (*jāti* or *ākṛti*) embedding or inhering in a particular, and its sameness or simulation to the class of universals (*sāmānya*) to which this belongs in a higher order *universal*, *sāmānya-lakṣaṇa*) which normally, and normatively, falls outside the range and scope of the extended senses and the deducing mind.[11] This implies that the mind extends

[9] SM: 63.

[10] *āsattirāśrayānāṃ tu sāmānyajñānamiṣyate*
tadindriyaja-taddharmabodhasāmgriyapekṣyate (BP 64)

Here it said that the *awareness* of the generic sameness structure is identified as the conjunct *(āsatti, pratyāsatti)* with the support-base (substratums) to which the particulars are associated. The complete commeasurement involved in the perception correlative to the *indriya*, sense instrument, is the unmitigated condition. (That is, the eye, the radiance, the mind, generic features, and contact, etc., must all be involved in this awareness-generation as well, to rule out any possibility of simple abstractions and conceptual elopements).

[11] BP 65 *Viṣayi yasya tasyaiva vyāpārao jñānalakṣaṇaḥ.* (also SM 64, p. 342) This verse underscores the facticity of the knowledge of the specific, unique and unusual universal as the transacting connection in the cognitive episode with its object cognized and via this connection *mutatis mutandis* knowledge of all object-substrata that possess this universal. A question is discussed in

beyond its ordinary capabilities to reach out in multiphase to regions of the world not accessible to the regular functions of the senses. What is being suggested is that the so-cultivated mind as a 'sixth + sense', or *sensus communis* (in Aristotelean terms), takes over and extends in time and space to categories of understanding that exceed the epistemic mediating-bounds of the senses and registers a knowing (*jñānagrahaka)* in the non-constraining epistemic environment. This, in brief, is then the phenomenology of *yogaja-pratyakṣa.* Hence, in this regard *yogaja-pratyakṣa* is both extraordinary and trans-sensory, or even 'extra-sensorial'; Stephen Phillips elsewhere has christened this uniquely peculiar transcendental 'a/perception' (*alaukika-cakṣu)* or 'extra-extrospection' following Matilal's 'mystical empiricalism.'[12]

Chakrabarti (2010), extending the cognition to *anumāna* or inference—in as much as inference embeds perception, and may implicate yogi-derived percepts—offers another rendering of the *universal*-based account of EP. In order to understand the *universal*-based account it is instructive to consider how one could be justified in believing the conclusion of the following argument, called SH:

> There is a fire on the hill over there; because I can see smoke above the hill over there; and wherever there is smoke there is fire, such as when I am in my kitchen cooking.

The conclusion of this argument is *that there is a fire on the hill over there*. The core premises are: (i) *I can see smoke above the hill over there* and (ii) *wherever there is smoke, there is fire*. However, while it is clear that one can use perception to gain knowledge of the presence of smoke above a hill, which is stated in premise (i), one must, in general, ask: how can one know (ii) that wherever there is smoke there is fire? The Nyāya maintain that the only way one can know such a claim is through *extraordinary perception*. Their reason for doing so is that the truth of such a claim requires grasping the universal *fire* and the universal *smoke* and understanding the special relation (*vyāpti*) between them, or the connectivity of aligned universals. In general one cannot infer from a finite set of observations of the absence of fire and the absence of smoke, and the presence of fire and the presence of smoke, that wherever there is smoke there is fire. A finite sample of co-variation

the commentaries: but how can you say such one knows all the smokes and fires, when these are not there; and is he therefore omniscient? The answer is smokes and fires do not have to be eternally present (somethings do), and what is known is not in any great detail, so no claims to omniscient in this condition is being emphasized. There are two further steps before this claim is possible, as described earlier.

[12] Phillips 1996: 175–8, Bilimoria (2011), but Matilal did not use this appellation as en endorsement but rather as a caricature of the position; he was a through-and-through realist and argued for the inclusion of universals within the operational features of ordinary perception (consistent with his direct realism thesis); in that regard Matial's non-nominalist view is the same as Jayanta Bhaṭṭa's on the direct perception of universals, but misses the further thesis of universals of universals, and unattached *sāmānya* (such as God's supreme knowledge and his over-arching bliss-state, *ānanda*). See Matilal *Perception*, p. 424 on ultimate real universals and their assimilation; while for Kant universals are known a priori; for Aristotle they are grounded in the physical, in Nyāya it is mixed up by a relation of inherence (*samavāya*).

of absence and presence of *x* and *y* cannot provide justification for the universal claim. On the *universal*-based model of EP the following occurs:

1. S has an ordinary sensory perception of a particular P.
2. When S has an ordinary sensory perception of a particular P, they also have an ordinary sensory perception of a universal U present in P. For example, when Renu perceives a cow before her in the pasture, Renu has a sensory connection through her ordinary perception of the universal *cowness* present in the cow before her.
3. A universal that is wholly present in a particular P has an extraordinary property: *what one comes to know of it in a particular* extends to all instances of the universal, past, present, and future.
4. So, by (1)–(3) S can have an extraordinary perception of what is true of all of the instances of a universal U simply by having an ordinary sensory perception of a particular P in which U is present.

We might further understand this kind of perception by looking at two points about it. First, does Nyāya philosophy take EP on the *universal*-based model to be a regular kind of perception stemming from the six features of mind it recognizes, the five senses and the unified *manas*? Chakrabarti notes that the answer is 'yes', and it might be further noted that in giving this answer about the nature of EP, it follows that extraordinary perception on the *universal*-based model is not the function of an *extra* sensory capacity, that is a capacity in addition to the six recognized sense functions. Second, it appears that *universal*-based perception is *not* the only kind of *extraordinary perception* for the Nyāya. In addition to *universal*-based perception, there is also perception of absence, or negative entities. Whenever one has a perception of absence, one has an extraordinary perception. Thus, the category of 'extraordinary perception' is investigated by the Nyāya in general.

The Vaiśeṣika Theory

Praśastapāda is one of the core contributors to the Vaiśeṣika School of Philosophy. *Ārṣjñāna* (*ṛṣi*-cognition) is one of the four kinds of *vidyā* (knowledge), some Indian philosophers treat it as being a state that is similar to *yogi-pratyakṣa* (yogic perception) and *Siddhadarśana* (siddhic vision). Sjödin (2012) provides a delineation and discussion of Praśastapāda's account of the distinction between the three states in his *Praśastapādabhāṣya*:

Yogī-pratyakṣa: Y-cognition

> But for the yogis, different from us, who are *yukta*, arises through the inner sense assisted by merit born from yoga, a correct vision of the own nature of their own self, [the self] of others, *ākāśa*, space, time, wind, atoms, inner sense [and] the qualities, motions, generalities, [and] particularities inherent in these [substances] and of inherence itself. For the one who are *viyukta* then, arises perception of the subtle, concealed and remote, by means of the fourfold contact when assisted by merits born from yoga. (Sjödin 2012: 473)

Siddhadarśana: S-cognition

> Siddhic vision is not a distinct (i.e. another) cognition. Why?
>
> This vision, which is preceded by effort [and] concerns subtle, concealed and remote objects visible to seers who are accomplished in [the practice of] eye and feet-ointment, the sword and globule, is just perception. Furthermore, the [distinctness of the] valid vision of matured merit and demerit of sentient beings in heaven, atmosphere and on earth, [being] grounded in the movement of the planets and stars, is just inferential. Furthermore, the [distinctness of the] valid vision of merit etc. [which is] independent of an inferential mark, is just included in either perception or *ṛṣi* cognition. (Sjödin 2012: 474)

Ārṣajñāna: A-cognition

> For the *ṛṣis*, the ones who arrange the transmitted, arises a cognition which is a presentation of the object as it is and which is appearing. [The cognition] arises from a contact between self and inner sense and from specific merit. [The cognition] is of past, future and present objects beyond the senses, like merit etc., [and of objects] discussed and not discussed in texts. This [cognition] is said to be "*ṛṣic*". Though this generally [occurs] for heavenly *ṛṣis* [it occurs] sometimes for worldly beings as well. Like in the case of a girl who says: -My heart tells me that my brother will come tomorrow. (Sjödin 2012: 477)

According to Sjödin's account of Praśastapāda theory, S-cognitions are not *distinct* from Y-cognitions because they are simply a form of *perception*. A-cognitions, by contrast, are distinct from Y-cognitions and S-cognitions because (i) A-cognitions involve the apprehension and presentation of an object as it is, (ii) they arise because of a peculiar merit on the part of the subject of the cognition, and (iii) they involve contact between the (*manas*) mind and the (*atman*) the self. On this account an *intuition* is a presentation of an object as it is due to a contact between the mind of the subject and the self of the subject that is a product of some kind of merit on the part of the subject. The merit comes from a practice that improves one's capacity to have A-cognitions. The notion of merit is not the kind of merit that is innate or due to a person's heritage. Rather, just as Y-cognition is a function of yogic practices, A-cognition is a function of a practice as well. It is merit that is a contributing cause to the production of an A-cognition. The merit derives from a practice that involves some components of yogic practice, but not all of them. In addition, it is important to note that A-cognitions are a form of *prātibha*, which means "shine forth", "shine upon", "come in sight", "appear to", and "burst forward". They have a strong presentational phenomenology. The word is typically translated as, "an instantaneous flash of insight or intuition". And by some, such as Bhartṛhari, it is articulated as the immediate understanding of sentences in one's own language whereby something is presented before the self as being *self-evident*.[13]

Finally, concerning *ṛṣi-cognition* the key questions are: is it taken to be a form of knowledge? Is it a form of knowledge that is reducible to another form of knowledge, such as perception, inference, or testimony? The answers to these

[13] See Sjödin (2012: 479–481) for discussion of these points; Bilimoria on *sphoṭa-pratibhā* in Bhartṛhari's linguistics (2008a: 18, 63, 96–8, 308).

questions are not uniform. However, it is clear that at least some philosophers hold that *ṛṣi-cognition* is a distinct category of knowledge, which is not generated from either a process of perception, a process of inference or verbal testimony. Moreover, it is a distinct kind of *pramāṇa*. The Profile for A-cognition is the following:

1. A-cognitions are caused by a merit that is not identical to yogic practice.
2. A-cognitions are not sensual/perceptual because there is no contact between the sense organs and the relevant object.
3. The experience of A-cognition is non-volitional. The subject does not try to have an A-cognition.

A Buddhist Theory

Dharmakīrti is one of the founding members of the Dignāga-Dharmakīrti school of Buddhist philosophy. He discusses *yogaja- pratyakṣa* in his *Pramāṇavārttika* (PV), the chapter on perception (PV3), the *Pramāṇaviniścaya* (PVin), and the *Nyāya-bindu*. J. Dunne (2006) offers the following characterization of the central components of Dharmakīrti's account:

1. A yogic perception is a cognition induced by a meditative practice (*bhāvanā*) (PV3.281; PVin1.28). The types of practice in question are ones that build to a "culmination" (*parinispatti*) (PV3.285 ≈ PVin 1.31). Specifically, these practices begin with learning about some object or idea, then contemplating it in a manner that involves reasoning; finally, one engages in the meditative practice itself, and when that practice reaches its culmination, a yogic perception will result (PVin *ad* 1.28).
2. The cognition that results from this type of process is vivid or clear (PV3.281 and PVin1.28 and 31); that is, the object appears with the same degree of vividness that accompanies cognitions involving sensory contact, as when an object is directly in front of one (PV3.282 = PVin1.29). This is indicated by the fact that, when persons have this type of cognition they react in an alert or excited manner that is absent when they believe themselves to be simply inferring or thinking of something that they do not take to be directly present (PVin1.30).
3. A yogic perception is similar to cognitions that occur when, for example, a person overtaken by grief repeatedly thinks of the departed person and eventually hallucinates that person's presence, or when an adept visualizes a colored disc and eventually sees it with complete vividness (PV3.282 = PVin1.29).
4. All cognitions of this kind—whether induced by meditation or by states such as grief—appear vividly; therefore they are not conceptual, since a conceptual cognition cannot present its content vividly (PV3.284ab = PVin1.32ab).
5. Although a yogic perception is induced by a process similar to hallucination, it is distinct from hallucinatory cognitions because the object of yogic perception

is "true" or "real" (*bhūta/sadbhūta*), whereas hallucinations have "false" or "unreal" objects (*abhūta/asadbhūta*). The only specific yogic objects mentioned are the Noble Truths (as is strongly implied by PV3.281 and 285, and is explicitly stated in PVin *ad* 1.28).

6. A yogic perception is trustworthy (*saṃvādi*), and it is a reliable cognition (*pramāṇa*) (PV3.286).

Dunne goes on to point out that Dharmakīrti's comparison of yogic perception with hallucination is intended to show that on Dharmakīrti's theory yogic perception should not be thought of as some kind of *mystical experience*. It is not presented as a mystical experience, and it is argued to be non-analogous with mystical experience. Rather, the process is, "designed to inculcate transformative concepts into the mind through an intense, vivid and *non-conceptual* experience that arises from learning, contemplating and meditating on those concepts Dunne (2006: 499)."

A core component of the Buddhist view that puts it in strong contrast with the Nyāya view is the disagreement between the two schools over the status of universals. The Buddhist denies that universals are ultimately real. They deny this on the following grounds: only what is causally efficacious really exists, universals, unlike particulars, are not causally efficacious, since universals cannot change; so universals cannot really exist. The Nyāya by contrast holds that universals truly do exist in the objects that we have a causal connection to, and that by coming into contact with them it is possible for us to have certain kinds of *extraordinary perception*. It is clear from the disagreement between the two schools on universals that the underlying theory of *yogaja-pratyakṣa* cannot be the same.

An important consequence of the Buddhist view of universals, according to Dunne, is that if yogic perception is a real kind of perception, then the objects it engages cannot be universals but must be particulars, since only the latter are ultimately real. However, as Dunne points out, Dharmakīrti does not delimit the scope of yogic perception to particulars. Rather, he opens it up it to universals, such as impermanence (*anityatā*) and emptiness (*śūnyatā*), as well as the to realization of the *four noble truths*.

Concerning yogic perception Dharmakīrti says that it is:

> A trustworthy awareness that appears vividly by the force of meditation – similar to cases such as the fear [induced by something seen in a dream] –is a perception; it is non-conceptual [PVin 1.28] (Dunne 2006: 507).

> [Some] adepts, having apprehended objects (*artha*) through cognition (*jñāna*) born of learning, and having established those objects through reason and a cognition born of contemplation, then meditatively cultivate [a realization of] those objects. When that meditation reaches its culmination, those adepts have a cognition with a vivid appearance, as in the case of fear [induced by a dream]. The adept's cognition is a perception that is a reliable awareness *(pramāṇa*); it is nonconceptual and has a non-erroneous object. That reliable perception is, for example, the seeing of the Noble Truths *(āryasatyadarśana*), as I explained in the *Pramāṇavārttika*. (Dunne 2006: 507).

Finally, as Dunne does, it will be useful to contrast Dharmakīrti's account of yogic perception with the view expressed by Vasubandu on grasping the Four Noble Truths. Dunne provides two important passages from Vasubandu:

> One who wishes to see the Truths from the beginning guards his ethical conduct. He then studies the teachings (*śruta*) that are conducive to seeing the Truths (*satyadarśana*), or he listens to [teachings about their meaning]. Having studied or listened, he contemplates. And having correctly contemplated, he applies himself to meditative cultivation. In a state of meditative concentration *(samādhi*), in him arises the contemplation-born discernment on the basis of his study-born discernment. And on the basis of his contemplation-born discernment, the cultivation born discernment arises in him. (Dunne 2006: 508).
>
> \- - - - - - - - -
>
> The study-born [discernment] is a definitive determination (*niścaya*) that arises from the reliability of a trusted person's statements (*āptavaca-naprāmāṇyajāta*). The contemplation-born arises from meditative concentration (*samādhija*)… (Dunne 2006: 508).

The general idea advanced by both thinkers is that yogic perception is the consequence of progression. The progression begins with a linguistically derived conceptual understanding, which is followed by a rationally derived conceptual understanding; the final culmination is a meditatively induced non-conceptual state that is vivid. However, to critically examine this state we might legitimately ask, what could this state be about, given the standard Buddhist rejection of universals?

To see a potential problem consider the Truth of Suffering that is an explicit object of yogic perception for Dharmakīrti, as well as other Buddhist thinkers.

1. To realize the Truth of Suffering, one must realize the impermanence of everything, since the impermanence of everything is part of what constitutes the Truth of Suffering by being a cause of suffering for each thing that does suffer.
2. The impermanence of everything is not something over and above all things that are particular and impermanent. There is no real universal of impermanence, which everything participates in. Rather, impermanence is abstracted from the particular impermanence that each and every thing undergoes.
3. Yogic perception, being a perception, is only of particular things that can be causally efficacious in the production of an image in the mind.
4. So, it cannot be that in having a yogic perception of the Truth of Suffering one is put into contact with the universal *impermanence*.

The problem that Dunne draws out here is extremely important from a cross-cultural point of view. In **our comparative exmination** we will show how this problem within Buddhist epistemology and metaphysics concerning *intuition* and its objects can be brought into contact with a well known problem in the philosophy of mathematics that extends to theories of how we can know moral truths.

The Yoga Theory

Patañjali is considered the founder of the Yoga School of Philosophy. His *Yoga-Sūtra* is considered the central text of the Yoga School. Puligandla (1970) describes Patanjali's view of intuition as follows:

> The three stages, *dhāraṇā', dhyāna'*, and *samādhi*, taken together constitute what Patañjali calls the *saṁyama*. According to [him], at the *samādhi* state the subject is freed from the brain-bound intellect and acquires intuition, known as *buddhi* or *prajñā*. It is through this intuition that the yogi grasps the subtler and profounder aspects of objects in the manifested universe.
>
> *Saṁyama* can be preformed on any object whatever and knowledge of it at different levels can be obtained. Thus Patañjali classifies knowledge as *śabda*, *artha*, and *jñāna*. *Śabda* is knowledge based on words alone. *Artha* is the knowledge which the yogi seeks, the true knowledge of any object whatever as grasped by intuition in the *samādhi* state. *Jñāna* is knowledge based on perception and reasoning, under which come all empirical sciences. Patañjali also distinguishes between *savitarka* and *nirvitarka samādhi* stages. In the former, the separation of knowledge into the above three kinds takes place; in the latter, which is the culmination of the *saṁyama*, the pure, real, internal knowledge regarding the object is obtained and the yogi then knows the real object by making the mind one with it. (1970: 25).
>
> The knowledge obtained through yogic meditation is not to be confused with ordinary kinds of knowledge, for instance, common sense and scientific knowledge. The latter are always based on pre-suppositions which cannot be validated within the disciplines themselves. Thus, Patañjali says that "The knowledge based on inference or testimony is different from direct knowledge obtained in the higher states of consciousness because it (the former) is confined to a particular object or aspect. (1970: 25)

To unpack the theory of intuition that Patañjali offers one must look carefully, as Puligandla does, at the notions of *dhāraṇā', dhyāna'*, and *samādhi* that constitute *saṁyama*. Yogic meditation as a source of intuition requires:[14]

1. Engaging in certain physical and mental practices known as the five *aṅgas* of yoga. The first two *yama*, *niyama,* are intended to eliminate distractions arising from uncontrolled desires and emotions. The second two, *āsana*, *prāṇāyāma*, are intended to eliminate disturbances arising from the physical body. The last, *pratyāhāra*, aims to prepare the mind for concentration by isolating the sense organs from the mind.
2. Engaging in certain meditative practices that prepare the mind for having a genuine intuition. *Dhāraṇā* is concentration. For Patañjali, "concentration is the confining of the mind within a limited mental area." In the dhārṇā stage the aim is to keep the mind continuously engaged in the consideration of one object, and to bring it back to that object if it wanders. *Dhyāna* is the uninterrupted flow (of the mind) toward the object (chosen for meditation). It is contemplation of an entity. This stage is reached only when a practitioner can hold their mind on a single object without any fluctuation. *Samādhi* occurs when there is only

[14] See Puligandla (1970: pp. 22–26).

consciousness of the object of meditation and not of the mind itself. The point of *samādhi* as a distinct state is that it enables one to remove a final distraction in the contemplation of an object: awareness of the self. In *dhyāna* one has complete concentration on the object, but one also has awareness of the self. In *samādhi* one removes awareness of the self.

3. According to Patañjali every manifestation has two forms: *rūpa* and *svarūpa*. The first is the inessential form and the latter is the essential form. In the transition from *dhāraṇā* to *samādhi* the *svarūpa* moves from being present in the background of one's consideration to disappearing completely as one's concentration on the object of contemplation increases.
4. In *samādhi* there is a fusion of subject and object. The fusion can be compared to that of the experience of flow when dancing. The dancer-dance distinction drops out for the dancer. Similarly, the contemplation of the object brought upon by the concentration of the subject leads to a loss of an awareness of the self in the concentration.

A most spectacular exemplification of the kind of luminously heightened perception that an adept steeped in yoga is capable of having, through her awakened yogic-epistemic capacities as per the *Yoga-sūtras,* occurs in (at least) two Books of the grand epic of the *Mahābhārata*. The first is reported in the celebrated *Bhagavadgītā* (BhG), where after hearing Krishna's detailed theory on the metaphysics of yoga (of various kinds), his bewildered warrior-friend Arjuna is moved to ask: what would one come to know at end of the practice of the yogas? Since it would be premature to presume that the amateur Arjuna could already 'be there', Krishna grants him an unusual gift: the momentary boon of *divyacakṣu*, 'divine vision', so that he could have a fore-taste at least of the immense knowledge the yogic process is capable of unleashing.[15] There follows an account of an epiphenous experience that Arjuna has wherein he reports seeing a thousand suns fulminating in intense radiance; gloriously encircling planets, galaxies and universes; bursting forth of energies and light-rays whose colors and playful swirling in curved spaces defy ordinary linear experience; *time* that stretches across infinite and parallel and spherical ranges, and renders all beings dead and extinct in eons ahead, but as if in the next moment, and more. Arjuna is so overwhelmed by this magnificent vision, unable to bear or understand its intensity fully, that he asks Krishna to return him to ordinary everyday perception. Krishna here would be identified as the super-*yukta*, and Arjuna as the aspiring-practicing yogi (were he to take a leaf from this theophanic experience and undertake upon himself the enjoined praxis).

The second episode occurs in the Book of Women following the 'Dead of Night' assault on the battlefield by one side of the cousin-brothers upon the other claimants to the sovereignty of the kingdom, which leads to a most horrific carnage, of the brutal deaths of the valiant commanders and fighters on both side of the warring clan. In the morning, the women, mothers, wives, daughters, mother-in-laws, sisters, and female mendicants arrive at the unsettling scene. The grand matriarch,

[15] Chapter 11, *The Bhagavadgītā.in the Mahābhārata.*

Gāndharī', is endowed on spot by Krishna with a 'moral intuition', an extraordinary vision that is enabled by certain divinely-endowed yogic-eyes (*divyena cakṣuṣā),* so that she could survey without breaking down in intolerable grief the full extent of the carnage that now defines the battlefield, and hence also gives expression to the moral improbity of the situation juxtaposed with the shocking grief of the women-in-tow. The description she is able to provide is a masterpiece of the work of pathos and empathy: 'Look at the array of widows, bewildered daughter-in-laws, newly-betrothed brides running hither and thither, with their braided hair down, soaking in the blood of their loved ones, some also looking for the heads severed from their fallen husbands. The jackals are out in daylight indifferent to this human noise, gnawing at every limb which only a few moon-nights before in deep conjugal embrace triggered many a pleasurable sensation to their beloved now distraught wives, screeching to the winds: How could this be—this pitiful slaughter? *Whose dharma, whose justice*'[16]

The Mīmāṃsā Critique

The Mīmāṃsā School of philosophy, like the Cārvāka, does not accept *yogaja-pratyakṣa* as a form of knowledge. However, the Cārvāka School only accepts normal sensory perception. They deny testimony, inference, memory, and all other commonly discussed potential *pramāṇa*. By contrast, the Mīmāṃsā are more liberal in the sources they accept. Perception, testimony, and inference are all acceptable, while memory and yogic perception are not. At least one argument they offer against yogic perception is, what we will call the exclusion-by-reduction argument.

Exclusion-by-Reduction Argument:

1. Yogic perception is an intuition that is the product of a sustained practice of meditation. The intuition that is produced through a sustained practice of meditation is a presentational flash of insight that is information bearing.
2. The information presented by an intuition either makes reference *only* to an event in the past that involved perception or testimony about something or the intuition presents itself as being about something *more than* that which has occurred in the past. What need is there for intuition?
3. If it apprehends something that is just about the past, then it is not distinct from what is found in memory. And since memory is invalid, intuition is invalid.
4. If it apprehends something more than that which was perceived in the past, then it is illusory, since it apprehends something that is non-existent.

So, yogic intuition is invalid either because it reduces to memory of prior knowledge or because what it purports to be about is illusory. Das (2002) notes that

[16] MhB. Clay Sanskrit Edition, Strīparvan: 281.

there are three additional reasons why the Mīmāṃsā do not recognize yogic perception as a source of knowledge.[17]

1. Sense organs by their nature have limitations. While it is true that the power of the sense organs can be increased by practice, there would appear to be a limit to what they can access.
2. While it is true that the power of a particular sense faculty can be increased, it is not true that a yogin that practices can see everything with his eyes. For example, his eyes cannot reveal sound nor his ears reveal color.
3. Although a person can possess a superior power of vision, the superior power of vision can only be applied to visible objects. *Dharma* is not visible, and is only knowable through the study of the Vedic texts. Thus, it cannot be the object of perception.

The 8th century doyen of the Mīmāṃsā, Kumārila Bhaṭṭa, was quite adamant that all perceptions involve a particular kind of contact, which is proper only to that perception, and he called this operational feature *saṃprayoga* in contrast to the *sannikarṣa* of the Nyāya. Indeed, he averred that 'contact' might not be the correct description of what transpires when an object comes into the vicinity of a sense-organ, but more like what the Buddhist protagonists also have in mind: an operational presentation at a distance. The brunt of the argument is that there cannot be perception without the *saṁyoga* of the sense-organ with its proper object given in its field: the mind, which by extension is a sense-instrument, is not in a position to, or has the capability of circumventing this process and 'grasping' the object *(jñānagrahak)* through some other, altered, state of consciousness. If that were the case, then inference and testimony would also involve this inexplicable perceptual knowledge, which would absurdly render inference and testimony otiose and reductively redundant. There would then be no need of the other *pramāṇas* as the so-called *yogi-pratyakṣa* or extraordinary perception would yield knowledge of matters esoteric (such as dharma or moral imperatives, and the exact size of the *apūrva* or deferred imperceptible *potentia* resulting from the māntric-effect of sacrifices and proportionate fruits (*phalas*) that can be expected).[18] This would go against the grain of the received tradition and the indispensability of *Śruti* in such transcendental matters. While he agrees with the Nyāya view that universals are perceived, he does not believe universals bear an 'inherence' relation (*samavāya)* to their objects but are rather identical (*tādātmya)* with them, and so he argues that what the Nyāya call (the further epistemic step of) *sāmānya,* 'sameness' is not something whose knowing requires a supernormal capacity but is instead a matter of inference: by observing smoke in its various occurrences (*k, l, m,* etc.) we infer that it is the same smokeness that pervades each of these instances. This is

[17] See Das (2002: 422). Das is of course, summarizing a rather barbed polemical discussion that the doyen of Mīmāṃsā, Kumārila Bhaṭṭa presents in his eminent work, notably the *Ślokavārtika* (Pratyakṣasūtram I, 53, 63–111).

[18] See Bilimoria (2014).

consistent with common sense understanding. Kumārila even denies that the understanding of the concomitant relation between two universals that is drawn upon in inference—the *vyāpti*—is a product of supersensuous or *extraordinary* perception, but rather a generalized conjunct of two perceived instantiations of a general (natural) kind (for which *ākṛti* rather than *jāti* is preferred), thus:

$$\text{smoke} \rightarrow \text{smokeness} + \text{fireness} \leftarrow \text{fire}$$

It follows that there is nothing elusively mysterious or mystical about perception of particulars or of composite perceptions (in inference), accumulative perceptions (in testimony), inverse counterfactuals (in presumption) and perception of absence (*abhāva*).[19]

Comparative and Constructive Commentary

It is now time to return to the central question that this essay began with:

> CQI: What can we learn about intuition, and how can our understanding of the purported phenomenon of knowledge by intuition be enhanced through a *cross-cultural philosophical investigation* of it?

We believe that there are two kinds of things one can learn from a cross-cultural investigation. On the one hand, one can learn what similarities and dissimilarities there are between different theories of *intuition* as found in different cultures. On the other hand, one can discover how a theory from one tradition might suffer from a cognitive blind spot by looking at other traditions. Our hope is that the cross-cultural-constructive engagement we will provide lays the foundation for future comparative studies. As an entry point into the discussion we will present a core question and a set core dimension questions for theorizing comparatively about uses of *intuition*, as well as a brief explanation of them. From that point forward we will move into an examination of each with respect to the literature on moral philosophy, the philosophy of mathematics, and/or classical Indian philosophy.

The core question that drives a cross-cultural and multi-disciplinary investigation into *intuition* is the *common kind question*.

> CK: Is there a common kind of experience that falls under the various uses and theories of *intuition* found in (1)–(5).[20]

The answer to CK itself depends on how a theorist aims to identify two mental states as falling under a common kind. There are at least two approaches one can take. The *phenomenological approach* maintains that two mental states fall under a common kind when and only when they share a common type of phenomenology at

[19] See Bilimoria 2008b, Part II *Abhāva,* and *Anupalabdhi.*

[20] CK can be expanded so that the question is about uses found outside the present essay.

the appropriate level of description. For example, a pain in my toe and a pain in my finger, though being distinct, both fall under the kind *pain* in virtue of the fact that at the appropriate level of description, they share a sufficiently similar phenomenology. While the locations of the two pains are different, and their intensity is different, their phenomenological similarity is strong. The *teleological approach* maintains that two mental states fall under a common kind when and only when they are the output of a common type of process. For example, an auditory perception and a visual perception, though being products of distinct processes, both fall under the kind *perceptual process* in virtue of the fact that at the appropriate level of description the processes that underlie them have a shared purpose: the acquisition of information. This shared, higher-type level, process can be contrasted against another process, such as making a decision about what to do, under which decision-making or choosing what to do would fall. It is an open question whether or not phenomenological ways of individuating mental states track teleological ways of individuating those same states. It is also a further question how the phenomenological and teleological approaches relate to the physical states that realize/cause the process and deliver the phenomenology.

A *common kind* theorist will maintain, regardless of whether the phenomenological or the teleological approach is taken, that the various uses of *intuition* found in classical Indian philosophy form a common kind. A *non-common kind* theorist will hold, for example, that although the experience of having an *intuition* across the various uses of *intuition* has a common phenomenology, that common phenomenology is not picking out a common kind at the relevant level of explanation. An analogy that can offer guidance here is the case of Jadeite and Nephrite. Although the two gems look macroscopically similar they are in fact distinct gems because of their underlying microscopic differences. The key point is that microscopic individuation is what matters, and not macroscopic resemblance. Similarly, a non-common kind theorist might maintain, for example, that although the Buddhist theory of *intuition* and the Nyāya theory of *intuition* share a common phenomenology, these uses of *intuition* don't pick out a common kind at a lower level of description.

In order to get traction on whether two or more theories of *intuition* might have common features or disparate features one must dig deeper into precise questions that can capture specific components on which one use of *intuition* might differ from another use of *intuition*. Here we provide a partial battery of questions that can be used to help capture how and when two or more uses of *intuition* are similar, and in what respect precisely they are similar or different. Where *I* stands for a gives use of *intuition* on a specific theory there are six main dimension questions:

Dimension Questions:

(i) Given a use of intuition *I*, what are the *proper objects* that *I* is directed towards?

(ii) Given a use of intuition *I*, what is the *phenomenal nature* of *I*?

(iii) Given a use of intuition *I*, what *process* or *processes* account for *I*?
(iv) Given a use of intuition *I*, is *I* a source of justification?
(v) Given a use of intuition *I*, is *I* a source of knowledge?
(vi) Given a use of intuition *I*, is the proper deployment or reception of *I* dependent on training or practice?

Concerning (i) it is important to begin analyzing a theory of intuition by trying to determine what the theory says are the proper objects of *intuition*. For it is possible that two theories of *intuition* maintain that the proper objects of *intuition* are distinct. In general, a theory of *intuition* could maintain that *intuitions* are only directed at mathematical truths, moral truths, philosophical truths, temporal truths (i.e. truths about the future, the past, or the present), particulars, universals, or a number of other entities. However, it is important to note that just because two uses of *intuition* differ in the objects they are said to be about, for example mathematical as opposed to moral truths, it does not follow that the two uses are *non-convergent*. Two or more uses of *intuition* are *convergent* when the theories presenting *intuition* are explicitly set over distinct objects, but because of the remainder of the structure of the theories of *intuition* the two theories could be extended to cover the remaining objects. For example, one theory might hold that *intuition* is about universals, while another theory says it is about moral truths, yet when one looks at the overall structure of the two theories there is no principled reason why the theory concerning universals could not be extended to cover moral truths, and the moral theory could not be extended to cover truths about universals. Moreover, the proper objects of two distinct theories of *intuition* need to be identified for the purposes of comparing the theories and for the purposes of determining convergence.

Concerning (ii) a theory of *intuition* could maintain that *intuition* has a strong phenomenal nature that is presentational or it could maintain that *intuition* has no distinctive phenomenology that is important, rather what is important is how *intuition* is related to a belief or an inclination to believe a proposition or to some other process. The general idea here is that some conceptions of *intuition* will maintain that it has a strong phenomenology that is important, while others will acknowledge the presence of the phenomenology, but downplay the significance of it. The phenomenology of *intuition* is important because some views might hold that the positive epistemic status of intuitions depends in part on the phenomenology of *intuition*, while others would deny it. For important work on the epistemic role of the phenomenology of intuition see Eli Chudnoff's (2014) *Intuition*.

Concerning (iii) a theory of *intuition* could maintain that there is a specific process that underlies *intuition* or that there are a variety of processes that account for *intuition*. At least some of the work that pertains to (iii) comes from cognitive science and neuroscience where theoretical models of cognition and FMRI are used to map which centers of the brain are responsible for the production of certain mental states, such as *intuition*. However, investigation of (iii) is also important for determining the issue of whether a use of *intuition* involves both reason and emotion. For example, a use of *intuition* might maintain that it is wholly rational and not dependent on any emotive or affective processes either for its production or for its evidential status. Another account might maintain that *intuition* is important when it is a function of an emotive processes. For important work on whether or not 'intuition' talk picks out a common process or distinct processes see Jennifer Nado's (2014) *Why Intuition*?

Concerning (iv) a theory of *intuition* could maintain that *intuition* is a source of justification or that *intuition* is not a source of justification. In addition, the important sub-questions here are: how is *intuition* a source of justification? Why is *intuition* not a source of justification? What kind of justification does *intuition* provide?

Concerning (v) a theory of *intuition* could maintain that *intuition* is not a source of knowledge or that *intuition* is a source of knowledge. The important sub-questions here are: If *intuition* is a source of knowledge, how exactly is it a source of knowledge? Is it a form of mediate or immediate knowledge? Is the knowledge a function of one possessing sufficient justification or is it a function of a direct connection or link to the truth-maker for the relevant truth that is known? Finally, if it is a function of a direct link how can that link be established with respect to the relevant objects it is set over? For example, if intuition is set over temporal truths, some of which are in the future, how can a subject have an intuition that is a direct connection to a future event?

Concerning (vi) a theory of *intuition* could maintain that *intuition* is operative in the relevant sense only when the subject has undergone some kind of training or practice. Here the idea is that some theories of *intuition* will say that a genuine *intuition* is present when and only when training and practice has taken place. Other theories will acknowledge the presence of *intuition* even in the absence of training or practice. One further division that can be found with respect to training is whether or not the theory holds that training improves the epistemic quality of the *intuitions* one has or whether training is simply what is necessary to prepare the mind to have an *intuition* in the relevant domain. Finally, how a theory of *intuition* treats question (vi) is often related to how it treats (iv) and (v).

We will now take the dimension questions into the literature we have surveyed.

Proper Objects

Within the work on classical Indian philosophy that has been surveyed here, it is clear that there are at least five discussions of proper domains for *intuition*. First, there is the use of *intuition* in Nyāya that focuses on our knowledge of universals. Second, there is the use of *intuition* in Nyāya that focuses on a yogic perception as aimed at objects in the material world that are either hidden, distant, or subtle. Third, there is the use of *intuition* in Vaiśeṣika that discusses *intuition* as a way of accessing past, present, and future objects beyond the senses. Fourth, there is the use of *intuition* in Dharmakīrti where it is mainly focused on our knowledge of the Four Noble Truths. Fifth, there is the use of *intuition* in Yoga, which is aimed at revealing ultimate truths of reality.

Phenomenal Nature

Within the work on classical Indian philosophy that has been surveyed here, relatively little is said about the phenomenology of *intuition*. We find that it is presented as being an excited state and one that involves a clear presentation of its object. The two most interesting features to point out come from the Buddhist thinkers. One main point they make is that *intuition* is to be related to states such as grief or hallucination. The other main claim that is made is that the vividness of *yogaja pratyakṣa* is used as a basis for arguing that the content of *intuition* must be *non-conceptual*. The argument for this is the following:

1. Representational states that are *vivid* are non-conceptual.
2. *Yogaja pratyakṣa* is representational.
3. So, *yogaja pratyakṣa* is non-conceptual.

That is, rather than talking about the relative strength of various *intuitions* the focus of discussion is on the vividness of it. In contrast to this view, many accounts of *intuition* in moral philosophy, within Western philosophy, would treat *intuition* as having *conceptual* content. The main reason for supposing this to be the case is that *intuition* in Western moral philosophy is directed at a moral truth that is conceptually articulated and brought to consciousness as a judgment to the effect that something is true.

The fact that the Buddhist conception, articulated under Dharmakīrti, treats *intuitions* as having non-conceptual content is, from a cross-cultural-constructive engagement point of view, quiet challenging. For those working on *intuition* in the Western tradition it may seem unimaginable how *intuition* *could* have non-conceptual content in the moral case. While it is true in Kant that concepts and intuitions are opposed to one another in the construction of experience, and thus that there is a notion of *intuition* that is non-conceptual, Kant's notion of an *intuition* is not the same one that is at play when we say, for example, that Bill has the intuition that Utilitarians give the correct answer to the Trolley Problem in which we are asked to determine whether we should kill one to save five. More importantly, though, the argument that Dharmakīrti offers for why *intuition* has non-conceptual content is very interesting and plausible. For, it is not uncommon to argue on the basis of the vividness of a representational state type that the kind of content that state type has is non-conceptual. Comes from the debate between John McDowell and Gareth Evans over whether perception has conceptual or non-conceptual content. McDowell (1996) argues that in order for perception to justify a belief, it must have conceptual content, since belief has conceptual content. However, Evans's (1982) *Varieties of Reference* contains an argument where he argues on the basis of the richness of our perceptual experience that perception cannot have conceptual content, since we don't have enough concepts to capture the fine-grained detail of our perceptual experience. One might wonder whether Dharmakīrti's argument from vividness to non-conceptuality for *intuition* in yogic perception similar to the founding idea of Evan's argument that fine-grained detail of perceptual experience leads to the view that perception must be non-conceptual.

We can examine the potential equipossiblity of these arguments by looking into the cases of moral intuition and mathematical intuition. Our question is: how plausible is it to hold that moral intuitions are non-conceptual in the way that Dharmakīrti seems to? A moral intuition, such as that it is wrong to torture an innocent individual for no other purpose than to cause suffering, is a strongly conceptual intuition, in that it would appear that *to* have the intuition one must possess the concepts of *torture*, *innocence*, and *suffering*. By contrast, the intuition generated by a visual representation of how a closed-concave figure can be modified into the shape of a circle would appear to require no concepts—at least in the sense of being linguistically tied. Rather, it would appear to involve only the ability to see in one's mind eye how a figure of a certain kind could be turned from a closed concave figure into a circle. Thus, at least some mathematical *intuitions*, arguably, would appear to have non-conceptual content. Note that in both cases at least one concept is required, the concept of truth, since in both cases the *intuition* has the partial content either that something is true, such as in the moral case, or that something is possible, such as in the mathematical case. So, we might conjecture the following. If Dharmakīrti thinks of yogic perception *intuition* of moral truths is like mathematical *intuition* rather than how contemporary Western philosophers think about moral intuitions concerning concrete situations, we would have a model to make sense of his account. In some ways the yogic perception *intuition* or suffering is like a mathematicians *seeing* with *survivability* the connectedness of various truths in mathematics in a given domain under a specific proof.

Finally, given this cross-cultural-constructive engagement, the most sensible position to take on the content of *intuitions* is that they can have both conceptual and non-conceptual content depending on the case in question. And that one must be sensitive to what kinds of *concepts* are being excluded or included in a treatment of the question: what kind of content does *intuition* have?

Process

The question of what process or processes underlie a specific use of *intuition* in most cases cannot be determined unless one looks at the actual psychological and neuroscientific data concerning the specific use of *intuition*. However, it is still possible to characterize three distinct views of the processes that generate *intuition*. The three views are: *the rational account*, *the emotional account*, and *the interactive account.*

The rational account maintains that *intuitions* are a product of a rational process. The rational account is most closely associated with mathematical discussions of *intuition*. It is so associated because it is thought that mathematical intuitions derive solely from a rational process. The more controversial case is that of moral intuitions as discussed in our knowledge of fundamental moral truths. Some philosophers would argue that our moral intuitions do not derive solely from a rational process. Rather, they derive from an emotional process working in conjunction with a rational process.[21] These points about the controversy over moral intuitions bring us to the other two theories of *intuition*.

The emotional theory of intuition maintains that the processes that generate *intuition* are emotional and not rational. It is clear that an emotional theory of intuition is unlikely to be coherent in the domain of mathematical intuition. It is far less clear that it could not be of substantial importance in the case of moral intuitions. The interactive theory holds that intuitions are a function of both rational and emotional processes. It is unclear whether a pure emotional theory of intuition is superior to an interactive theory for the case of moral intuitions. In addition, one could argue that all three theories are correct for different domains in which *intuition* occurs. And thus, as a consequence, there is no common kind that falls under the use of *intuition* in rational theories vs. emotional theories.

Justification and Knowledge

We will take the issues surrounding justification and knowledge together. The main reason why is that the biggest contrast between classical Indian and Western philosophical discussions about the epistemic status of *intuition* is that while there is an issue concerning justification on the basis of *intuition* in moral discussions in Western philosophy, there is no discussion of it in classical Indian philosophy. In classical Indian philosophy the main and central question concerns whether or not and how *yogaja pratyakṣa* is a *pramāṇa*. At least one reason for the absent discussion concerning justification is that within classical Indian philosophy the following two views about sources of knowledge (*pramāṇa*) typically hold: (i) they are factive; and (ii) knowledge is non-componential. Because *parmāṇa* are factive it must be the case that if *intuition* is a *pramāṇa*, then it is when one has an intuition the content of it is true. In addition, it will follow on this view that we can have *intuition* like experiences that are phenomenally similar, but not genuine *intuitions* because their content is not true. Because knowledge is non-componential it must be the case that if *intuition* is a *pramāṇa*, it is not mediated by an intervening mental state. For example, in Western epistemology many philosophers take knowledge to factor into the following components: (a) truth, (b) belief, (c) justification, and (d) some anti luck condition strong enough to rule out Gettier cases. By contrast, in Indian philosophy, sources of knowledge are not typically taken to factor into distinct components. Rather, knowledge is take to be a relation between the mind and the proper object it is set over.

Now, in the Western discussion of ethical intuitionism there is room both for discussion of justification and knowledge of basic moral truths on the basis of *intuition*. On the classical Indian side of the discussion the central question surrounds knowledge of moral truths by way of *intuition* as a way of

[21] One should note however that there is literature within experimental philosophy and cognitive science that discusses the possible ways in which *intuitions* about moral cases depend not on rationality, but rather emotions or affective processes. See for example work by Joshua Greene and Jonathan Haidt.

distinguishing between knowledge that is gained through a teacher's instruction versus knowledge that is gained on one's own. (The episodes narrated from the *Mahābhārata* under the Yoga section bring this out most poignantly.) *Intuition* serves as a possible route for an individual to gain knowledge of moral truths independently of a teacher's instruction. In addition, the knowledge of moral truths that one acquires has the features of being (i) direct, (ii) unmediated. However, with respect to the non-conceptuality of *intuitions* in the case of moral truths, it is not clear that every school of classical Indian thought would subscribe to this account. The Buddhist, under Dharmakīrti's articulation, would. However, it is plausible that the Nyāya would hold that an *intuition* of a moral truth is conceptual.

Finally, the most interesting connection we can draw cross-culturally is that there is a connecting line through a famous problem in 20th century discussions in the philosophy of mathematics to discussions of ethical intuitionism all the way to Dharmakīrti's discussion of *intuition* of the Four Noble Truths.

The central problem that one encounters in the Western context for thinking about how intuitive perception can be a source of knowledge or justification is the contact-problem, elsewhere known as the Benacerraf problem.[22] The problem is initially presented for the case of mathematics, and then can be altered for the case of morality. The initial set up is based on a set of inconsistent claims.

(Causal Isolation) The truth-makers for mathematical statements, such as that $1 + 2 = 3$, are abstract objects which are causally isolated from humans.

(Causal Connection) Both justification for believing that *p* and knowledge that *p* require some kind of causal contact between the subject and the truth-maker for *p*.

(MAJK) We do possess some knowledge of mathematical truths, and we do have justification for believing many mathematical claims.

The claims above are inconsistent. For if (MAJK) is true, then either the truth-makers for mathematical statements are not causally isolated, or, neither justification nor knowledge require a causal connection between the subject and the truth-maker of the proposition. On the basis of taking a certain line in the philosophy of mathematics, namely that the truth-makers are causally isolated from human subjects, one could argue that mathematical intuition is useless as a basis for justifying mathematical beliefs or providing one with knowledge of mathematical truths. For one could argue that without a causal connection between humans and mathematical objects, which are the truth-makers for mathematical statements,

[22] See Benacerraf (1973) for the original articulation of this problem for the case of mathematics. One should note that the problem is more general than the one articulated by Benacerraf. Because the problem is more general it is being discussed here under title 'the contact problem.

mathematical intuitions could not be reliable. The core idea is that reliability requires connection.[23] The contact problem can be extended to the case of moral cognition. A simple formulation of it would be the following:

(Causal Isolation) The truth-makers for moral statements, such as that *causing innocents to suffer is morally wrong*, are universals understood as non-concrete entities.

(Causal Connection) Both justification for believing that *p* and knowledge that *p* require some kind of causal contact between the subject and the truth-maker for *p*.

(MOJK) We do possess some knowledge of moral truths, and we do have justification for believing many moral claims.

On the moral account, if one maintains (MOJK), then it appears that we cannot know moral truths on the basis of moral intuition. For how does moral intuition put us into contact with the truth-makers for moral statements, abstract universals.

Now although this version of the problem is not present in classical Indian philosophy, Dunne's discussion of Dharmakīrti engaged a certain problem within Buddhist philosophy. Recall the problem:

1. To realize the Truth of Suffering, one must realize the impermanence of everything, since the impermanence of everything is part of what constitutes the Truth of Suffering by being a cause of suffering for each thing that does suffer.
2. The impermanence of everything is not something over and above all things that are particular and impermanent. There is no real universal of impermanence, which everything participates in. Rather, impermanence is abstracted from the particular impermanence that each and every thing undergoes.
3. Yogic perception, being a perception, is only of particular things that can be causally efficacious in the production of an image in the mind.
4. So, it cannot be that in having a yogic perception of the Truth of Suffering one is put into contact with the universal *impermanence*.

In the case of our potential knowledge of the Truth of Suffering through yogic perception the problem is that we only have yogic perceptions of particulars, and not universals, since on the Buddhist ontology there are *no universals*. Dunne's formulation of the problem is not done by way of a contact problem. But it can easily be formulated as such by the following argument.

(a) It is possible to have a yogic perception of the Truth of Suffering only if one has a connection to the universal of impermanence, which would provide them with contact to the relevant truth-maker for the truth of suffering.
(b) On the Buddhist ontology there are no universals.
(c) So, it is impossible to have a connection with the universal of *impermanence*.

[23] For an account of rational intuition that challenges the problem presented via *the contact-problem* see Chudnoff (2014).

(d) So, it is impossible to have contact with the relevant truth-maker for the Truth of Suffering, which is necessary for having a yogic perception of the Truth of Suffering.

Thus, exploration of the moral case, cross-culturally, reveals a general problem concerning *intuition* and its objects. Namely: How can *intuition* be a source of knowledge, if the truths it is supposed to provide us knowledge of rest on a domain of objects that are inaccessible to human minds?

Training and Practice

The final dimension that must be explored is that of training and practice. And it is here that we find a feature that stretches across all schools of classical Indian philosophy. The core debate is over two features of training and practice: (a) whether training or practice is relevant to the issue of generating *intuitions* that can provide one either with justification or with knowledge, and (b) what kind of training or practice is relevant? In the classical Indian context there are several important features of the discussion on training.

First, the training is to be (a) ethical, (b) physical, and (c) mental. In the Yoga school, yogic perception requires that one both *act* in certain ways and *abstain* from acting in other ways. These ethical practices prepare one to have yogic perceptions. Yoga, itself, requires a physical practice of *asana*. These practices play a role in training the mind to be *still*. It is theorized that the physical practices place the body in a position that allows one to train the mind in being focused because the body being in that position makes the mind want to *move about*. Of course *asana* practice is not just about the stilling of the mind, since *asanas* also provide other positive benefits. But in relation to *intuition* this is the primary purpose. And finally meditation is necessary in order to still the mind when the body is not in a difficult position. The core idea across all three of these is that certain practices prepare the mind by training it to have *intuitions* that are of a certain epistemic quality.

Second, the training would appear to be *domain general*. In the case of mathematics, one would argue that training in mathematics alone is what is relevant to having mathematical intuitions that are reliable and trustworthy for forming mathematical beliefs and gaining mathematical knowledge. But perhaps one would hold off on arguing that training the mind in general is an important step towards gaining reliability within the mathematical domain.[24] By contrast, in the case of the Yoga School there are two ideas of relevance. Training the mind in general is relevant for having *intuitions* in general. But also training the mind to focus has a spill over effect into many other aspects of one's life. Not just in areas that pertain to

[24] The claim we make here about the relation between training the mind in general versus training the mind in mathematics is a conjecture about what some philosophers of mathematics might say. We take it that some, perhaps influenced by Husserl, would say that training the mind in general is also an important step towards having reliable mathematical intuitions. And that those influenced by work in philosophy of mind on the role of attention in perception, would likewise claim that training the mind to be attentive in general is an important step toward having reliable mathematical intuitions.

domain specific *intuitions* such as in the areas of morality, mathematics, or metaphysics.

From a cross-cultural point of view the key insight that is to be gained is that a theory of how *intuition* can be a source of justification or knowledge might be further investigated by looking into how training the mind in general either positively or negatively effects one's *intuitions* in specific domains. It could be that by training the mind in general to be attentive and focused, and to engage extended concentration one is able to have *intuitions* with a stronger phenomenal and epistemic quality.

Concluding Remarks

Both intuition and perception are prominent features of our cognitive lives. It is striking to find so many treatments of perception from a cross-cultural-constructive point of view in comparison to the total absence of any in the field of intuition.[25] The present study aims to rectify that problem with the hope of encouraging more cross-cultural-constructive works on intuition. Some areas in which there could be more development are the following.

From a historical point of view it would be interesting to see work comparing historical discussions of *intuition* in Western philosophy with specific schools of classical Indian philosophy. Potential comparisons could engage various members of the Nyāya School, such as Gaṅgeśa and Udayana, with various Western rationalists, such as Descartes and Spinoza.

From a cognitive science point of view it would be interesting to see work that brings *intuition* as discussed by Kahneman (2011) into contact with any of the schools of Indian thought that discuss different ways in which one can train the mind to have trustworthy *intuitions*.

From a comparative philosophical point of view it would also be interesting to see more work exploring the different processes that bring about *intuition* as discussed here under the topic of reason and emotion. It is likely the case that comparative examination of both emotion and intuition would be highly useful to enhancing our understanding of *intuitions* as generated by emotion as opposed to those generated by reason. Rational intuition has received far more attention in the recent literature than the topic of emotional intuition.

Finally, it is important to close this study by engaging the question: why is a cross-cultural-constructive engagement of a phenomenon useful? There are perhaps many answers to this question, both positive and negative. We will close with a positive answer reflecting our own theory.

[25] For an example of an excellent recent work on perception from a cross-cultural-constructive point of view see Coseru (2012). In this work Coseru develops a Buddhist account of perception while also engaging work from Western epistemology and philosophy of mind as well as neuroscience and cognitive science.

A cross-cultural-constructive engagement of a phenomenon is useful because it plays an instrumental role in enhancing our understanding of the phenomenon. Understanding a phenomenon fully and robustly requires approaching the phenomenon from as many points of view as that phenomenon allows for. **Intuition** being a pervasive feature of the human condition surely admits of a cross-cultural investigation, as opposed to simply a scientific investigation or theoretical investigation. A cross-cultural investigation sits alongside an empirical and a theoretical (a priori or non-experimental) investigation. It does not override the latter two investigations. Rather, it complements both. As we seek a theory of *intuition* we seek it from every corner from which it has been investigated and theorized about.

Another significant outcome of a cross-cultural engagement might be the critical pay-offs. What we mean by this is that—as we began to argue in the introductory section—far too often philosophers and intellectuals generally hold an extremely polarized picture of the West and the East, roughly paralleling the erstwhile distinction between reason, on the one side, and passions, mystical forebodings, meditation, and such other esoteric pursuits or predilections, on the other side. Philosophy in the West is supposed to be built on a solid foundation of science, reason and argumentation.[26] The truth of course is that these distinctions and the cleavage painted are suspect and ultimately misleading. Form our discussion of the Indian approaches to *intuition* (and from our other work on metaphysics and epistemologies of India) it should be apparent that there is as much difference and diversity within Indian philosophical traditions (or for that matter Chinese) as there is in Western traditions, from Ancient Philosophy to Continental and post-secular philosophies. Second, that reason, logic and argumentation are not just the provenance of the West; these art-forms were rigorously practiced and developed in India (and in the works of Indian philosophers spread-out more globally today). Thirdly, what is lost in the accounts of the history of philosophy is that much of the influences that aided if not propelled the growth of philosophy in what is nowadays considered as the 'West' came from the infiltration of or borrowings from Indo-Aryan ideas (to the East as much as to the West of their original home in Central Asia). Many parts of Europe extending out from Russia were considered as integral to the "Orient" (hence the use of 'Oriental' for 'Orthodox' that is still current in some parts of that world); and Germany was very much part of the 'East' or the 'Orient' until it was transformed in the Middle Ages through an infusion of Indo-Aryan and Hellenistic ideas.

Furthermore, mysticism of many kinds, and in some instances of quite wild varieties, was still rife between 16–19th century Germany, and few philosophers of the period could escape the temptations, including Kant, Nietzsche, Hegel, and Schopenhauer, to name a few. One might even venture to suggest that the contemporary (rekindled) interest in the West in ESP, Psi, and paranormal cognition, have their roots in the 'Western' cultures of the Romantic period; that phenomenon such as psycho-kinesis, clairvoyance, precognition, including psychic-spiritual

[26] Solomon 1995: 253.

mediumship, were known and widely practiced in these cultures (frowned upon, of course, by the churches, that led to gruesome ‘witch-hunting’). Their roots in the academic thinking went back to Paracelsus, whose epistemology was based on the integration of three modes of ‘knowing’: empirical, scientific, and intuitive—where the last two supervene sequentially on each of the prior methods: thus one has an intuitive understanding (*experientia*) of the property of the object known (*scientia),* that is encountered by the senses (*experimentum*). This surfaces more in Kant’s schemata of the categories of understanding than in his use of the trope ‘intuition’ (for sensation qua *experimentum*).[27] Kant, who was influenced in no small measure by the mystic ‘first German philosopher’, Boehme, is reported to have pondered on the possibility of a ‘sixth’ (extra-ordinary) sense, and invited the Swedish seer Swedenborg to show him the apparent workings of the occult sense, though the experiment failed and Kant remained unconvinced (in theory at least). Several 19th century scientists, such as Wolfgang Pauli, drew on spiritual or occult archetypes to even explain Kepler’s configurations of the heavens. 20th century physicists such as Julian Huxley, Schrödinger, Eddington, Oppenheimer, perhaps also Einstein—not to mention the renown Indian mathematician in Cambridge, Ramanujan—showed strong leanings towards transcendental metaphysics (some drawing on the Indian Upaniṣads). From this basis they arguably ventured scientific conjectures and proffered predictions as well that awaited empirical or mathematical verifications; such ‘non-scientific dabbling’, some might call it, nevertheless influenced their scientific thinking as much as their regular life-styles.

Constructive comparative philosophy, then, can be seen to play a crucial role in disabusing the moderns of the simplistic and over-determined view that *because* theories of intuition are (historically in the West) grounded in occult metaphysics, they have no relevance to or impact upon how one does science, or philosophy for that matter, and that the discourse should, if not already has been, relegated to the dustbins of the ancient world to the ‘East’. When in fact the history and career of intuition in the East has been quite the opposite.

Drawing on an analogy, at one time this also was said of emotions and passions, and also more generally of ethics or moral philosophy; but this all changed in contemporary times, and as a result of interventions from many quarters, some rather interesting work has been done in cross-cultural studies of emotions, not just by anthropologists and psychologists, but by philosophers also, so that we begin to better understand and appraise the claims of the universality of emotions, or at least of certain of the emotional responses, sentiments and passions, and how these form part of moral judgments (an area we cannot go further into here, but touched upon in an earlier section) (Solomon 1995). So it behooves modern-day philosophy not to regard constructive comparative philosophy to be a tangential or irrelevant pursuit in our quest towards understanding some of the common threads that just might run through the cultures and philosophies of the plural worlds, near and far.

[27] See Gibbons 2001: 11, 15, 52, and 91.

Raimon Panikkar, writing in the 1980–90s, drew attention to a further virtue of CQI, by adding another methodological element to the erstwhile practice, and this he labeled, the ‘imparative hermeneutic’. ‘Imparative’ is derived from the Latin *impayare,* to implore, to confront. He explained that in this method a real space of mutual criticism and fecundation is opened up for genuine encounters between different philosophical and cultural traditions. One ‘enters’ into another’s dimensions of intellectual or cultural ‘meaning’, and allows that to speak to (and reappraise) one’s own convictions in a dialogical situation. One then assumes a neutral vantage point from which assessment is made of the comparative worth of the aspects investigated. In Panikkar’s view, this provides a needed antidote to the kind of ‘mono-formist’ culture of philosophy that hitherto has all but sounded the death-knell of the rich and varied particularities of the various philosophical and cultural traditions, globally extant, each of which may have something unique to offer. Thus, he views the larger objective of the Imparative-hermeneutic program to draw into dialogue different perspectives (from among the various traditions) to address real–life and global issues in such a way that comparisons can become relevant to the human condition, to the problems and crises that face humankind regardless of whether religions are implicated or not. Imparative philosophy proposes that ‘we may … learn by being ready to undergo the different philosophical experiences of other people’ Pannikar (1988: 127). Associated with such imparative work is the recognition that nothing is nonnegotiable Pannikar (1988: 128). Panikkar suggests that imparative philosophy employs in this regard diatopical hermeneutics. Departing from morphological hermeneutics—distance within one single culture—and diachronic hermeneutics—temporal distance *á la* Hegel—diatopical hermeneutics is ‘the required method of interpretation when the distance to overcome … is … the distance between two (or more) cultures, which have independently developed different spaces (*topoi*) their own methods of philosophizing and ways of reaching intelligibility along with their proper categories’ Pannikar (1988: 130).

Panikkar is suggesting that there is a phenomenology implicit in this cross-cultural enterprise, and this calls upon the researcher’s conscious engagement with empathy and a preparedness to bracket-out belief in the truth of one or the other position that does not allow for a possible third position suggested in the *imparare* encounter that takes into account the universal range of human experience in as much as it is possible to do so in any concrete situation. Imparative philosophy as an alternative to comparative philosophy may be the antidote to overcoming parochialism, as well as to cultivating tolerance and understanding of the richness of human experience. And here diatopical hermeneutics has the functional role of forging a common universe of discourse (not a common ground through assumed equivalences) in the dialogical dialogue that is taking place in the very encounter. So, Panikkar basically argues that cross-cultural philosophy is a ‘mature *ontonomic* activity of the human spirit, contrasting everything, learning from everywhere, and radically criticizing the enterprise itself’ Pannikar (1988: 136).

Works Cited and Further Readings

J. Alexander, *Experimental Philosophy: An Introduction* (Polity Press, 2012)

P. Benacerraf, Mathematical truth. J. Philos. **70**(19), 661–679 (1973)

Bhāsā-Pariccheda (BP) with Siddhanta Muktavali (SM), ed. by Pt. H. Sukla Sastri (Chokhamba Sanksrit Series Office, Varanasi, 1972)

Bhāsā-Pariccheda (BP) with Siddhanta Muktavali (SM). trans. Swami Madhavananda (Advaita Ashram, Calcutta, 1977)

P. Bilimoria, (ed. with Introduction). *Essays on Indian Philosophy: J. N. Mohanty* (Oxford University Press, New Delhi, 1993)

P. Bilimoria, Radhakrishnan—saving the appearance in Plato's academy, in *New Essays in the Philosophy of Sarvepalli Radhakrishnan*, ed. by S.S. Rama Rao Pappu (Indian Books Centre, Delhi, 1995), pp. 327–344

P. Bilimoria, Abhāva: Negation in logic, real non-existent, and a distinctive pramāṇa in the Mīmāṃsā, in *Logic, Navya-Nyāya and Applications Homage to Bimal Krishna Matilal, Studies in Logic*, vol. 15, ed. by M. Chakraborti, B. Lowe, M.N. Mitra, S. Sarukkai (College Publications, London, 2008), pp. 43–64

Bilimoria, *Śabdapramāṇa: Word and Knowledge as Testimony in Indian Philosophy* (D.K. Printworld (P) Ltd, New Delhi, 2008a)

P. Bilimoria, Extra-sensorial liaisons of 4D yogins: enigma extolled by Nyāya; impeachable to Mīmāṃsās, in *World of Philosophy: A Harmony, Christopher Key Chapple*, ed. by S. Sunanda, M. Intaj, C. Dilip, D. Sri Prashant (Delhi, Shanti Prakashan, 2011), pp. 58–68

P. Bilimoria, Mantric effect, efferverscent Devatās, Noetics of supplication, and the Apūrva in the Mīmāṃsā, in *Sanskrit Studies* (Sanskrit Studies Series of JNU), ed. by S. Kumar, vol. III (DKPrintWorld, New Delhi, 2014), pp. 222–247

K. Chakrabarti, *Classical Indian Philosophy of Induction: The Nyāya Viewpoint* (Lexington Books, NY, 2010)

E. Chudnoff, *Intuition* (Oxford University Press, Oxford, 2014)

E. Chudnoff, Intuition in mathematics, in *Rational Intuition: Philosophical Roots, Scientific Investigations*, ed. by L.M. Osbeck, B.S. Held (Cambridge University Press, Cambridge, 2014)

C. Coseru, *Perceiving Reality: Consciousness, Intentionality, and Cognition in Buddhist Philosophy* (Oxford University Press, Oxford, 2012)

B.C. Das, The Mimāmskas on Yogaja Pratyakṣa: A critique. Indian Philosophical Quarterly XXIX **4**, 419–431 (2002)

J. Dunne, Realizing the unreal: Dharmakīrti's theory of yogic perception. J. Indian Philos. **34**, 497–519 (2006)

G. Evans, Varieties of Reference. Clarendon Press (1982)

B.J. Gibbons, *Spirituality and the Occult From the Renaissance to the Modern Age* (Routledge Publishing, 2001)

Giner-Sorolla, Intuition in twenty-first-century moral psychology, in *Rational Intuition: Philosophical Roots, Scientific Investigations*, ed. by L.M. Osbeck, B.S. Held (Cambridge University Press, Cambridge, 2014)

M. Hawley, *Sravepalli Radhakrishnan. Internet Encyclopedia of Philosophy* (2006), http://www.iep.utm.edu/radhakri/SH2bi

M. Huemer, *Ethical Intuitionism* (Palgrave Macmillan, 2005)

D. Kahneman, *Thinking Fast and Slow* (Farrar, Straus, and Giroux, New York, 2011)

C. Kidd, Husserl's phenomenological theory of intuition, in *Rational Intuition: Philosophical Roots*, ed. by L.M. Osbeck, B.S. Held (Cambridge University Press, Scientific Investigations, 2014)

J. McDowell, Mind and World Oxford University Press (1996)

J. Mohanty, Kalidas Bhattacharyya as a metaphysician, in *Essays on Indian Philosophy: J. N. Mohanty*, ed. by P. Bilimoria (Oxford University Press, 1993b) pp. 33–44

J. Mohanty, The concept of intuition, in *Essays on Indian Philosophy: J. N. Mohanty*, ed. by P. Bilimoria (Oxford University Press, 1993a), pp. 26–33

B.K. Matilal, *Perception An Essay on Classical Indian Theories of Knowledge* (Oxford University Press, New Delhi, 1991)

J. Nado, Why Intuition?. Philos. Phenomenol. Res. **89**(1), 15–41 (2014)

Nyāyamañjarī (NM) of Jayanta Bhaṭṭa, Part I. Sūrya Nārāyaṇa Śukla (edited text). Vārānasī: Chowkhamba Sanskrit Series Office, 95 ff. (*Yogajapratyaksa*); 1971; Part II under *apavarga and ātmaparīkṣā*

Nyāyasūtra (NS) of Gautama, *Nyāyadarśanam* with *Vatsyāyana's Bhāṣya, (*NSBh*) Uddyotkara's Vārttika, Vācaspati Miśra's Tātparyatika* and *Viśvanātha's Vṛtti* (Munshiram Manoharlal, Delhi, 1985)

L. Osbeck, B. Held, *Rational Intuition: Philosophical Roots, Scientific Investigations* (Cambridge University Press, Cambridge, 2014)

R. Pannikar, What is comparative philosophy comparing?, in *Interpreting Across Boundaries: New Essays in Comparative Philosophy*, ed. by G.J. Larson, E. Deutsch (Princeton University Press, Princeton NJ, 1988)

S. Phillips, Counter Matilal's bias: the philosophically respectable in indian spiritual thought, in *Studies in Humanities and Social Sciences*, vol. III, no. 2, ed. by A. Chakrabarti in Honour of Bimal K Matilal, pp. 173–183 (1996)

S. Phillips, *Epistemology in Classical India: The Knowledge Sources of the Nyāya School* (Routledge Press, 2012)

R. Puligandla, Phenomenological reduction and yogic meditation. *Philosophy East and West*, vol. 20.1, pp. 19–33 (1970)

J. Rawls, *A Theory of Justice* (Harvard University Press, MA, 1971)

R.C. Sinha, *Concepts of Reason and Intuition with special reference to Sri Aurbindo, K. C. Bhattacharyya, and S. Radhakrishnan*. (D.K. Printworld LTD. Publishers of Indian Traditions, New Dehli, 1981)

R.C. Solomon, The cross-cultural comparison of emotion, in *Emotions in Asian Thought*, ed. by J. Marks, R.T. Ames (SUNY Press, Albany, NY, 1995) pp. 253–308 (commenting on chapters by P. Bilimoria, G. Parkes, J. Kupperman et al.)

Siddhānta-Muktāvalā (SM): *Nyāyasiddhāntamuktāvalī* with *Bhāṣā-Pariccheda* of Viśvanātha, with Hindi Commentary by Dharmendranāth Śāstrī, (Delhi: Motilal Banarsidass, 1977) (see also BP above)

A. Sjödin, The girl who knew her brother would be coming home: Ārṣjñāna in Praśastapādabhāṣya, Nyāyakandalī, and Vyomavatī. J. Indian Philos. **40**, 469–488 (2012)

Ślokavartika (ŚV) of Kumārila Bhaṭṭa with Commentary *Nyāyaratnākara* of Pārthasārathi Miśra. Svāmī Dvārikādāsa Śāstrī (text ed.) Vārānasī: Tārā Publications (1978)

The Bhagavadgītā (BhG). *In the Mahābhārata,* trans. and ed. by J.A.B. van Buitenen (Chicago University Press, Chicago, 1985)

The *Mahābhārata* (MhB). (Strīparvan) vol. 7. Book XII. The Book of Women. Trans. James L. Fitzgerald (The University of Chicago Press, Chicago, 2004)

The *Mahābhārata* (MhB), Books X and II. Dead of Night/Women. Trans. Kate Cosby. Clay Sanskrit Edition (New York University Press, JJC Foundation, New York, 2009)

A. Vaidya, Philosophical methodology: the current debate. Philos. Psychol. **23**(3), 391–417 (2010)

Chapter 4
The Map and the Territory

John R. Searle

I have in my hand a road map of the state of California.[1] Like all such ordinary objects it is philosophically astounding and I am going to explore some of its astounding features. The interest that the map has for me is not just in the specifics of map productions and cartographic representations, but I have a series of questions of a much more philosophical and indeed almost metaphysical kind about the relation between representation and reality and the implications that these have for our relations to the world. For science in particular and knowledge in general, how does the map represent the territory? First of all, we have to make an assumption that there is a territory with more or less determinate features. The map represents that territory in at least certain essential features. In the case of the map of California, there are all sorts of features of California that are left out of the map such as the number of blonde people living in Los Angeles, or the amount of rainfall that occurs in the Central Valley during the winter months. None of these are represented in the map. What is represented? In order to answer that question, I am going to say a bit about the representing relation. There are series of entities in California, call them cities, mountains, roads, coastline, etc. These are represented how? For each of these entities, there is a mark or area on the map and typically a mark or area with a name next to it. Next to one marked area is "Sacramento" and next to another, "San Francisco". These areas actually stand for Sacramento and San Francisco or whatever else is designated on the map. However, a map is not the same as a list of marks and names. What is added to the lists of marks and names that makes it a representational map? What is added is a *method of projection* of the features of the map to the entities in reality. Naively, we can say that the method of projection is such that, given the method, the relations on the map are identical with

[1] *California*, AAA, 12/16-3/18 Printed in the USA.

J. R. Searle (✉)
Department of Philosophy, Willis S. and Marion Slusser Professor Emeritus
of the Philosophy of Mind and Language, University of California, Oakland, CA, USA

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_4

the relations in reality. So, Sacramento is north of Los Angeles. In reality, and on the map, it is exactly the same: Sacramento is north of Los Angeles. However, in the map, there is literally no north and south, there is simply the representation of north and south. That introduces a question of what is the method of projection. Well, the method of projection in this particular map is that we assume that the Earth can be represented at least in portions on a flat surface. We assume that the map has a top and a bottom. We assume that north is at the top and south is at the bottom, west to the left as we look at the map, and east to the right. Now, given those relations on the actual sheet of paper, we can say that the relations of the marks, "Sacramento", "San Francisco", "Los Angeles", etc., have to be exactly the same on the map as they are in real life. So, on the map, there is an area on the top half of the map that represents Sacramento and it is nearer to the top than the big blotch that represents Los Angeles. That is exactly what is meant by saying the map represents Sacramento as being north of Los Angeles.

How then does the map represent? Well, it is tempting to say, and to an extent it is indeed true to say, that the map is a kind of picture. There is a picturing relation between the map and the territory. However, it only forces the question back: What is a picturing relation? It is not enough to say that the map looks like the territory, because, of course, from most points of view it does not. However, there are locations in airplanes and rockets from which looking at the territory will be somewhat like looking at the map. The map is a picture, in a sense, of the territory. How? We could say as a start that the relation between the elements of the map is isomorphic to the relation of the corresponding elements of reality. That is right. Now we have to explain "isomorphic". We already started with that when we said that each of the elements on the map represents an element in reality and the relations on the map, given the method of projection, are identical with the relations in the real world. That is what is meant by saying that it is isomorphic and in that sense the map is a kind of picture of the territory. I have in fact an aerial photograph of the Pacific coast line south of San Francisco and use it as a map showing the relations between my home in Berkeley and my coastal place south of Half Moon Bay.

There are other features of the map that are not matters of picturing but more like language. For example, national highways are in red, state highways are in black. This is not because of the different colors of the roads but as a conventional, language like, way of representing the difference.

Wittgenstein in the *Tractatus*[2] tried to make the picturing relation of the sort that I have described in maps essential to the nature of meaning and representation. He thought ordinary language sentences disguised the actual logical structure of both the representation and the reality that it represented. Under analysis, he thought that the sentences of ordinary language would be disguised, complex versions of the most basic, elementary sentences, that these sentences consisted of arrangements of names, and that the arrangement of names in the sentence pictured the arrangement of objects in the fact. The basic components of reality for Wittgenstein are not

[2] Wittgenstein, L. *Tractatus Logico-Philosophicus*, London, Routledge and Kegan Paul, 1951.

objects but facts. The object just is constituted by its possible combinations with other objects to exist in facts. Wittgenstein has a problem: what do you do about false statements that are nonetheless meaningful? He says, in order to account for that, you need a distinction between the *Sachverhalt* and the *Tatsache*. Sachverhalt is a possible state of affairs. Tatsache is an actual state of affairs. If the representation of the possible state of affairs represents an actual state of affairs, then the statement or proposition, the Satz is, true. If it does not, that is if there is and a Sachverhalt that is not actual, then the Satz is false.

It is fair to say that Wittgenstein's effort to get a general account of language using this apparatus failed. Why? The most obvious answer is that there are all sorts of relations represented in language which the picturing model does not work. Think of the sentence, "Trump's elections revealed dissatisfaction among the white middle classes." How would you draw a picture of that? Or even a simpler component of it, "Trump was elected." How do you draw a picture of that? Even if you break it down into individuals voting, how would you have a picture of the individuals that could amount to saying, "Trump was elected". The interesting thing is to see how far the Wittgenstein model does work for maps. Can we think of an actual arrangement of objects in the world as a Tatsache and the arrangement of elements in the map as a proposition, a Satz, that represents the Tatsache? Up to a point, I think it works. The problem is, it does not yet account for the essential thing, the representing relation. The idea that Wittgenstein has is that the fact of isomorphism already constitutes representation, but of course, it does not. There are various ways of showing this. One is, if you think that the isomorphism between map and territory was sufficient to guarantee representation, then why is the territory on the earth not a representation of the map? That is, isomorphism is symmetrical. A is isomorphic to B implies that B is isomorphic to A. But the representing relation is not symmetrical. The fact that the map represents the territory does not imply that the territory represents the map. The isomorphism does not yet guarantee the representing relation. What fact about the map makes it a representation of a territory, given that the isomorphism is not sufficient? The answer, I believe, to that question is to invoke the fundamental notion implicit in all of this and that is the intentionality of the user. It is only a map if it is intended to have certain conditions of satisfaction. Indeed, that is the case with meaning in general. Meaning is the imposition of conditions of satisfaction on conditions of satisfaction. The production of the map is the condition of satisfaction of the intention to produce it, but in addition to the production of the map, we and have a further set of conditions of satisfaction. Namely that there should be a matching relation between the elements of the map and the elements of the territory. Don't worry if you do not understand this jargon of "conditions of satisfaction". I will explain it later.

Wittgenstein's effort to reduce meaning to isomorphism is one of a long history of efforts to explain meaning in nonsemantic, non-intentionalistic terms. Like all other such efforts it fails. Meaning cannot be reduced to something non-intentionalistic. Why would anyone want to do this reduction? The feeling is that if meaning really exists then it must be reducible to some non-intentional phenomena. In a world consisting of basic phenomena, as described for example by physics and chemistry, meaning cannot be one of the basic phenomena. In short the reductions are motivated by the traditional reductionist urges. I want to argue on the contrary that we have to recognize that intentionality is a basic feature of reality, not reducible to something else. Along with life and consciousness it is a biological phenomenon. Like consciousness it is not reducible to something else. But why should it be? It is just a fact about how nature works that human and some animal brains create consciousness and intentionality. Meaning is a form of derived intentionality in a way I will shortly explain. The *derived* intentionality of maps, pictures, sentences and signs can be explained in terms of the more basic *intrinsic* intentionality of perceptions, beliefs, desires, etc.

Once we have introduced the notion of intentionality, we then get a much simpler analysis of meaning from the one in Wittgenstein. It does not solve all of our problems by any means, but at least it avoids the obvious counterexamples and inadequacies of the *Tractatus*. The obvious counter examples are that there are lots of representing relations that are not isomorphisms. But furthermore, if the map represented the territory, the territory would have to represent the map, and that is how the *reductio ad absurdum* works. Isomorphism by itself is neither necessary nor sufficient for representation.

If the map model does not work for meaning in general, how does meaning work? I think of linguistic meaning as an extension of a more biological basic phenomenon of the capacity for human minds to represent objects and states of affairs in the world. The unfortunate name we have given to this is "intentionality", and I will continue to use that word with the usual proviso that there is no special connection between intentionality and intending. Intentionality includes not just intending, but also beliefs, hopes, desires, perceptions, the emotions and lots of other mental phenomena. Intending is just one kind of intentionality, among many others.

To understand intentionality[3] you need a few basic notions: first*, the distinction betwee*n *content and type*. The three types of Intentional states—beliefs, perceptions and desires have the same content when I believe that it is raining, wish that it were raining and see that it's raining,. We can represent these as Bel (It is raining), Des(It is raining) and Visual Experience (It is raining). The general form is S (p), where the "S" marks the type and the "p" the propositional content. Second, the distinction between different directions of fit applies. Beliefs and perceptions are supposed to fit how the world is: they have the mind–to-world direction of fit. Desires and intentions are supposed to represent how we would like the world to be or intend to make it be. They have the world-to-mind direction of fit. Third, the notion of conditions of satisfaction: If the fit actually comes about, if the belief is true, the intention carried out, and the desire fulfilled we can say in each case that the intentional state is *satisfied.* We can say that every intentional state with an entire

[3] The account which follows is a summary of the account in Searle, J.R. *Intentionality: An Essay in the Philosophy of Mind,* Cambridge, Cambridge University Press, 1983. For more details see the original version.

propositional content and the direction of fit of either mind-to-world or world-to-mind is a *representation of its conditions of satisfaction.* This is the key to understanding intentionality. Intentionality is mental representation and the most important intentional states, those that have an entire propositional content and a direction of fit, are representations of their conditions of satisfaction.

If we assume that the mind represents the world in perception as well as in thought, then we have a notion of intentionality as a matter of representing and we can think of linguistic meaning as an extension of that more biologically basic notion. How exactly is it extended? Well, if I look outside and see that it is raining then I have a visual experience whose intentional content is that it is raining and the experience is caused by that fact. I put this by saying that the *conditions of satisfaction* of my visual perception are that it should be raining and that this very experience should be caused by the fact that it is raining.

VisExp (It is raining and the fact that it is raining causes this VisExp)

Because the Causally Self Reflexive feature is common to all visual experiences I find it useful to put the notation CSR into the intentional type. The notation above might mistakenly suggest that you have to see the causal relation. So I prefer this notation:

VisExpCSR (It is raining.)

This means exactly the same as the original, but I hope it is clearer.

So there is a self reflective or self referential component in the intentional content of the visual experience: the conditions of satisfaction of my visual experience require reference to the visual experience itself. Now, if I then, on the basis of my seeing that it is raining, I form the belief that it is raining, the belief that it is raining has the same condition of satisfaction but without the causally self reflexive feature. The condition of satisfaction of the belief is simply that it is raining.

Bel (It is raining)

But suppose I utter the sentence “It is raining”, then what have I done? What is in common to the utterance of the sentence and to the belief that it is raining? Well one thing, obviously, is that the utterance itself has the same condition of satisfaction as the belief. The condition of satisfaction of the belief is that it is raining. The mental state has the form S (p). The utterance has the form F (p) Where the “F” marks the Force of the utterance, that of a statement, command, etc. and the “p” the propositional content. Thus the form is

|- (It is raining)

We uses Frege’s assertion sign, “|- ”, to mark the force of assertion.

In this case the utterance, an assertion, has the same conditions of satisfaction as the belief. But how did that utterance get those condition of satisfaction?

The intentionality of beliefs, perceptions, desires and intentions is as I said earlier *intrinsic* or *original*. It could not for example be that very belief if it did not have that intentionality. The intentionality of sentences, pictures, maps etc. is *derived*. That very sentence, "It is raining" could have meant something completely different if a different meaning had been imposed on it. The name for such derived intentionality is "meaning."

The utterance gets the condition of satisfaction, gets its meaning, because we have intentionally imposed that condition of satisfaction on the utterance. In this case we are using an existing sentence, "It is raining" which already has that it is raining as its standard sentence meaning. But imagine a case where we don't have that. Where I just make a gesture or signal to someone that it is raining and I expect that the hearer will understand what I *mean* when I make the gesture. Then what it is about that gesture that makes it meaningful? What makes the otherwise meaningless gesture meaningful? It has the same condition of satisfaction as did the original belief, just as the belief had the same condition of satisfaction, minus the causally self reflexive feature of the visual experience. This is the essence of speaker meaning. We intentionally impose conditions of satisfaction on our marks and on our utterances, and we can say that meaning consists of the intentional imposition of conditions of satisfaction on conditions of satisfaction. Why do we say it that way? Because the production of a mark or the utterance is the condition of satisfaction of the intention to produce it. Thus the intention to produce it has the condition of satisfaction that it should be produced. But if it is meaningful, then it has additional conditions of satisfactions, it has truth conditions in this particular case.

You can see this more clearly if you look at the use of actual sentences in existing languages. Suppose I am learning French, and I practice saying to myself, "il pleut, il pleut, il pleut", then the conditions of satisfaction of my intention is just to produce the utterance. I am practicing pronunciation, but I do not mean what I say. But if I say "il pleut" and mean it, I actually mean that it is raining, then the conditions of satisfaction are not just that I produce the sentence "il pleut", but that the utterance has the additional condition of satisfaction that it is raining; and this reveals the point I was trying to make earlier, the essence of speaker meaning is the intentional imposition of conditions of satisfaction on conditions of satisfaction.

Not all meaningful utterances have an entire propositional content. "Hurrah for the team" or "Down with the Fascists" do not have an entire proposition. Their form is not F(p) but F(n). Some speech acts have no propositional content at all, "Hurrah" "Damn" or "Ouch" but the general phenomenon of language is representation and meaningful speech acts with direction of fit and entire propositional content have conditions of satisfaction, and represent, in the different possible speech act modes, states of affairs in the world.

We can now understand the "meaning" of the map. The conditions of satisfaction of the map are that the objects and relations in the world should be isomorphic to the objects and relations in the map. The map has the map-to-world direction of fit, and given its meaning we can use it to contain real information about the world.

This gives us a very general account of meaning in language.

I said that Wittgenstein's efforts in the *Tractatus* to explain the essence of language failed. In his later work the Philosophical Investigations[4] he gives up on the idea that there is an essence of language. He thinks that there are countless (Unzählige) different ways of using language, different kinds of language games, and he thinks it is a mistake to look for an essence of language in the way that I have been doing. I think his later account is also mistaken. It is true that there are lots of different uses of language but the culturally and biologically most fundamental are in the performance of speech acts that have a propositional content. These come in large numbers: consider some names in English. There are statements, assertions, questions, orders, commands, requests, hypotheses, promises, avowals, pledges, apologies, thanks, congratulations, vows, threats, etc. But though large the numbers are by no means infinite nor even so large as to be unmanageable. When you consider in detail how it works in the speech acts that have the structure F(p) it turns out that there are five and only five basic speech act types and I will simply list these;

First, Assertives. Their purpose is to tell us all things are in the world. The philosopher's favorites are statements and assertions. These have the word-to-world direction of fit and they take any propositional content. Using Frege's assertion sign "|-": For the Assertive type we can say they have the form:

|- (p).

Second, Directives. Their purpose is to attempt to get the hearer to do something. Favorite examples are orders requests and commands. They have the world-to-word direction of fit and their propositional content always refers to a hearer and a voluntary act by the hearer. Using the shriek mark for the type, the general form is

!(H does A).

Third, Commissives. Their purpose is to commit the speaker to doing something. The philosopher's favourites are promises, but vows threats and pledges should be included. The propositional content is always that S does A, and the direction of fit is world-to-word. So the general form using "C" for the speech act type is

C (S does A).

Fourth, Expressives. The purpose of these is just to express some feeling or attitude typically about a state of affairs which is presupposed to exist. In almost all cases

[4] Wittgenstein, L. *Philosphical Investigations*, Oxford, Basil Blackwell, 1953.

these are about the speaker or hearer. You *apologize* for something you have done and *thank* for something the hearer has done. Wecoming and congratulating are other famous examples. Typically the fit is presupposed. So the general form is to attribute some property to S or H and express an attitude. The general form is thus:

E (S/H + property).

Fifth, Declarations. The purpose of these is to create a new state of affairs by representing the state of affairs as existing. Adjourning the meeting by saying "I adjourn the meeting" or declaring war by saying "War is declared" are examples. They have both directions of fifth because they make something the case—and thus achieve the world-to-word direction of fit by representing it as being the case, by the word-to-world direction of fit. Any propositional content in principle can occur in the Declaration but for humans the possibilities of what we can create by Declaration are severely limited. Not so for gods. When God says "let there be light" that is a Declaration. It makes light exist by declaring it to exist. The general form is:

D (p).

Contrary to Wittgenstein's claim that uses of language are "countless" and that there is no essence of language we see that the representing relation is pervasive in language and it is marked by the occurrence of a propositional content in just about all of the most important uses of language: Assertives, Directives, Commissives, Expressives and Declaratives. It doesn't matter whether we call this an "essence". The important thing is to see that in an understanding of the functions of language it is essential to see that the representing relation, which is the biologically essential feature of intentionality, is extended in language and made much more powerful than in the prelinguistic forms. Prelinguistic animals in cooperation can do a lot. But they cannot create nation states, operate universities, organize wars, stock markets, literary festivals or write books on philosophy.

Back to the map and the territory: We can see that the representing relation, though a paradigm, is not a model for all language but is a special case based on resemblance.

I have tried to explain some more general properties of meaning and intentionality. In intellectual life one of the worst things we can do is give our readers the impression they understand something when they do not. Based solely on reading this article you do not have a thorough understanding of intentionality and meaning, but I hope you do understand two things that I'm trying to get across. Intentionality is a basic biological phenomenon, as much a part of the natural world is digestion or photosynthesis. Linguistic meaning is a form of derived or imposed intentionality.

Chapter 5
Iconic Representation: Maps, Pictures, and Perception

Tyler Burge

Maps and realist pictures comprise prominent sub-classes of *iconic* representations. The most basic, most important sub-class is perception. Other types are drawings, photographs, musical notations, diagrams, bar graphs, abacuses, hieroglyphs, and color chits. I will say something about what it is to be an iconic representation and why a prominent way of thinking about iconic representation is misconceived. Although I am primarily interested in what it is to be iconic, and in the iconic nature of perception, what I have to say will, I hope, illuminate the iconic nature of maps and pictures.[1] Both rely on iconic aspects of visual perception.

A primary theme of this article is that, like all representation, iconic representation gets its representational structure from the nature of the representational functions and competencies that underlie its use. In fact, representational structure marks aspects of representational functions and competencies. Iconicity is an aspect of representational format. Although it affects how a subject matter is represented, it is not an aspect of *basic* representational structure or function. The basic representational structure and functions of iconic representation are also present within the structure and functions of non-iconic language and non-iconic thought.

The key intuitive idea underlying the notion of iconic representation is that it is marked by natural correspondences between units of representation and entities in

I am indebted to Ned Block for comments on an earlier version.

[1] For a fine discussion of differences between maps and language, see Elizabeth Camp, 'Thinking with Maps' *Philosophical Perspectives* 21 (2007), 145–182.

T. Burge (✉)
Department of Philosophy, UCLA, Los Angeles, CA, USA
e-mail: burge@ucla.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_5

the represented subject matter. Here is a somewhat fuller characterization of iconic representation.[2] A representational content, or representation, R, is *iconic* by virtue of meeting the following three conditions:

(1) There is a natural, systematic 1-1-into mapping from one or more types of structural representational units, or constituents, of R, or from a repertoire that includes R, to corresponding types of entities in the subject matter that R functions to represent.[3]

(2) The mapping in (1) preserves correlations between some *relations* among structural representational units of R, or within the repertoire in which it is embedded, and natural relations among entities in the represented subject matter.

(3) R represents the relevant entities, and relations among them, in the represented subject matter, partly by way of the mapping cited in (1) and (2).

Condition (1) allows a whole representation to count as a constituent. Condition (2) allows identity as a limit case of a relation.[4] These points accommodate *very*

[2] Representations are not the only sorts of things that can be iconic. In more extensive, forthcoming work on iconicity, I explicate iconicity for *information registration*. A state X informationally registers state Y if and only if (a) instances of states X and Y statistically co-vary in a significant way, (b) instances of X tend to be caused by instances of Y, and (c) X's meeting conditions (a) and (b) is functional. Information registration is not representation. In my terminology, truth is propositional veridicality; accuracy is non-propositional veridicality. Representation requires having either accuracy conditions or truth conditions as part of the nature of the state that represents. Initial registration of the retinal image in visual systems does not have, and is not taken in science to have, accuracy or truth conditions. A bacterium informationally registers light. Although the occasional scientist attributes seeing to bacteria and even trees, no bacterium's states are explained in the statements of laws of any science as having accuracy conditions. Information registrations, however, can and commonly do meet conditions for being iconic. Registrations of the retinal image have a structure and function that map iconically to spatial aspects of the retinal image, and degrees of light intensity. In such cases, the function of the natural mapping is entirely biological, not representational. Non-representational, non-perceptual sensory states commonly bear iconic relations to sensed aspects of the environment. For more on the distinction between representation and non-representational information registration, see my *Origins of Objectivity* (Oxford: Oxford University Press, 2010), chapter 8.

[3] That the mapping is functional implies that it could fail to match structural elements in the subject matter. So there can be non-veridical and purely fictional iconic representational mappings. Fictional pictorial mappings are parasitic on real mappings. Non-veridical mappings are parasitic on veridical ones.

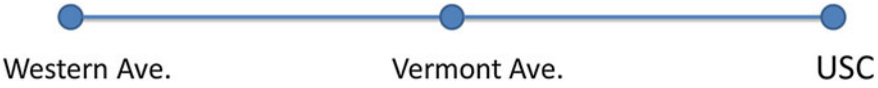
It is possible to allow minor divergences from strict 1-1 mappings. Perhaps for convenience two representational elements could be mapped to a single represented item, as when two circles occur on a subway map for stops at the same station on different subway lines.

I am not fully convinced by such examples. Commonly, different circles represent different positions within the same station. When they do not, it is commonly possible to regard the two different circles as the same representational element, repeated for convenience–or as occupying different maps (one for each subway line). I owe the example to Ned Block. Although I do not insist on strict 1-1 mappings, I take them to be paradigmatic.

[4] I state the first two conditions separately, although condition (2) could be taken to be implicit in what is meant by a natural, systematic mapping in condition (1).

simple iconic representation. For example, a color chit lacks a relational structure, ordinarily understood. A color chit might represent iconically through its color's being the same as the color that is represented.

Take a slightly less simple example of iconic representation. Suppose that the following map represents the light-rail line between the Western Avenue stop and the USC stop:



The dots and the positions of the names iconically represent the stations and their relative positions, and the lines iconically represent the relevant portions of the light-rail line. There are non-iconic elements in this iconic representation. The names for the stops are non-iconic.

In accord with (1), there is a natural, systematic 1-1-into mapping from dots to stations, and from lines to tracks between stations. The natural mapping is spatial. Relative positions of dots and names and the relative compactness of dots are mapped naturally to the relative positions and relative compactness of the stations. The extended nature and linearity of the lines and the relative positions of the lines are mapped to the extended and linear natures of the rails and their relative positions.

In accord with (2), spatial relations among the dots and lines preserve some spatial relations among the stations and tracks. Thus the between-ness relation among the dots preserves the between-ness relation among the stations: the middle dot is between the outer dots, and the Vermont Ave. station is between the USC and Western Ave. stations. On the other hand, distance relations are not preserved under the mapping.

In accord with (3), the map represents relevant spatial relations partly via the mappings cited in (1) and (2).

So the map represents the light-rail line and its stations iconically.

The central notion in the explication (1)–(3) is that of a *natural* correspondence. I have no definition. Paradigmatically, natural relations, including 1-1 mappings, are of the sort that natural science (physics, chemistry, biology, geology, and so on) or mathematical science represents.[5] In the light-rail map example, spatial mappings are evinced as natural by the fact that natural sciences study spatial relations.

Metrical or topological relations in spatial arrangements, relations of intensity among light reflectances or among sounds, temporal relations, relations of greater or lesser size or speed, relations of natural parts to natural wholes, relations of sound pitch or degree of pressure are examples of natural physical relations. The idea of

[5] This list of sciences is paradigmatic, not definitional. I take the notion of naturalness to be intuitive. The key point is that the mappings are not *in themselves* representational or intentional. Natural mappings are close cousins of what Grice called natural meaning. See H.P. Grice, 'Meaning', *The Philosophical Review* 66 (1957), 377–388.

natural mathematical relations is less obvious. Simple operations on the natural numbers (doubling, adding two, dividing by two, factorization) are clear examples of natural operations. Inevitably, what counts as natural in mathematics depends on degree of expertise and amount of background knowledge.

Let me say something about what a natural mapping is *not*. Natural mappings are not *in themselves* representational mappings. (See notes 2 and 5.) Being iconic is a non-representational feature of representation or representational content. When *representation* is iconic, iconicity is an aspect of how a *representatum* is represented. The natural mappings that make representation iconic are prior and independent of whether they are capitalized upon in representation.

Natural mappings are also not established by convention. They exist independently. An iconic representation's being a representation can depend on convention. The use of dots and lines to represent stations and tracks, in the map example, is conventional. The natural mapping relations have been selected conventionally for representational use. But the natural mapping relations that the convention utilizes–the mapping between spatial relations among the dots and lines, and spatial relations among the stations and tracks–are not conventional. They do not result from agreement, or unconscious but non-compulsory coordinations.[6]

A consequence of condition (3) is that an individual that uses an iconic representation (in producing it or in receiving it) must be sensitive to and competent in using the natural mappings. The individual must respond to the mappings "naturally". Thus an individual representer must be sensitive–perhaps unconsciously–to the fact that relations among structural elements of the representation are analogs of relations among some structural elements of the represented subject matter. The natural mapping must not only be, in itself, a non-representational, objective relation. It must be *natural for* users of the mapping–at least natural enough to allow relatively easy use. What is natural for a user can vary with the user's species and learning history.

I will say more about iconicity and naturalness in other work. I turn here to some ways of thinking about iconicity that are seriously mistaken. The views that I criticize are centered in Jerry Fodor's claims about iconic representation.

Fodor maintains, 'it is having a canonical decomposition that distinguishes discursive representations from iconic ones'.[7] Fodor understands compositionality in language, which he takes to be non-iconic, discursive representation, as follows:

> A…representation in L is syntactically compositional iff [if and only if] its syntactic analysis is exhaustively determined by the grammar of L together with the syntactic analyses of its lexical primitives. A…representation is semantically compositional in L iff its semantic interpretation is exhaustively determined by its syntax together with the semantic interpretations of its lexical primitives.[8]

The characterization of decompositionality in language is unobjectionable.

[6] David Lewis, *Convention* (Cambridge: Harvard University Press, 1969).

[7] Jerry A. Fodor, *Lot 2: The Language of Thought Revisited*, *op. cit.*, 173.

[8] *Ibid*, 172.

A *canonical* decomposition is a privileged, correct decomposition. Fodor takes representation to be iconic in that it lacks a canonical syntactic or semantic decomposition. He infers from this view that

(a) iconic representations have no constitutive structure;
(b) constituents of iconic representations are homogeneous ('each constituent contributes in the same way');
(c) iconic representations lack logical forms;
(d) iconic representations lack a distinction between semantic constituents that contribute individuals and constituents that contribute properties;
(e) iconic representations lack truth [or accuracy] conditions;
(f) iconic representations lack ontological commitments;
(g) iconic representations do not impose principles of individuation on domains in which they are interpreted;
(h) iconic representation is not representation-as.[9]

Fodor takes pictures to be paradigms of iconic representation.[10] He mentions graphs. But other than certain psychological states, pictures constitute the only case of iconic representation that he discusses in any detail. Fodor's discussion of pictures is supposed to constrain how one thinks about iconicity in perception, and presumably maps.

Fodor rests most of his reasoning on what he calls 'the picture principle':
PP(1): If P is a picture of X, then parts of P are pictures of parts of X.
Fodor notes that according to PP(1), '*all* the parts of an icon are ipso facto constituents'.

He argues for the principle as follows:

> (ARG) Take a picture of a person, cut it into parts whichever way you like; still, each picture part pictures a person part. And the whole that you have if you reassemble all the picture's parts is a picture of the whole person that the parts of the picture are pictures of.[11]

As far as I can tell, (ARG) derives from Stephen Kosslyn. Kosslyn writes: 'Depictive representations convey meaning via their resemblance to an object, with parts of the representation corresponding to parts of the object. In this case, a "part"

[9] *Ibid*, 174–177. I think that Fodor intends the claim more broadly, to mean that iconic representations lack any structure relevant to being veridical.

[10] *Ibid,* 173.

[11] *Ibid*, 173. The principle and the argument for it are also stated in Fodor's 'The Revenge of the Given', in B. McLaughlin and J. Cohen, *Contemporary Debates in Philosophy of Mind* (Blackwell, Oxford, 2007), 108.

can be defined arbitrarily, cutting up the representation in any way; no matter how you cut it, the part will still correspond to a part of the object…'.[12]

PP(1) is false. The argument for it, based on (ARG), is unsound. The conclusions about the semantics of pictures that Fodor draws from the principle are mistaken. The claims (a)–(h) and the claim that iconic representations lack canonical decomposition are all false.

I start with some small critical points. Though in themselves small, they are connected to deeper issues. In the second sentence of PP(1), Fodor can be charitably taken to mean that each picture part is a picture of a person or person part.[13] This claim is clearly mistaken. Nearly all pictures of persons have parts that picture things that are not persons or person parts. Most pictures of people do not picture them naked. Parts of a picture that picture the buttons on the person's shirt are not pictures of parts of the person. If a part of a picture represents a highlight or shadow on the person's forehead, it does not represent a part of the person.

Further, nearly all pictures of a person picture a background for the person. PP(1) holds that for every picture of a person, the parts of the picture are pictures of parts of the person. But parts of the picture of a person that picture parts of the background do not picture parts of the person.

PP(1) is false for many further reasons. For example, indiscernible micro-parts of the picture–molecules either beneath or on the surface–do not depict anything. Parts of surfaces that result from losses of paint usually do not represent anything.

I lay these problems aside. Analogs of them will return, because they connect to fundamental difficulties. One might think that, so far, I have just shown how careless Fodor has been. One might think that his position can easily be repaired.

One partial repair would be to construct a principle for the whole *scene* that the picture depicts. The point of the repair is to show that every part of the picture pictures a part of the scene. Even highlights are parts of the scene, even though they are not parts of anything else pictured in the scene. Take a realist painting of three real giraffes. One might illustrate the principle by claiming that the top half of the

[12] Kosslyn's idea is expressed in his *Image and Mind* (Cambridge, Mass.: Harvard University Press, 1980), 33; and *Image and Brain* (Cambridge, Mass.: London, 1994), 5. Fodor does not credit Kosslyn. See also Kosslyn's 'Mental Representation' in *Tutorials in Learning and Memory*, J. Anderson and S. Kosslyn eds. (New York: W. H. Freeman and Company, 1984), 105–107. Kosslyn's syntactical and semantical ideas are vulnerable to the same points I make against Fodor's. For further expressions of the Kosslyn idea, see M. Tye, *The Imagery Debate* (Cambridge, Mass.: MIT Press, 1991), 44; D. Braddon-Mitchell and F. Jackson, *Philosophy of Mind and Cognition*: *An Introduction* 2nd edition (Oxford: Blackwell, 2012), where, 179ff., they claim, 'there is no natural way of dividing a map at its truth-assessable representational joints'. Although Braddon-Mitchell and Jackson do not mention Fodor or Kosslyn, they in effect echo the view, basing it on the claim that 'there is no natural *minimum* unit of truth-assessable representation in the case of maps'. They present this view as if it were obvious. I discuss minimality of size toward the end of this article.

[13] Taken literally, the second sentence in (ARG) implies that all parts of a picture depict the person. This view is clearly mistaken. A tiny picture part that depicts the left side of a mole on the person's cheek is not a picture of the person. I assume that here Fodor is simply being careless in his formulation.

picture depicts a part of the scene. This is claimed to be so, even though the top half cuts across the middle of the giraffes' bodies, cuts across trees and their branches, and depicts an amalgam of foreground body parts and background tree parts. One might add, aping ARG, that one can "cut up" the picture any way one likes; and the cut-up parts will picture parts of the scene. Any arbitrary cutting could be reassembled to produce the original picture. One might take this argument to support 'Picture Principle (2)':

> PP(2): Every part of a picture pictures [or represents] a part of the scene that the picture pictures [or represents].[14]

One might take PP(2) to show that pictures lack canonical decompositions and to show (a)–(h).

PP(2) and the argument for it are intuitive for some. But intuitiveness does not vindicate them, or support conclusions about pictures and iconic mental representations that Fodor infers from PP(1). The basic problem for both PP(1) and PP(2) is that the arbitrary representational units (whether primitive or combinations of primitives) that they allow correspond to no units grounded in use and understanding of pictures. Any serious semantics for pictures–like any serious semantics for any representation–must be grounded in representational usage and representational competence. I will try to make this problem vivid by developing it slowly.

Let us be more specific about what a *part* of a picture is. Let us focus, as PP(1) and PP(2) should have, on parts that are on the surface and intuitively relevant to understanding the picture.

There is a notion of a Goodmanian part that includes any aggregate of scattered parts of the picture as making up a part.[15] For example, the part that depicts the upper half of the left-most giraffe's left ear and the part that depicts the right-most third of the highest leaf on the right-most background tree are not contiguous. One Goodmanian part of the picture consists of these non-contiguous picture parts.

Fodor does not rule out Goodmanian parts. His phrase 'cut it into parts whichever way you like' and his talk of 'reassembling' the parts do not stipulate that the assemblies, short of the whole reassembled picture, must be among erstwhile contiguous parts. Whether or not Fodor intended to include Goodmanian picture parts, let us pursue these matters a step further.

PP(1) and PP(2) retain *some* intuitive force on the Goodmanian understanding. The illustrated scattered "part" of the picture can be taken to depict a scattered part

[14] E. J. Green and J. Quilty-Dunn, 'What is an Object File?', forthcoming *British Journal of the Philosophy of Science*. Their principle PP(2) is clearly inspired by Fodor, although they do not present it as a repair of Fodor's mistakes. They argue for PP(2) in ways nearly identical to the way in which Fodor argues for PP(1)–again using Kosslyn's example of cutting up a picture in arbitrary ways.

[15] N. Goodman and H. Leonard, 'The Calculus of Individuals and Its Uses', *Journal of Symbolic Logic*, 5 (1940), 545–55; N. Goodman, *The Structure of Appearance* (Cambridge, Mass: Harvard University Press, 1951; 2nd ed. Indianapolis: Bobbs-Merrill, 1966).

of the scene. Why should one not allow Goodmanian parts to count as parts, in understanding these principles?

It is not credible that so broad a notion of depiction is relevant to a serious semantics. It is not credible that a serious semantics takes arbitrary scattered parts of the picture as representational units. The picture's semantics hinges on its use–and on the psychological competencies, processes, and types of understanding that figure in its production and appreciation. Its use and the associated psychological competencies reside in the perceptual segmentation of pictures, the intentions of the painter, and the conventions of interpretation for realist paintings. Nothing in the use, production, or appreciation of the picture corresponds to, or is explained in terms of, any such unnatural representational units. Usage, production, and appreciation treat the part of the picture that pictures a whole giraffe as a unit. They do not treat as a representational unit the scattered Goodmanian picture part that I cited.

The part-whole relation for pictures that is relevant to representational units, or representational constituents, for a semantics for pictures is constrained. It does not follow off-the-cuff intuitions about the parts of pictures.

Suppose that PP(2) is understood to *exclude* Goodmanian scattered parts. Every part of a part is to be taken to be contiguous to some other part of the part. The same problems remain.

Take a part of the picture whose left side corresponds to a small sliver of a giraffe's right flank, and whose right side corresponds to a melange of a part of a tree trunk, parts of a couple of branches, parts of leaves, and parts of patches of sky behind the foliage. Take the left and right sides of the part to be contiguous. One might find it intuitive that this picture part represents a part of the scene. It is the part consisting of that sliver of the giraffe, that part of the tree trunk, the mix of branch and leaf parts, and the visible parts of sky behind the foliage. One might grant that that is a part of a scene. One might grant that that part gets represented in a rough intuitive sense. But a serious semantics of the picture should not and does not follow such intuitions.

The semantics of the picture hinges on use of the picture and relevant psychological competencies, processes, and understanding. Nothing in the use, production, appreciation, or understanding of the picture corresponds to such "units", or suggests that such units are otherwise consequences of these factors. Perceptual segmentation, intentions of the painter, and conventions of interpretation for realist paintings simply do not cut the painting up in that way. That part of the picture is not a semantical or "syntactical" unit.

Consider the following analog of arguments for the intuitiveness of PP(2)–supporting respectively the Goodmanian and contiguity notions of part. Someone might argue as follows.

> (Goodmanian) The scattered part of the sentence 'The dog nuzzled the cat' that consists of the words 'The dog' and 'the cat' represents the dog and the cat. The dog and the cat, together, make up a part of the state of affairs that the sentence represents. So the part of the sentence consisting of 'The dog' and 'the cat' represents that part of the state of affairs consisting of the dog and the cat. Any combination of any two words or word-combinations

> in a sentence, each of which represents a part of a state of affairs that the sentence represents, is a semantical unit and itself represents a part of that state of affairs. So 'the dog the cat' is a semantical unit that represents a part of the state of affairs.
>
> (Contiguous) The part of the sentence 'The dog nuzzled the cat' that consists of 'The dog nuzzled' represents a part of the state of affairs represented by the sentence. It represents the dog and the nuzzling. These make up a part of the state of affairs that the sentence represents. So the part of the sentence consisting of 'The dog' and 'nuzzled' represents the part of the state of affairs consisting of the dog and the nuzzling. Any combination of any contiguous words in a sentence, each of which represents a part of a state of affairs that the sentence represents, is a semantical unit and itself represents a part of that state of affairs. So 'The dog nuzzled' is a semantical unit that represents a part of the state of affairs.

Naively, both arguments are intuitive. But anyone who knows anything about the semantics of language knows that these are bad arguments. Neither 'The dog the cat' nor 'The dog nuzzled' is a semantical or syntactical unit in the sentence. 'The cat' is embedded in a verb phrase that is independent of 'the dog'. 'Nuzzled' dominates that verb phrase, and is again not a part of any semantical or syntactical unit with 'The dog', except the unit of the sentence. The sentence is built from a noun phrase and a verb phrase. Decomposition of the sentence does not cut across these units. 'Nuzzled' and 'the cat' are embedded in one unit, with 'nuzzled' dominating 'the cat'. 'The dog' is another unit. One cannot mix and match. We know these things via reflection on patterns of linguistic usage, competence, production, and understanding.

Fodor recites such facts about language. But analogous points undermine his claims about pictures. The idea that there are no semantically natural joints in pictorial representation that depict natural joints in the scene has nothing to be said for it. Arbitrary combinations of picture parts can seem naively to represent parts of the scene. But most such combinations correspond to no units that figure in usage or understanding. Usage and understanding ground any serious semantics for pictures. They ground postulating picture parts or aspects that correspond to natural perceptual, intended, or conventionally demarcated units in the scene.

Parts of the picture are involved in representation of a giraffe's body and body parts. Parts of the picture represent each natural constituent of the background and that constituent's parts. Each of these scene parts is represented in the picture–just as nuzzling, the cat, and the dog are each represented in the sentence 'The dog nuzzled the cat'. One can compose arbitrary "parts" of a scene out of such materials. But there is no representational *unit*, short of the whole picture, formed by combination of arbitrary picture parts–anymore than 'The dog the cat' is a representational sub-unit of a sentence. Representational sub-units in the picture correspond to units in usage and understanding. Sub-units that represent the giraffes are thus grounded. Sub-units that represent leaves in the background are thus grounded. But an alleged butchered sub-unit that represents an amalgam of a giraffe's upper half and arbitrary slivers of sky and foliage is not thus grounded. The whole picture parses into representational sub-units grounded in patterns of usage and understanding.

As indicated, some representational sub-units induce a part-whole structure. For example, the picture will have sub-units that represent parts of a giraffe's visible flank's surface, as well as sub-units that represent the whole visible surface. But the picture has no sub-unit that represents an amalgam of a flank part and a part of a background leaf.

Competencies associated with perceptual, intentional, and conventional patterns in making and interpreting realist paintings do not support the idea that every complex part of the picture is a representational unit. Representational units are determined by psychological use, processes, functions, and competencies. Only picture parts that correspond in some way to psychological kinds are constituents.

Similar points apply to maps. Suppose that cities larger than a certain size are represented on a road map by a standard-sized circle. Cut a piece of the map that includes part of one of the circles, and conjoin it with a non-contiguous, arbitrarily chosen piece also cut out from the map. The conjunction forms a Goodmanian part of the map. Alternatively, cut one of the circles in half; then include in the same cut an arbitrary part of the region contiguous with the circle half. In both cases, one would have map parts that intuitively map parts of the terrain. But those parts would not be representational units. Representational units are representations of cities, roads, and spaces between roads–and parts of roads, spaces, and (sometimes) cities.

Lines that represent roads represent, iconically, the roads' length and direction. Spaces between the lines map into spaces between the roads. Line parts map parts of roads. Parts of spaces other than lines and circles map parts of terrain not occupied by roads or cities. But no map usage, perceptual capacity, convention, intention, or representational understanding takes combinations of circle parts and parts of surrounding space as a representational unit, any more than 'the dog nuzzled' is a representational unit. Arbitrary map parts, whether scattered or contiguous, are not expressions of the conventions or the perceptual or conceptual competencies that ground representational content for the map.

Similar points apply to iconic perceptual representation. Perceptual structure is not determined by intentions, conventions, or understanding. It is determined by perceptual processes, functions, and competencies. The representational units in any iconic perception must correspond to natural psychological kinds. For example, the processing of the edge that forms a representation of the boundary of a giraffe's ear is a different process from the process that forms representation of the color, or the size and shape, of the branches behind the giraffe. The computation of a representation of the farther-than relation between the branch-bodies and the giraffe-bodies hinges on forming separate representations of giraffe surfaces and branch surfaces. Representational units mark representational competencies, states, and formation processes. The idea that one can cut representational states and competencies in perception "any way one likes" is out of touch with the computational and kind-explanation practices of perceptual psychology. Representational content marks representational states and competencies. Perceptual states and competencies are explained in perceptual psychology via principles for forming units of representational content. There is massive evidence that perceptual representation is iconic–makes use of natural mappings between the format of perceptual representation and elements in the physical environment.[16]

The master mistake in Fodor's reasoning is methodological. It is the mistake of reasoning to a semantics for pictures from armchair reflection on the format of pictures, rather than from the psychological and conventional capacities that underlie their use. One cannot understand the semantics of *anything* by starting from reasoning about its format–iconic or otherwise. One must begin by reflecting on usage and underlying psychological competencies.

Before criticizing Fodor's conclusions more specifically, I set out the most generic representational kinds that ground semantics for perception, maps, and realist pictures. The representational function of all these types of representation is referential identification.[17] All function to pick out particulars (specific surfaces, bodies, places, and so on) partly through contextual, causal relations to those particulars and partly by discriminating them from other contextually relevant particulars by characterizing them in terms of properties, relations, or kinds.

In all cases, this function is grounded in representational psychological competencies.

These two functions–picking out and characterizing–are fulfilled by the three basic types of semantical primitives in iconic representations that we have been discussing. One type is comprised of *referential applications*. Referential applications are event types whose representational function is to apply the characterizers so as to pick out or refer to particular entities. Referential applications are analogous to specific, referential uses of demonstratives or indexicals. They are individuated through occurrent events. In that sense, they are not freely repeatable.[18]

A second basic type of semantical primitive marks general, freely repeatable competence in referential application. This type consists of *referential schemas* for demonstrative-like or indexical-like application. The repeatable competence to apply 'this' or 'here' on particular occasions is marked by repeatable words 'this' and 'here'. These words do not in themselves refer to anything. They refer through events of referential application. Similarly, referential applications in pictures, maps, and perceptions are exercises of schematic, repeatable referential competencies that are also marked by referential schemas.

[16] Citing and explaining in detail why visual psychology routinely takes visual representations to have the format of a picture-like array would take up too much space for this article. For examples of work that either illustrate or help motivate the approach, see S. Murray, H. Boyaci, and D. Kersten, 'The Representation of Perceived Angular Size in Human Primary Visual Cortex', *Nature Neuroscience* 9 (2006), 429–434; M. Silver and S. Kastner, 'Topographic Maps in Human Frontal and Parietal Cortex', *Trends in Cognitive Sciences* 13 (2009), 488–495; T. Poggio, 'The Computational Magic of the Ventral Stream: Towards a Theory', *Nature Precedings* (2011), http://dx.doi.org/10.1038/npre.2011.6117.1.

[17] Not all iconic representations represent particulars. Some graphs represent only correlations among properties. Such iconic representations lack referential applications.

[18] For more discussion see my 'Five Theses on *De Re* States and Attitudes' in J. Almog ed. *The Philosophy of David Kaplan* (Oxford: Oxford University Press, 2009); *Origins of Objectivity*, *op. cit.*, 83–84.

The third type of semantical primitive is comprised of *attributives*. Attributives are representational contents that are kinds of abilities, states, or events that function to indicate a property, relation, or kind, and to attribute it to entities referred to by the referential applications of attributives. Attributives are characterizers. When they are applied, attributives function to characterize entities picked out by the referential applications.

For example, a visual perceptual state can single out two surfaces, characterize them as surfaces (applying the attributive surface), and characterize one as farther than the other (applying the attributive farther than):

(that x_1)(that x_2)(ego-here$_3$)[Surface(x_1)Surface(x_2)Farther-than(x_2, x_1, ego-here$_3$)]

Read: that$_1$ surface farther away than that$_2$ surface from ego-here$_3$.[19] The subscripts stand in for referential applications. 'that x' and 'ego-here' stand in for referential schemas. 'Surface' and 'Farther-than' stand in for attributives. In perception, and in a picture, the attributives (unlike the linearly ordered, convention-dependent language) occur in an iconic format, mapping naturally to a subject matter. Similarly, a map can refer iconically to three stations, characterize them as stations (with iconic markers, and perhaps, but not necessarily, also with words), and characterize one iconically as between the others.

All iconic representations that we have discussed– perceptions, pictures, maps– that function to represent particular entities have a noun-phrase-like representational structure. The structure is scope- dominated by one or more applied referential schemas, including at least one applied demonstrative-like schema. The applied referential schemas apply one or more attributives.

Thus, in the blocked off illustration, the applied referential schemas are (that x_1), (that x_2), and (ego-here$_3$). There are two applied demonstrative-like referential schemas–(that x_1) and (that x_2). The applied referential schema (ego-here$_3$) is indexical-like, rather than demonstrative-like. As noted, there must always be at least one applied demonstrative-like schema in every perceptual state, in every map, and in every picture that functions to represent a real subject matter. In the illustration, the applied demonstrative-like referential schemas function to pick out a surface, each a different surface. The applied indexical-like referential schema functions to pick out a place. These three applied schemas scope-dominate the whole structure–insuring that the whole structure is noun-phrase-like, not propositional. The applied demonstrative-like schemas each apply the attributive Surface.

[19] Ego-here$_1$ is an application of an *ego-centric index*, marking the position of the perceiver. A spatial ego-centric index marks the origin of a spatial mapping from a perception to spatial structures, and does so in a way that privileges the origin as being of special psychological ("ego") significance. Nearly all perception contains such applications of spatial or temporal ego-centric indexes. See *Origins of Objectivity*, *op. cit.*, 187, 199, 287. Most commercial, paper maps lack ego-centric indexes and are *allocentric*. They map space in a way that is independent of the position of the map or the map's user. Many allocentric maps still have origins established by referential applications. See note 25.

The applied indexical-like schema, (ego-here$_3$), applies the attributives Place and ego. The three referential schemas jointly apply Farther-than.

The identificational function of iconic representations is embodied in the demonstrative-governed noun-phrase-like structure. The relevant iconic representations constitutively have an identificational function. Noun phrases in language are usually not iconic. But many share the abstract representational structure just articulated. This noun-phrase-like structure first arose in perception, which pre-dates language.

I believe that the points just made about the representational constituents and representational structure of realist pictures, maps, and perception are apriori. They derive from reflecting on the representational functions and competencies involved in these types of representation. In the case of perception, the practice of perceptual psychology accords with the account. The science of perceptual psychology takes perceptual states both to refer to particulars and to characterize them. Specific representational units with these functions are delineated empirically. The science postulates no propositional states.

Present purposes do not demand explaining this structure in detail.[20] The important point here is that any such noun-phrase-like structure has a canonical decomposition, no less in iconic representation than in non-iconic language. The structure decomposes into its constituent referential applications, referential schemas, and attributives. Complex attributives decompose into their constituents. Farther-than is almost certainly primitive. Surface (like edge) is probably also primitive. There are interesting issues here about how representations of surfaces and edges relate to representation of surface- and edge-parts, issues to which I shall return. Perceptual primitives are determined not by intuitions, but by science's discovery of basic perceptual competencies. Both primitives and complexes are non-arbitrary representational units.

Since I am less interested here in elaborating a semantical analysis than I am in discrediting the arguments that attempt to show that there is no semantical structure in iconic representations, I do not expand on these remarks. The main point is that realist paintings, maps, and perceptual states function to identify their *representata* by picking them out referentially, and to characterize them by indicating and attributing attributes. The attributives function to characterize particulars, which are picked out occurrently and contextually. Representational units are not arbitrary, but correspond to psychological competencies.

Let us return to (a)–(h), the remaining claims inferred from the false principles PP(1) and PP(2):

(a) iconic representations have no constitutive structure;
(b) constituents of iconic representations are homogeneous ('each constituent contributes in the same way');
(c) iconic representations lack logical forms;
(d) iconic representations lack a distinction between semantic constituents that contribute individuals and constituents that contribute properties;

[20] I do so in a coming book, tentatively titled *Perception: First Form of Mind*.

(e) iconic representations lack truth [or accuracy] conditions;
(f) iconic representations lack ontological commitments;
(g) iconic representations do not impose principles of individuation on domains in which they are interpreted;
(h) iconic representation is not representation-as.

Contrary to (a), (b), and (d), the constituents have constituent structure, in that they are typed as singular representations applying attributive representations. Contrary to (a), (b), and (c),[21] they have the grammar-like and semantical forms of contextual determiners dominating attributives. Contrary to Fodor's understanding of (e), they have accuracy conditions. Accuracy conditions are instances of this general scheme for pictures, maps, and perceptual states: If every singular referential application in a given representation picks out a particular and every attributive is accurate of particulars to which they are attributed, the representation is accurate. Otherwise, it is not accurate. It is obvious that realist pictures, maps, and perceptions–paradigms of iconic representation—can be accurate or inaccurate.

Where iconic representation is appropriately committal, it carries "ontological commitments"–contrary to (f). Perception is *committal*, in that it presents the world as being a certain way, and undergoes a certain sort of failure–a representational failure–if the world is not as the perception represents it.[22] Whereas perception is constitutively committal, maps and realist paintings are not. They can be presented whimsically or as fictions. But most maps, and paintings presented as depicting real entities are committal.

Fodor does not explain 'impose principles of individuation'. On any normal construal, contrary to (g), some iconic perceptual states pick out bodies as bodies, and are capable of tracking them over time. Picking out and tracking bodies as such requires operating according to principles that determine when bodies are the same and when they are different.

Contrary to (h), iconic representation is representation-as. *Every* attribution to a particular in a realist painting, map, and in a perceptual state, is a form of representation-as.[23] Attributions are characterizations. Characterizations are representations-as. Attributions just *are* forms of representation-as.

[21] Fodor does not explain his notion of 'logical form'. See note 9. I regard logic as an account of propositional validity by virtue of propositional structure. Pictures and perceptions are not propositional. Regardless of how one uses the term 'logical form', there are certainly forms of pictorial and perceptual representation that have veridicality conditions and a semantical structure, together with something analogous to a grammar.

[22] *Origins of Objectivity*, *op. cit.*, 74–75.

[23] For detailed discussion of attribution and representation-as in perception, see *Origins of Objectivity*, *op. cit.*, 379–381; 'Origins of Perception', *Disputatio* 4 (2010), 25–28. Fodor makes the further fundamental error of conflating information registration with genuine representation. He calls both 'representation'. He assimilates all iconic representation to information registration. See note 2. For detailed discussion of the distinction, see *Origins of Objectivity*, *op. cit.*, chapter 8; 'Origins of Perception', *op. cit.*, 2–5; 'Perception: Where Mind Begins', *Philosophy* 89 (2014), 385–403; reprinted in T. Honderich ed. *Philosophers of Our Times* (Oxford: Oxford University Press, 2015).

I noted the master methodological error underlying PP(1) and PP(2). It is that of inferring semantical structure directly from iconic format, instead of inferring it from underlying use and representational competencies.[24]

[24] Fodor argues that some psychological states, for example those in what he calls an 'echoic buffer', are non-perceptual iconic representations with semantic content. He applies (a)–(h) to such states. I will not discuss this argument in detail. But it is as off-hand and unsound as the main argument that I have discussed. To begin with, the argument confuses information registration and representation. (See notes 2 and 23.) Although Fodor does not cite examples of what he means by an "echoic buffer", one can assimilate what he says about it to information-registrational states like the first registration of the retinal image. The first registration of the retinal image is iconic, but it is not representation. His claim ('The Revenge of the Given', *op. cit.*, 113) that such states must have semantical content because categorization (attribution) is extracted from them is clearly mistaken. Perceptual categorizational attribution is extracted from the initial registration of a retinal image. But that registration lacks semantical content. It is purely information registration.

Fodor argues that an "echoic buffer" is not subject to the "item effect". The *item effect* is 'the rule of thumb that, all else being equal, the "psychological complexity" of a discursive representation (for example, the amount of memory it takes to store it or to process it) is a function of the number of individuals whose properties it independently specifies', *Ibid*, 110–111.

Fodor does not explain 'discursive' clearly, but his explanation, *Ibid*, 107, takes discursive representation constitutively to have all the properties (a)–(h) that he denies of iconic representations. I believe that his accounts of iconicity and discursiveness are both defective. If Fodor's description of the "item effect" were correct, one would expect perceptual representation as well linguistically expressed conceptual representation to show it in memories of such representation. Then Fodor's argument would divide non-representational, information-registrational states (registration of the retinal image, the "echoic buffer")–which do not show an "item effect"–from representational states–both iconic and non-iconic, both perceptual and conceptual–which do. The argument would then fail to bear on the distinction between iconic and non-iconic psychological states.

But the argument has yet further defects, scientific defects. Limitations on memory, even in retaining complex representational states, including perceptual states, vary with the type of memory, not just the representational complexity of the state. Certain types of very short-term, iconic memory retain virtually the full complexity of perceptions' representational content. M. Coltheart, 'Iconic Memory and Visible Persistence', *Perception and Psychophysics* 27 (1980), 183–228. Certain types of unconscious iconic long-term memory are also virtually unlimited in their capacity to retain the complexity of perceptual states or of beliefs formed most directly from perception. Cf. T. Brady, T. Konkle, G. Alvarez, and A. Oliva, 'Visual Long-term Memory has a Massive Storage Capacity for Object Details', *Proceedings of the National Academy of Sciences of the United States of America* 105 (2008), 14325–14329. Even visual working memory, the original poster child for the item effect, is not limited in the way that Fodor assumes. Other factors besides the number of items and representational complexity determine even visual working memory's limitations. For reviews and explanations of why the item effect is not a basic explanatory notion, see C. R. Sims, R. A. Jacobs, and D. C. Knill 'An Ideal Observer Analysis of Visual Working Memory', *Psychological Review* 119 (2012), 807–830; D. Fougnie and G. Alvarez, 'Object Features Fail Independently in Visual Working Memory: Evidence for a Probabilistic Feature-store Model', *Journal of Vision*, 11 (2011), 1–12; G. Bae and J. Flombaum, 'Two Items Remembered as Precisely as One: How Integral Features Can Improve Visual Working Memory', *Psychological Science* 24 (2013), 2038 –2047; K. Hardman and N. Cowan, 'Remembering Complex Objects in Visual Working Memory: Do Capacity Limits Restrict Objects or Features?', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41 (2015), 325–347; T. Brady and G. Alvarez, 'Contextual Effects in Visual Working Memory Reveal Hierarchically Structured Memory Representations', *Journal of Vision* 15 (2015), 1–24. Fodor's argument that there are iconic representational states in perception is laced with both conceptual and scientific errors.

There is a corollary error–that of identifying representational constituents too closely with parts of a picture, where 'part' is understood in a commonsense way that is not guided by serious semantical reflection. I have criticized versions of this error that take parts of the picture to be representational units, when they are not. The Fodor-Kosslyn claim that one can cut a picture any way one likes embodies this error. Another version of the error is to assume that every representational constituent is a part of the picture, in an unqualified, intuitive sense of 'part'.

Recall some basic points about the relation between parts of *sentences* and semantical constituents. I stipulate that letters and parts of letters are not parts of sentences. Let us assume that words or morphemes are basic parts of sentences. Some parts, so understood, are not representational units. Words embedded in some idioms are not semantic constituents of sentences. Conversely, some representational constituents are not morphemes, words, or word combinations. In context-dependent uses of sentences, representational constituents include occurrent, contextual, referential applications. These are the referential events that constitute occurrent uses of demonstrative-governed phrases. No word or symbol, in the language, expresses or stands in for the occurrent, referential application.

Words like 'that' are schemas. Such words do not represent, demonstratively, any given entity. Occurrent use is needed to carry out such representation. The occurrent event of referential application, not the word itself, is the semantical unit that is central to referential representation. There is no separate part of the sentence that is specific to the occurrent event of application. Ultimately, language *must* rely on applications that are not themselves terms or parts of sentences. They are not themselves symbols. They are events of application.

The same point applies to pictures and perceptual states.[25] In a picture, reference to a particular is effected by an occurrent use. Singular reference to a particular is effected by the intentional act of putting a picture part onto the canvass or interpreting the picture, not by any picture part, token or type, taken on its own. The same colored shape could have represented a different particular, or no particular at all, in a different context. There are no two picture parts, one of which indicates a repeatable color property and another refers to a concrete instance of the property. In a perceptual state, reference to particulars is effected by occurrent events that instantiate a repeatable representational ability-type. No separate symbol or part of

[25] Allocentric maps can avoid context-dependent referential applications of the maps' spatial origin. But many allocentric maps use ordinary proper names, which do involve context-dependent determiners. And even non-context-dependent, canonical names must, in the end, be explained in terms of context-dependent referential applications. Ego-centrically anchored maps all involve referential applications to the "home" anchor position.

the content, of either a picture or a perceptual state, distinguishes the occurrent application from the representational, attributive type(s) that it applies.[26]

Depiction always involves a combination of occurrence-based singular application and a characterizing attribution. The intuitive notion of depiction is not a semantical primitive–at least in cases of pictures that pick out particulars. Similarly, for maps and perceptual representation. All perception is via combinations of singular applications and attributions.

A central respect in which parts in an iconic representation do not correspond to constituents resides in representation of *relations*. In a picture no *part* of the picture specifically represents the depth relation between an object in the foreground and the background. Yet the picture pictures the foreground object as in front of the background.

It is a mistake to identify a specific part of the picture that serves as a representational constituent that represents *any* relation that a picture depicts. If one object is depicted as to the left of another object, with some distance between them, there is no answer as to what *part* of the picture specifically represents the relation *to the left of*. The spatial relation is depicted, but no part of the picture corresponds specifically and proprietarily to the space between the entities. The part of the picture between

[26] Both Fodor in *Lot 2: The Language of Thought Revisited*, *op. cit.*, 175, and J. Quilty-Dunn in ‘Iconicity and the Format of Perception’, *Journal of Consciousness Studies* 23 (2016), 255–263, take a *term* that symbolizes the singular reference to be required, if an iconic representation (the whole perception or picture) is to represent a particular. Quilty-Dunn writes, ‘…icons lack the representational apparatus to bind features by picking out an object and attributing those features to the object’ (261). But that is exactly what paradigmatic icons, like realist paintings of individuals and perceptions, do.

Green and Quilty-Dunn, ‘What is an Object File?’, *op. cit.*, make much of the supposed non-iconicity of memory index files. Anaphoric applications in memory–in index files—that derive from referential applications in perception may or may not have a ‘symbol’ that effects the anaphora. There being such a symbol would not prevent perceptual memory from being iconic.

Indeed, if symbols occur in relations that bear natural correspondences to relations in a subject matter, the arrangement of symbols is iconic. Note the iconic representation of positions of light-rail stations by the arrangement of their names. As the light-rail map example indicates, many iconic representations have non-iconic symbolic elements. Moreover, being a symbol does not preclude being iconic, as hieroglyphs and some Chinese words show. But such a symbol is not needed in perceptual memory that relies on anaphoric retention of perceptual singular reference, any more than a symbol for a referential application is needed in perception, or indeed natural language. Anaphora can be effected through occurrent events that are causally and functionally connected to the occurrent referential application event involved in the original perception. Such a causal link can underlie the changing iconic perceptual attributives that support the application's use in the index file. There need be no symbol for the referential application in the index file. But since iconic representations can have symbolic elements that are non-iconic, the presence of such a symbol in index files would not prevent such files from being iconic.

A significant error in Quilty-Dunn's article is the mis-attribution to me in *Origins of Objectivity*, *op. cit.*, of the view that the difference between perception and cognition consists (sometimes he says ‘partly consists’, sometimes he does not) in perceptual representations' being iconic, and in cognitive representations' being discursive, or language-like. I think that some cognitive, even propositional, representations are iconic. In fact, I think that *all* propositional beliefs immediately formed from perception are iconic.

the parts that represent the two entities represents other particulars–a background wall, for example. That part is not specific or proprietary to the relation *to the left of*. No picture-part (scattered or not) is a representational constituent specific to the relation. The relation is represented through the relations among the parts, not through a symbol for the relation. The relation *to the left of* is depicted by situating the part of the picture that represents one object to the left of (from the viewer's perspective) the part of the picture that represents the other object. Language has a separate symbol for the relation *to the left of*. The picture does not.

Of course, the iconic arrangement in realist paintings never does the representing all by itself. The arrangement must derive from appropriate uses and competencies.

Analogous points apply for perception and maps. Trying to locate, in non-representational terms, the *part* of a perceptual state, or the underlying neural state–or the part of a map–that indicates and attributes even a simple spatial relation is a quixotic enterprise. The science of visual psychology has suggested no analog for a single symbol (a semantical or syntactical constituent) that in language commonly would indicate a relation and, in the context of a sentence, attribute it. Computations operate on representations' *being* in certain relations. Nothing in computational theory of perception requires that there be a symbol for every relation that is computed.

A simple identification of picture parts with semantical constituents does not work even for iconic indication and attribution of non-relational properties. If asked what part of a picture represents the color, or size, or shape of a surface, it can seem obvious that one can draw a boundary around a relevant picture part. It is intuitive that that part represents the color, size and shape of the surface, as well as the surfacehood of the surface. But this answer is only good enough for off-hand remarks about the relation between parts and semantic constituents. It is the *color* of the picture part that figures in representing the color in the scene. It is the scaled *shape* or *size* of the part that figures in representing the shape or size of the surface. None of these properties of a part of the picture is itself a part of the picture. Each is an aspect of a part.

A similar point applies to visual perception. Different aspects of the perceptual content represent different attributes. Some sensitivity to light intensity, reflected in a natural way in intensity of neural activity, figures in iconic perceptual representation of color. Some space-like arrangement, within a sensory-registration state or perceptual state, of the effect of neural firings figures in the iconic representation of shape. Computations work as well on properties or relations as on object-like "parts". Privileging parts is the effect of thinking of computation in perception too much on an analogy with linguistic computation.

In realist pictures, maps, and normal visual representation, properties are represented in packages. A realist picture and a normal visual perception represent a surface's color, size, shape, orientation, location, and so on, all at once. A road map does not represent a road without representing its position, length, and shape. Such packaging is a normal feature of many types of iconic representation. By contrast, a language user routinely represents a surface without representing its size or color, or a road without representing its position or length. Which properties one talks about is a matter of choice. The packaging-of-properties aspect of some types of perception is a *type* of iconicity. It is not essential to iconic representation. A color-chit represents a color-shade iconically. It normally represents no other properties. Packaging is not even essential to perception. Certain pathological types of perception may iconically represent color without representing shape or any other spatial attribute.

Even though packaging is usual in iconic representation, loss of some unit in the package, or even depackaging, does not in itself undermine the iconicity of a representation. A drawing could have its color removed and remain an iconic drawing. A map's use of larger circles for larger cities could be discontinued, and the map would remain iconic. Review of the explication of iconicity near the beginning of the article supports this point. The point is also intuitive.

Analogous points underlie the obvious fact that some aspects of an iconic perceptual representation may be better retained than others in perceptual memory, even though the memory is iconic, and even though the different aspects originally came in a "packaged" iconic perceptual representation.[27] Even the complete loss to

[27] E. J. Green and J. Quilty-Dunn, 'What is an Object File?', *op. cit.* state that PP(2) and the following principle are 'the signature markers of iconic format': (H) (for Holism) 'Each part of the presentation represents multiple properties at once, so that the representation does not have separate vehicles corresponding to separate properties and individuals'. Their exposition follows Fodor in conflating information registration and representation. Further, not all iconic representation represents multiple properties. So packaging is not a signature marker of iconic format. As noted, a color chit can be used to represent a color shade, and nothing else. As I indicated, it is problematic to claim that when multiple properties are represented in a package, the representation goes by way vehicles. If vehicles are (say, picture) *parts* or object-like entities, then it is not true that a single vehicle represents multiple properties. The parts are not, strictly, the representing units. If the properties or property-instances count as vehicles, then different "vehicles", not one, effect representation of different properties: the color of the picture part represents the color; the shape of the part represents the shape (scaled appropriately); and so on. In perception, different aspects of a perceptual state represent different environmental attributes, even though properties are represented in a package. So (H) is mistaken in various ways.

Green and Quilty-Dunn argue that since perceptual working memory sometimes retains different properties to different degrees, perceptual working memory cannot conform to (H). They conclude that perceptual working memory is not iconic. They further infer from considerations of simplicity that perceptual representation is not iconic either. Even apart from the mistakes in PP(2) and (H), this train of reasoning seems to me a *reductio* of their conception of iconicity and their conception of how representation takes place in pictures and perception. The fact that various properties are iconically represented in a package in perception does not begin to show that they cannot be remembered iconically to different degrees.

memory of some aspects of a packaged group of representations is compatible with the memory's remaining iconic.[28]

Iconic perceptual memory could, for example, retain iconic representation of shape better than color. A visual perception normally represents size, 2-D and 3-D shape, color, orientation, location, surfacehood, bodihood, all as a package. But since these properties are represented by different aspects of the perceptual state, and some derive from different, dissociable formation processes, remembering these properties to differing degrees, or even forgetting some, does occur. Such occurrence bears not at all on the memory's remaining iconic. Iconicity depends on there being one or more natural mappings. Even when the memory of one property in a package is weaker than that of another property, there remain natural correspondences between the aspects of the memory state that represent the properties and the properties themselves. Iconicity in memory does not depend on memory's preserving all aspects of pictorial representation.

Normal visual perception represents its basic attributes in packages. Doing so is an aspect of the iconic nature of normal visual perception. It is a way in which visual perception is normally picture-like or map-like. But neither perception nor iconicity is essentially picture-like or map-like in this respect. As noted, a visual perception and visual memory could be of a single unlocated color shade, with no packaging and no spatial representation. They would remain iconic, as a color chit is. Auditory perception is iconic but not at all picture-like. The etymology of 'iconic' is connected to visual imagery. But standard dictionary meanings of the term generalize beyond the visual.[29] The connection between iconic representation and pictures or maps is real and paradigmatic. It is not constitutive. What is constitutive is natural mapping from some units of representation to corresponding entities in the subject matter.

Let us return to the idea that because one can cut up a picture into small parts each of which represents something, the semantics of a picture is arbitrary. It is sometimes, though not always, correct to think of pictures, maps, and perceptions as

[28] There is evidence that iconic perceptual working memory in fact retains different features to different extents, as one would expect. Packaging is one factor in retention. But many factors bear on how well different iconically represented attributes are retained in memory. Attention might affect different attributes differently. Facts about how different properties are registered differently in neural coding can also ground differential retention. Relationships among the types of properties retained can affect retention. For papers that bear on these issues, see M. Wheeler and A. Treisman, 'Binding in Short-term Visual Memory', *Journal of Experimental Psychology: General* 131 (2002), 48–64; Y. Jiang, M. Chun, and I. Olson, 'Perceptual Grouping in Change Detection', *Perception & Psychophysics* 66 (2004), 446 – 453; D. Fougnie and G. Alvarez, 'Object Features Fail Independently in Visual Working Memory: Evidence for a Probabilistic Feature-store Model', *op. cit.*; G. Bae and J. Flombaum, 'Two Items Remembered as Precisely as One: How Integral Features Can Improve Visual Working Memory', *op. cit.*; K. Hardman and N. Cowan, 'Remembering Complex Objects in Visual Working Memory: Do Capacity Limits Restrict Objects or Features?', *op. cit.*; T. Brady and G. Alvarez, 'Contextual Effects in Visual Working Memory Reveal Hierarchically Structured Memory Representations', *op. cit.*.

[29] Here is a definition from Merriam-Webster: a sign (such as a word or graphic symbol) whose form suggests its meaning.

containing smallest representational parts–pixel-like representations. It is easy to infer, mistakenly, that the semantics of pictures has no structure. Even if a realist picture is built up from small, primitive, representational parts, it does not follow that the parts can be combined in arbitrary ways. Representation depends on usage, and usage never evinces or allows wholesale arbitrary combinations among representations.

A further point is even more important. Presence of smallest, pixel-like representational parts in a picture, map, or perception does not imply that those parts are representationally primitive. Often the representational content of smallest-parts depends on the representational content of larger wholes. In such cases, pixels need not be primitives. For example, a pixel may have the content: part, in such and such a location, with such and such size and shape, with such and such color, of such and such surface. Such characterizations are consistent with a picture's having compositional, representational structure. (Of course, this characterization is intended as the non-iconic, linguistic counterpart of an iconic characterization.) I think that they often correspond to what is fundamental psychologically. Small parts of an edge are often represented only *as* such parts. So a representation's being analyzable into small representations of small parts (each with a potential use), where the representations occur in a part-whole structure that maps naturally into a part-whole structure in the *representata*, does *not* imply that the pixel-like representations of small parts are representationally, or semantically, primitive.

For example, smallest useable/discernible iconic representations of very small parts of a line, which certainly occur in maps, pictures, and visual perception, need not be primitive representational units. They need not be basic building blocks for composing representations of longer lines. Rather, representations of longer lines with natural end points can be representationally primitive. They accommodate iconic representations of smaller, contained line segments, which lack any natural endpoints, by representing them as parts of the "natural" longer line. Each represented part is distinguished by its represented position within the longer line.

I think that perceptual representation of spatial parts usually works that way. For what is representationally primitive depends on what representational competencies are fundamental. And perceptual competence to distinguish lines with natural endpoints is certainly more fundamental than competence to distinguish the various parts that lack natural boundaries. Representing spatial parts of natural wholes by representing them as parts of, or cuts in, the wholes is compatible with the known fact that, in visual perception, representations of lines with natural endpoints are formed through combining non-perceptual information registrations of smaller edge segments.[30] One must distinguish *information registration* of very small edge segments in proximal stimulation (in the visual image) from *representation* of very small parts of environmental lines with natural end-points. The former are probably

[30] See W.S. Geisler, J.S. Perry, B.J. Super, and D.P. Gallogly, 'Edge Co-occurrence in Natural Images Predicts Contour Grouping Performance', *op. cit.*; J. Frisby and J. Stone, *Seeing: The Computational Approach to Biological Vision*, *op. cit.*, chapter 6.

first in the order of perception formation. The latter are derivative in the structure of perceptual representation.

Iconic representation occurs in language and thought, as well as maps, pictures, and perception. Language sometimes incorporates pictorial elements. Propositional thought often incorporates perceptual elements. Although pictures and perceptions are not propositional, their iconic formats can be coopted in propositional structures, making units in those structures iconic. The most basic perceptual beliefs, for example, are iconic. The popular view that iconicity is disjoint with propositionality is mistaken.

Although most iconic representation is iconic partly in its relational structure, iconic representation does not essentially represent by way of relational structure: one could invent an iconic word for red that consists in a red color chit.

Iconic representations, like all representations, represent by being the product of a use and competence. Representational structure marks the representational functions of the psychological competencies of the users of representation. Representational structure is found by investigating the uses by and competencies of those who use the representations. Iconicity is an aspect of representational format, not a determiner of basic semantical functions. The basic semantical functions of iconic representation–reference and attribution–are shared with non-iconic representations in language and in non-iconic thought. Many, but not all, linguistic modes of presentation and representational content differ from iconic modes of presentation and representational content. Modes of presentation and representational content mark types of competence. Iconic and non-iconic competencies are psychologically different. But basic semantical functions and basic semantical or representational structures are shared.

What is distinctive of iconic representation is that it relies on natural relations between the representation and the represented subject matter. This reliance is evolutionarily very old. It is where representation begins. Representation begins in nature's stamping itself directly into a creature's capacities. The structural correspondences that result from the stamping–inasmuch as they are law-like and reliable–contribute not only to primitive representation, but also to primitive ancestors of epistemic warrant and knowledge. The effects of this stamping remain in language and abstract thought, which in many ways have outgrown iconic representation.

Chapter 6
Scientific Realism in the Post-Kuhnian Times

–Beyond Structuralism and Historicism

Tian Yu Cao

Introduction: Kuhn's Challenges to Scientific Realism

Scientific realism is a special position in philosophy of science. According to a naive version of it, unobservables in a mature scientific theory—no matter these are about the real essences of things, or about the causal agents and the hidden mechanisms of the world—are only the representations of what exist in the objective (mind-independent) world, existing in the way just as the theory describes as they should be. Thus what a mature scientific theory provides is the true knowledge of the objective world, and the rationality of scientific developments simply lies in the accumulation of objective knowledge. Thus, according to this version of scientific realism, the sole basis of objectivity and rationality of scientific enterprise is the correspondence between the representative knowledge and the objective world to be represented.

Almost all practicing scientists, and a large number of philosopher of science too, take this version of scientific realism for granted. Yet, philosophically, it is vulnerable to close examinations. Beginning in the 1950s, the notion of theory-ladenness of data, which deprives data of its innocence and authority in dictating the theoretical statements, and the notion of underdetermination of theory by data, which assigns the theory (theoretical statements and theoretical entities) an inescapable conventional status, became increasingly influential. Both notions have removed possibility for theorists to have any direct and reliable access to reality, and thus made any attempt to take unobservables in theoretical science as true representations of what exist and happen in the world extremely dubious.

The idea of conventionality rather than veracity of theoretical entities was captured by Rudolf Carnap's notion of linguistic framework: meaning and truth of

T. Y. Cao (✉)
Department of Philosophy, Boston University, Boston, USA
e-mail: tycao@bu.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_6

any scientific statement make sense only within a linguistic framework. Carnap (1934, 1950) A hidden assumption underlying this notion is that experiences and knowledge are possible only through the mediation of a linguistic framework, nobody has direct access to reality. The constitutive role of linguistic framework in knowledge production betrays its affinity with Kantianism. However, different from Kant's a priori constitutive (and regulative) principles, which are timeless, immutable, universal and absolutely apodictic certain, and in line with the neo-Kantian notion of relativized a priori principles, Carnap's notion of linguistic framework laid down the conceptual foundation for framework-relativism. This relativist position was further strengthened by his principle of tolerance, which endorses the framework pluralism and was justified by his general logical empiricist stance, rejecting any judgment external to a linguistic framework, such as those privileging one framework over another, as metaphysical and meaningless, thus has to be expelled from scientific and philosophical discourses.

Without noticing the step already taken by Carnap, Thomas Kuhn in the 1950s and early 1960s reached a similar position through a very different route, namely through an innovative examination of certain period of history of physical sciences. (Kuhn 1962) Kuhn's notions on paradigm, on scientific revolutions as paradigm-shifts, and on incommensurability are quite well-known, and thus require no more than a few brief remarks for the purpose of this paper.

First, Kuhn's notion of paradigm, similar to Kant's a priori constitutive principles, neo-Kantian's relativized a priori constitutive principles, and Carnap's linguistic framework, functions as a constitutive framework (CF for short), both as a necessary precondition for scientific knowledge and for specifying the space of conceptual possibilities in scientific explorations.

Second, the distinctive feature of Kuhn's work is his historical sensitivities. Scientific knowledge is always produced in a historically constituted CF,[1] which, in turn, facilitates, conditions and constrains its production. Highlighting the historicity of knowledge production has emancipatory potential, freeing us from dogmatism: if ideas, concepts and norms taken for granted today are in fact the product of historical conditions rather than natural truths, then any claim for them to have unchallengeable authority is groundless. The historicity thesis also raises the task for properly understanding the transition from one set of historical conditions to another for knowledge production, whose significance will be briefly explored in section "SHASR—Beyond Historicism and Structuralism".

Third, Kuhn's historical relativism, incarnated in his incommensurability thesis phrased in terms of paradigms before and after a paradigm-shift, although persuasive and influential due to its being underpinned by historical facts, conceptually is only a special case of Carnap's framework-relativism if we remember that both paradigm and linguistic framework are just different forms of CF mentioned above. Thus the trans-paradigmatic rationality issue facing Kuhn is similar to the

[1] How a CF is historically constituted will be addressed below, although Kuhn himself did not have proper conceptual resources to address it.

metaphysical issue of privileging one linguistic framework over another facing Carnap, and thus has to be addressed by going beyond their common framework of CF, and seeking underlying rationales constraining the transition from one CF to another. The task is much easier in Kuhn's case of historically generated paradigms than in Carnap's case of arbitrarily defined and chosen linguistic frameworks.

Fourth, in addition to historical relativism, the devastating implications of the incommensurability thesis against scientific realism is also clear and heavily exploited by all stripes of anti-realists. Kuhn (1970) claims that "I can see no coherent direction of ontological development" in the history of science. If fundamental ontology of physics, which is assumed by realists to be the representation of the fundamental ontology of the physical world (the foundation of what exist and happen in the physical world), has undergone radical changes over centuries, from Aristotle's natural places in his finite cosmos to Newton's forces in his infinite universe, to Einstein's cosmology dictated by his gravitational fields, with each incommensurable to others, Kuhn contested, then how could we take any of them seriously or realistically? Even more serious is the prospect that since there will be no end to scientific revolutions in the future, no fundamental ontology investigated in any theoretical science can survive radical changes, thus no fundamental ontology in any theoretical sciences at present or in the future can have better chance than those in the past. If fundamental ontology in any theoretical science has no referent in the world, then scientific realism is no more than a faith. With the destruction of scientific realism, the grounds for objectivity and rationality in science also become quite shaky.

For the convenience of philosophical analysis, let us reformulate Kuhn's philosophy of science centered round the incommensurability thesis in terms of fundamental ontology rather than paradigm. A paradigm as a CF is a holistic structure which itself is constituted by certain historically available presumptions: metaphysical ones crystallized from entrenched commonsense beliefs and scientific ones from well-established scientific principles. These presumptions stipulate what exist and happen and what could exist and happen, or what the ontology or the content is, in the phenomenal world and representable by scientific statements. In a holistic structure, all are connected, causally or otherwise, and inseparable. It is reasonable to assume that among all that exist (happen) or could exist (happen), some are more fundamental than others in the sense that all others are dependent upon, or at least connected to, those fundamental ones. Let us call those fundamental ones fundamental ontology (FO). It is clear that a specific FO (such as massive particles moving in an absolute space and evolving along absolute time, or a set of quantum entities transiting from one state to another according to a certain quantum dynamic laws) embodies the constitutive principles of a specific CF, and thus is the core of a CF: each CF is in a sense characterized by its ontological commitment to a specific FO, and in another sense is constituted by the latter, which, however, is susceptible to changes, radical or otherwise, in science. The dialectics between CF and its FO renders CF a dynamic framework rather than a static one, and thus opens a space for investigating the rationales underlying its change.

A decisive advantage of the reformulation as compared with the original one is that the debate on the incommensurability thesis and its implications can be analyzed in a more manageable way in terms of the precisely definable notion of FO than the vague and evasive notion of paradigm (or its variations, including lexical structure Kuhn used in the last stage of his career, Kuhn 2000), as we will see in the following sections.

Finally, Kuhn's Kantian connection—in terms of the noumenal world, the phenomenal world, a priori principles [CF], human knowing actions, and the relationships among them—which has attracted more and more attention in recent decades and has been exploited heavily by Friedman (2001, 2009, 2010, 2011, 2012, 2013, 2014) and many others (e.g. Massimi 2008) can be summarized as follows.

First, Kuhn acknowledges the existence of a stable and permanent world which, however, similar to Kant's thing-in-itself, is ineffable, indescribable and undiscussible. Second, for Kuhn, a phenomenal world is constituted, in a neo-Kantian way, by a given lexical structure, a refined version of paradigm Kuhn adopted in his later reflections, consisting of patterns of similarity/difference relations shared by members of a linguistic community, which makes communication possible and binds members of the community together. Third, Kuhn maintains that a structured lexicon, as a CF, is the embodiment of the stable part of human knowledge about the world, which underlies all the processes of differentiation and change, thus is a precondition for describing the world and cognitively evaluating truth claims. As a corollary, here is the fourth point, phenomenal world is changeable if CF has undergone radical changes.

Insofar as the structure of phenomenal world can be experienced and the experience communicated, it is constituted by the structure of the lexicon of the community inhabited in the world. Yet, Kuhn stresses over and again in his later responses to his critics, the world is not constructed by the inhabitants with their conceptual schemes (hypotheses and inferences, etc.). For this fifth note-worthy point, Kuhn gives two arguments. First, phenomenal world is empirically given. That is, people born into a phenomenal world constituted by a lexical structure must take it as they find it: it is entirely solid, not in the least respectful of an observer's wishes and desires, quite capable of providing decisive evidence against invented hypotheses and conceptual schemes which fail to match its behavior. That is, conceptual schemes are conditioned and falsifiable by phenomenal world. Kuhn resolutely rejected radical social constructivism his self-claimed followers advocated, mainly because he was firmly in line with Kant's phenomenal realism. Second, although people with their conceptual schemes are able to interact with the given phenomenal world and change both it and themselves in the knowing-practicing process, what they can change is not the whole phenomenal world, but only some parts and aspects of it, with the balance remaining as before. The result may or may not be the emergence of a new CF from the more tentative schemes, which would constitute a new phenomenal world. What exactly makes "may" rather than "may not", however, is a strikingly dark gap yet to be filled, upon which Kuhn offers no illuminating light. What is also missing in Kuhn's philosophy of science is a clear understanding of the dialectics and dynamic relationships, such as constraining and adaptation, between noumenal world and phenomenal world, to which Kant himself also kept silent.

Thus for Kuhn, both phenomenal world with its structure and the subject with his CF-empowered ability to act are historically situated: they are empirically given and practically changeable. Kuhn's notion of objectivity is entirely constrained by his notions of phenomenal world and the knowing subject, and contains two components: intersubjectivity in terms of community consensus resulted from interactions (communications) among members of community, and the constraints posed upon the consensus by phenomenal world. This notion of objectivity was the historical and conceptual background, both positively and negatively, against which structuralist conceptions of objectivity and thereupon versions of scientific realism, have been developed.

Structural Realists' Responses to Kuhn's Challenges

Since the central argument in Kuhn's challenge to realism is his claim of ontological discontinuity in scientific development, one line of realist response to the challenge pursued by some structuralism-oriented realists is to address the inter-theoretical relations directly in terms of mathematical structures (the involved theories adopted), taking as the form of the world, without involving physical ontology (what physically exist and happen), or the content of the world, as a medium.

It should be noted that in mature theoretical science, the presence and activities of fundamental ontology are the ultimate resource the theory can utilize for describing, explaining and predicting empirical phenomena. The fundamental ontology in a theory can take various categories, such as objects or entities (an extension of objects to non-object physical entities, such as fields) or physical structures, properties and relations, events and processes. But since there is no bare entity without any property, without being involved in relations with other entities, in events and processes, and there is no ontological category that has a free floating existence without being anchored in some physical entity, traditionally more often than not, physics assume some fundamental physical entities as its fundamental ontology, although property, process and other non-entity ontology were also occasionally adopted, such as energy in energetics and process in S-matrix theory.

It is clear that the place and roles of fundamental entities (FE) in a theory is similar to FO in Kuhn's CF. Furthermore, since FO must incarnate in some form of FE, the two levels of Kuhn's incommensurability thesis, at the theory-level in terms of FE and at the CF-level in terms of FO, are closely connected. Thus, although the contemporary debates around structural realism are often narrowly focused on the legitimacy or illegitimacy of ignoring or rejecting the existence of physical objects (entities, fundamental or not fundamental ones) in the structure and change of scientific theories, rather than addressing Kuhn's much wider conception of CF, FO and their radical changes, the bearings of the former on the latter is not deniable.

The realist urge of the structuralist line of responses to Kuhn's challenges to scientific realism, namely his relativism and anti-realism, can be clearly seen in its attempt to establish a cognitive continuity in scientific development through a referential continuity between mathematical structures (involving law-like statements which go beyond knowledge of empirical regularities) used by physical theories at different historical stages, such as those used by Newton and Einstein. The structuralist-holistic nature of the line is exhibited in its replacement of physical entities (with intrinsic, causally effective properties) by formal relations, its rejection of atomistic metaphysics, according to which entities characterized by their essence and intrinsic properties exist independently of each other, and its taking entities as merely the names of images we substituted for what really exists.

Structuralism as an influential intellectual movement of the 20th century has been advocated by Bertrand Russell, Rudolf Carnap, Nicholas Bourbaki, Noam Chomsky, Talcott Parsons, Claude Leve-Strauss, Jean Piaget, Louis Althusser and Bas van Fraassen, among many others, and developed in various disciplines such as linguistics, mathematics, psychology, anthropology, sociology and philosophy. As a method of inquiry, it takes a structure as a whole rather than its elements as the major or even the only legitimate subject for investigations. Here, a structure is defined either as a form of stable relations among a set of elements, or as a self-regulated whole under transformations, depending on the special subject under consideration. The structuralist maintains that the character or even the reality of a whole is mainly determined by its structuring laws, and cannot be reduced to its parts; rather, the existence and essence of a part in the whole can only be defined through its place in the whole and its relations with other parts. Thus the very notion FE grounding a reductive analysis of science is in direct opposition to the holistic stance of structuralism. According to this stance, the empirical content of a scientific theory lies in the global correspondence between the theory and the phenomena at the structural level, which is cashed out with mathematical structures without any reference to the nature or content of the phenomena, either in terms of their intrinsic properties, or in terms of the unobservable entities, FE included.

In the epistemically interesting cases involving unobservable entities, the structuralist usually argues that it is only the structure and the structural relations of its elements, rather than the elements themselves (properties or entities with properties), that are empirically accessible to us. It is obvious that such an anti-reductionist holistic stance has lent some support to phenomenalism. However, as an effort to combat compartmentalization, which urge is particularly strong in mathematics, linguistics and anthropology, the structuralist also tries to uncover the unity among various appearances, in addition to invariance or stable correlation under transformations, which can help discover the deep reality embodied in deep structures. Furthermore, if we accept the attribution of reality to structures, then the anti-realist implications of the underdetermination thesis is somewhat mitigated, because then we can talk about the realism of structures, or the reality of the structural features of unobservable entities exhibited in evidence, although we cannot directly talk about the reality of the entities themselves that are engaged in structural relations. In fact, this realist implication of structuralism was one of the starting points of current interests in structural realism.[2]

In the philosophy of science, structuralism can be traced back to Henri Poincare's idea about the physics of the principles (1902). According to Poincaré, different from the physics of central force, which desired to discover the ultimate ingredients of the universe and the hidden mechanisms behind the phenomena, the physics of the principles, such as analytic dynamics and electrodynamics, aimed at formulating mathematical principles. These principles systematized experimental results achieved by more than two rival theories, and expressed the common empirical content and mathematical structure of these rival theories. Thus they were neutral to different theoretical interpretations, but susceptible to any of them. The indifference of the physics of the principles to ontological assumptions about the ultimate existence was approved by Poincaré, because it fitted rightly into his conventionalist view of ontology. Based on the history of geometry, Poincaré accepted no fixed ontology that is constituted a priori by our mental faculties. For him, ontological assumptions were just metaphors, they were relative to our language or theory, thus would change when the change of language or theory is convenient for our description of nature. But in the transition from an old theory with its ontology to a new one, some structural relations expressed by mathematical principles and formulations, in addition to empirical laws, may remain valid if they represented the true relations in the physical world. Poincaré's well-known search for invariant forms of physical laws has its roots in the core of his epistemology, namely, that we can have objective knowledge of the physical world, which, however, is structural in nature; we can grasp the structures (form) of the world, but we can never reach the ultimate ingredients (content) of the world.

The inclination for structuralism was reinforced by the rapid development of abstract modern physics (relativity theories and quantum theories) in the first quarter of the twentieth century, and was reflected in the writings of Moritz Schlick and Bertrand Russell. The logical empiricist Schlick (1918) argued that we cannot intuit the unobservable entities of mathematical physics since they were not logical constructions of sense data, but we can grasp their structural features by implicit definitions, and this was all scientific knowledge required. In accord with this trend, Russell (1927) introduced objects with structural and implicit definitions and claimed that in science we can only know structures, which can be expressed by terms used in mathematical logic or set theory, but not properties and essences of objects.

It should be noted that the mathematical structure taken so seriously by structuralists, as a structure of relational statements, is both causally inert, lacking structuring agent for structuring laws between causally effective properties, and neutral to the nature of relata, and thus cannot exhaust the content of the relata. For example, classical mechanics and quantum mechanics share many mathematical

[2] This implication is explored in details in Cao (2010).

structures, and thus these structures can tell us nothing about the classical or quantum nature of the physical entities under investigations. A permutation group, after certain interpretation, can help to tell boson from fermion, but it cannot tell scalar from vector. Surely with the introduction of more and more refined mathematical structures together with relevant interpretations, the content and nature of the relevant relata can be gradually revealed and increasingly approached. But then this is achieved by introducing additional ontological posits, and thus goes beyond the confine of structuralism.

Poincare's followers in our times, the so-called epistemic structural realists, have made three central claims (Worrall 1989, 2007). First, it is claimed that objects may or may not exist, in any case, they are totally inaccessible, they are hide forever from our eyes. Thus as far as objects are concerned, we have nothing meaningful to say. Second, it is claimed that scientists are able to discover true relations and structures in the real world, and these relations and structures are accumulating, retained across radical theory changes. It is argued that the structural knowledge, in particular those expressed in mathematical structures, have globally reflected the true structure of the world. This stance justifies the term structural realism. Third, it is also claimed that structural realism is the only defensible realist position in understanding science; no stronger position would be defensible or have any real meaning; certainly not the position in which an item by item realistic reading of scientific theories, or the referential continuity in traditional referential semantics, is pursued.

There are two serious defects in this position. First, if entities, in particular those central to a scientific theory (such as the ether for Fresnel's optics or photon for Einstein's theory of photo-electric effects, the FEs) are merely the names of images we have substituted for what really exist, then scientific revolutions would be relegated into a status of illusion. But scientific revolutions have existed and played great role in the evolution of science. We can give different accounts of what scientific revolutions actually are, but it is hard to swallow that such an important phenomenon in the history of science is merely an illusion. The reason why FE is so important for the understanding of scientific theory lies in the fact that the ontological commitment of a theory to it specifies what is going to be investigated in the theory, dictates its theoretical structure and the direction of its further development, and thus constitutes what Lakatos called the hard core of a research program (another name and form of CF).[3] Can we use structure to give an acceptable account of scientific revolution? No. Surely, structure in the discourse of structural realism can be mobilized to offer an account of the continuity in revolutionary theory-changes. But then the discontinuity in the changes would be invisible. The inability to see the discontinuity is inherent to structuralism because the mathematical relations without physical interpretation—which, as argued by Redhead (2001), as an additional ontological posit is a taboo to structuralism—are neutral to

[3] For a detailed discussion of the important roles of FE in theoretical physics, see Sect. 9.1 of (Cao 2010).

the nature of relata, and thus cannot exhaust the content of the relata, as we just indicated above. For this reason, concentrating on the shared mathematical structures, though helpful for conceiving the history of physics as a cumulative and progressive process, would render, for example, the quantum revolution invisible.

Second, the item by item realist reading is impossible, it is argued, because there is no way to have direct access to an isolated entity in the world. That is true. But it is also true that no one can have any direct access to the structure in the world either. Ultimately speaking, the only thing that is directly accessible to us would be sense data or introspection, not even concepts or percepts, let alone the structures of the world. If, cognitively, structures are accessible to us through our reasoning faculties, then the ingredients of the structure should also be accessible to us through the similar reasoning faculties. The arguments for the existence, comparison, and continuity of structures can be applied to entities in the same manner.

Take the electron as an example. Is it legitimate to discuss the referential continuity of the electron? Surely we can. The very existence of the electron is indicated by pointing to those with lightest mass and smallest negative charge, which qualities can be taken as its quiddity. Whenever some particle has shown to possess this quiddity, we know that it is a member of the natural kind "electron". Now surely the notion of electron makes sense only within a particular theoretical framework, J.J. Thompson's, Rutherford's, Borh's, or Heisenberg's framework. We don't have direct access to an electron. Any conception of electron must be specified by some theory. But no matter how radical a change happened between the electron in Thompson's theory and that in Heisenberg's, in case both possess the same quiddity, we know that there is a referential continuity across theory-changes.[4] Then a deeper question is: is it possible to claim the existence of an entity when the description of this hypothetical entity undergoes radical change (Radical in the sense that the essence of the hypothetical entity is also changed)? The answer depends on which position you take. If entity is conceptualized in structural terms, as will be discussed below, the answer is yes. In any other position in which entity is conceptualized in non-structural terms (such as haecceity or substance), then the answer must be no, as the Kuhnians have argued over decades.

Joseph Sneed (1971) and Wolfgang Stegmuller (1979) adopt a different form of structuralist approach. In their informal set-theoretical approach, structure refers to the structure of the whole theory, including mathematical formalism, models, intended applications and pragmatics, and thus has enough room for empirical content. But what this meaning-holism-based structuralism lends support to is not scientific realism and related view of continuous scientific development, but the Kuhnian anti-realism and the related disruptive view of history of science. For other

[4] It goes without saying that the whole discussion above remains within the discourse of structuralism because the notion of the electron as a kind of entity is completely formulated in structural terms: the notions of mass and charge are (and can only be) defined in relational and structural terms: no entity would have any mass if it exists lonely in the world in which no other masses exist and thus no gravitational relations with other masses exist; the same can be said to the notion of charge.

structuralists, the notion of structure is generally narrowed down to mathematical structures, and the empirical content, as suggested by ontic structural realism (OSR), can only be smuggled into structural knowledge through data models, the structure of which is targeted by the mathematical structure of the theory.

OSR claims that there is nothing but structure. What does it mean? When it means—as a response to the difficulty in conceiving a relation without releta while maintaining a purely structuralist metaphysics—that the notion of entity has to be reconceptualized in structural terms, it is essentially correct. Any haecceitist notion of entity completely detached from network of relations and structures has to be rejected. It has to be rejected because no such entity could have existed, or at least we humans have no access whatsoever to such kind of mysterious entities. There are good reasons to believe that all entities have their internal structures and are themselves embedded in various networks of structures. It is this involvement in structural networks that have provided the possibility for humans to have cognitive access to them.

But when the claim means, as French and Ladyman (2003a, b) frequently emphasized, that entities are merely the nodes in the structure, it is ambiguous and does not differentiate two cases in the structuralist discourse: there are things between relations, and there are relations between things. The first case refers to the holistic structure, in which relata are merely the place-holders, their existence and meaning are constituted by their place and role played in the structure. The second case refers to the componential structure, in which the existence of structure depends on the existence of its constituents and the way they are structured together.

It should be noted that the second case remains within the structuralist discourse: the constituents of a structure themselves are embedded in various structural relations and have their own structure. Certainly they can never exist by themselves alone. Still, the second case justifies the notion of atomicity at each and every level of componential structure. A hydrogen atom is a structure of an electron and a proton glued together by electromagnetic forces; but both electron and proton are not constituted by hydrogen as a structure. This kind of atomicity, however, does not contradict the general claim that for the co-existent ontological categories, (structure and constituents, or relation and relata), no ontological primacy of one over the other can be justified, although in terms of epistemic access, structural relations definitely enjoys primacy over relata (ingredients).

So there is a tension in the notion of reconceptualizing entity in structural terms. It can mean, as the OSR advocators intend to mean, the dissolution of (physical) entity into (mathematical) structure, or it can mean the constitution of an entity by the structural relations it involves in (Cao 2003a, b). The dissolution view would suffer from the same difficulty the no-entity position (thus no-FE) suffers in its impotence in understanding scientific revolutions. Note that when an entity (as a member of a natural kind) is said to be constituted by the structural relations it is involved in, it means conceptually we form our conception of an entity by knowing the structural relations it is involved in. But more importantly, it also means that an entity is metaphysically, ontologically, constituted by the structural relations it is involved in. Without these relations, no entity would exist in the first place.

In the Ramsey-sentence version of structural realism, the structure means the structure of observable content. But in order to structure the observable content properly, scientists need FE central to the theory that is Ramsified. Scientist working in any fundamental theory simply cannot formulate the structure of his theory's empirical content such as the Casimir effect or the three-jets phenomenon without adopting certain explanatory fundamental entities, such as the vacuum field or quarks and gluons. Then the crucial question in the structural realism debate is: should one take FE, a theoretical term in the Ramsey sentence formalism for sure, as a way of organizing the observable content, or as having existence in the real world? My own position on this issue is: FE as a natural kind term must refer to kind of things that are to be found in nature, and the kind itself is constituted and individuated by some underlying factors existing in the world. How to justify this position and how to settle the further question regarding the validity or invalidity of the incommensurability thesis in terms of FE are the topics of the next section.

SHASR—Beyond Historicism and Structuralism

Structuralism in philosophy of science, due to its commitment to structure-only metaphysics rejecting the existence or relevance (to science) of entities, restricts scientific knowledge to the forms of the world rather than its content, and thus is unable to offer any convincing account of scientific revolution, which is nothing but a radical (incommesurable) shift from one CF to another, each is committed to and constituted by a special FE. Its reconceptualization (of entity in structural terms) thesis, however, has offered an important insight. Although its original intention is to dissolve physical entities into mathematical structures, the intimate connection highlighted thereby between entity and its structural features can be interpreted in a constitutive way and support the claim that entities are constituted solely by its structural properties and relations, (which are scientists'only epistemic access to objective reality), and thus-constituted entities exist objectively. With this interpretation, structuralism is able to characterize science as an open and cumulative process in which objective knowledge of the world, both its form and content, can be obtained.

However, the absence of the notion of CF within the structuralist discourse has rendered it unable to capture the constitutive and constructive nature of scientific knowledge: the very fact that all structural knowledge is obtained within a CF is simply ignored. The ignorance of CF's constitutive role is one of deep reasons why structuralism cannot properly understand scientific revolution.

In contrast, historicism in contemporary philosophy of science, Kuhnian or otherwise, characterizes scientific knowledge as necessarily constituted and constructed within a CF, and thus is able to accommodate important features of science, such as science having objective content within a CF (internal realism), scientific revolutions as CF-shifts, and many others.

Difficulties with historicism have their deepest roots in the very notion of CF itself. Constitutive framework (CF), in its original Kantian form, or its various neo-Kantain off-springs, is characterized by its closeness: it specifies the space of possibilities, and thus all knowledge produced within the CF cannot go beyond it and, in particular, could not be in contradiction to it. Knowledge and truths are definable only within the CF, either a priori or by convention. For Kant, the notion of CF is self-consistent, it is rooted in human nature, human mental faculties. Once the Kantian apriorism is rejected, however, difficult questions for all forms of neo-Kantianisn arise: where did a CF come from? If it is not given a priori, then by what is it constituted? By arbitrary conventions or something else? How can it be possible to have different CFs? Why some CFs are mutually incommensurable?

So the conceptual situation is this: structuralism offers a relatively solid foundation for obtaining objective knowledge of the world, on the basis of objective structural knowledge about the form of the world, but is unable to accommodate the historical fact of scientific revolutions due to its blind sport on CF; historicism captures all implications of CF, yet is unable to understand its very nature, especially its origin and the causes for its change. Thus it seems impossible to have a coherent picture about science without going beyond structuralism and historicism while assimilating the insights from both of them.

It is crucial to realize that a CF, a priori or conventional or otherwise, is itself constituted by a set of presumptions. Some of them are metaphysical in nature, namely based on well-entrenched commonsense beliefs; others obtain their authority from successful scientific theories. This constitution of CF certainly was the case for Kant's apriorism with or without his realization, and was clearly revealed by the collapse of Kantian apriorism in the mid-19th century, when non-Euclidean geometries and non-Newtonian (non-object based) field theory had arisen. The relativization of apriorism in neo-Kantianism and the historicization of it in current neo-Kantian post-Kuhnian philosophy of science advocated and defended by Michael Friedman and others, are all driven by the changes in the presumptions constituting CF, deep changes in fundamental physics, such as the rise of relativity and quantum theory, in particular.

How can the changes of presumptions constituting a CF be possible if all knowledge, scientific or commonsensical, is constituted within the CF? Translated into the notions we used in this paper, how an EF constituted within a CF can change to a different FE which would constitute and characterize a new CF? Of course, this kind of changes actually occur, as vindicated by the collapse of Kantianism and the historical fact of scientific revolutions. The question is how to conceptualize it. Thus the necessity of going beyond historicism as well as structuralism.

Taking into consideration the role of science in constituting CF, which is vindicated by the rise and fall of Kantianism, the developments of fundamental physics in the 20th and 21st centuries are crucial in motivating our new understanding of the nature of CF and of the rationale and mechanism for the change of CF. Most important among these developments are the rise of general theory of relativity concerning the nature of spacetime, the rise of quantum theory concerning the probabilistic nature of what exists and what happens, the rise of quantum chromodynamics concerning the reality of permanently confined entities, and, in particular, the rise of quantum gravity concerning the deepest layer of reality which presumes no spacetime structures, yet underlies the microscopic and macroscopic realms with spacetime structures.[5] These developments have clearly shown that objective knowledge is structurally constructed within a historically constituted CF, and that the increasing of structural knowledge thus constructed will sooner or later change the configuration of the structural knowledge accumulated so far that are responsible for the constitution of a CF, resulting in the emergence of a new CF.

Motivated by these significant developments, I propose a structuralist and historically constitutive and constructive approach to scientific realism (SHASR), accordinng to which objectivity and progress of scientific knowledge can be conceptualized in structuralist and historicist (historically constitutive and constructive) terms.

Central to the approach is the understanding of constructing a FE by using structural knowledge within a CF and reconstructing, still within the same CF, a new FE in terms of increased structural knowledge, which new FE goes beyond and may be in conflict with the given CF, and characterizes a new CF. How could this be possible? Surely it would be impossible if CF is closed as historicism assumes it to be. Yet the historical constitution of CF mentioned above points to a dialectics between science and CF (a mutual constitution between them), which makes the closeness thesis untenable. More pertinently, we should take CF as a mediation in science's exploration of the world, which is a necessary condition for science to be possible. Yet, as a mediation it is a window rather than a curtain for science to see what exist and happen in reality, and thus is itself conditioned by the exploration and has to adapt itself to the situation created by the exploration if it turns out that the window is not proper for seeing what have already emerged from the exploration, looming increasingly clearer and larger, even though the inappropriateness can only be sensed through the improper window (the current CF). The situation just mentioned is familiar to historians of science, namely the occurrence of anomaly, which is the midwife of new scientific theory and new CF.

Let us look at construction first. Crucial to the mutual constitution of entity (FE as a special case) and structure is a proper understanding of the very nature of entity. As a causal agent an entity is endowed with a certain group of basic properties which dictate its nomological behaviors and thus render it embedded into various causal-hierarchical structures. The identities of different kinds of entities and the individuality of each member of a kind of entity are, for this reason, constituted by relevant groups of structural properties. Thus a kind of fundamental entity is constructed when a combination of basic factors (the constituting structural

[5] Detailed descriptions and analyses of these developments can be found in Cao (1997, 1999, 2001, 2006, 2010, 2014a, 2016).

features) realized together as a being, while a concrete entity may be considered a nexus consisting of an inner core of tightly co-dependent structural features constituting the entity's essence and a corona of swappable adherent structural features allowing it to vary its features while remaining in existence.[6]

In more detailed terms, when we have a set of empirically adequate and qualitatively distinct structural statements (all of them involve an unobservable entity and describe some of its features scientifically discovered), the constraints on the organization of the given set of structural statements (structural knowledge) necessary for the objective constitution of an entity (or FE) can be formulated as follows.

If within the given set there is a subset such that (i) it is stable within a configuration of the set and is reproducible in variations of the configuration; (ii) it occupies a central place (the core) in the configuration; (iii) it describes some physically specific features that can be interpreted as the intrinsic features[7] of the entity, which are different from its accidental (context-dependent) features described by those situated on the periphery of the configuration; among these features, (iiia) some of them (e.g., spin) are common to various physical entities, (iiib) others (e.g., fractional charges) are qualitatively specific to the entity (e.g., quark), thus can be taken as its essential features and used as identifying references to characterize the entity and distinguish it from other entities; and (iv) some of its statements describe the causal efficacy of the intrinsic features (essential features in particular), these causally effective features can be taken as a basis for explanation and prediction; then we are justified (i) to claim that there is an unobservable entity constituted by the set of structural properties and relations; (ii) to take it as ontologically inseparable from the structural properties described by each and all statements in the set, and is responsible for the general mechanism (underlying empirical laws) resulted from these properties (especially dynamical properties); and (iii) to take the objective structural statements in the set as providing us with objective knowledge of the unobservable entity.

It should be stressed that the objectivity of thus constituted entity has two sources, one from the objectivity of the constituting structural knowledge (statements), the other from the objectivity of the holitstic feature of the constitution. That is, the set of structural statements constituting the physical entity has a new feature that is absent in what is involved in each and all structural statement in the set. Different from an amalgamation of structural statements, which itself is structureless, the constituting set is hierarchically structured. Most importantly, the set has a stable core subset which provides feature-placing facts about the hypothetical entity, and thus can be used as identifying references to the entity, rendering the entity referentially identifiable. As a crystallization of holistic characteristics of a hierarchically structured configuration of the set, which are prescribed, in a specific theory, by a specific allocation of roles [essential or not] and places [core or

[6] Cf. Simons (1994).

[7] In the sense of context-insensitive, not of existing lonely without connecting to others.

periphery] to the statements involved, (in addition to the coherent existence of the structured configuration at specifiable spacetime locations), the entity thus constituted is relatively stable against all changes except for those which have changed the role of core statements.

What is important for scientific realism is that a thus constructed FE as a natural kind term must have its referent in the world. Here the necessary and sufficient conditions for the kind membership are more than the structural statements themselves, but also involves the holistic characteristics or the specific configuration of the set of the structural statements just mentioned above. This is important for the reference-fixing and objectivity of the constructed FE, but is also crucial for understanding the reconstitution of FE and thus the emergence of new FE and scientific revolution, as we will see shortly.

But there is an issue of underdetermination of the thus constituted FE by the constituting set of structural knowledge, which undermines the uniqueness or even the reality of the constituted FE by opening the door for conventionality. It seems difficult to escape from this problem because whatever satisfies the constituting set must be considered as a referent for the FE thus constituted. But the satisfaction does not pose any constraints on the internal causal composition or functional organization of the FE other than those only on its upward accessible relations, certainly not on its downward compositions, for any entity sitting at the interface with other theoretical entities and accessible relations which scientists are interested in investigating. A truism that is often forgotten is that the nature of an entity is always much richer than any specific description of its structural involvements. The reason is simple. Many of its properties and relations may not be realized in any situation or known to the describers or scientists.

Since radical under-determination of FE by the constitutive set without any empirical consequence is scientifically uninteresting and can be fixed by revising metaphysical scheme, while those with conflicting empirical consequences can be resolved by further investigations in more differentiating contexts, the only philosophically interesting cases of under-determination are those with compatible entities. Here I find the idea of generative entrenchment (GE) highly helpful. According to Wimsatt (2003), GE of an entity in a complex system is a measure of how much of the generated structure or activity of a complex system depends upon the presence or activities of that entity. Entities with higher degree of GE are more conservative in evolutionary changes of such system. Thus GE acts as a powerful and constructive development-constraint on the course of evolutionary process. Now science is clearly an evolving and highly complex system, and FE in a theoretical science is the one with highest degree of GE (all phenomena described by the theory depend on its presence and behavior), thus it would be virtually impossible to replace an FE with anything else without changing the whole theoretical description and structure. Uniqueness cannot be ultimately established, but practical uniqueness can be assumed by taking ever increasing number of structural descriptions as identifying features for fixing, or more properly constituting, the identity of the FE. The uniqueness and reality of a theoretical entity can be established, or constructed in a positive sense, this way, to the extent reached by the structural knowledge involving this entity. If the idea of GE can be deployed as a strong constraint in arguing against multiple realizability in philosophy of mind, and for the idea that mind can only be a brain phenomenon, then it would be much easier to argue that QCD as a complicated conceptual scheme can only be realized in quarks and gluons. That is, the reality of quarks and gluons are almost uniquely fixed by the structural description.[8]

The idea of GE can be easily extended from entity to constitutive factors of entity, which in fact is an important conceptual resource for our understanding of the reconstruction of FE which is crucial for SHASR.

Now come to reconstruction. Since the door to any direct access to unobservable FE is closed, any construction of FE has to be reconsructed time and again with the unavoidable changes of the configuration of the set of structural statements from which a natural *kind* (FE) is constituted. With the increase of structural knowledge (statements), the reallocation of some core and peripheral statements, and the change of the role of some core statements (describing essential features or not), the defining features of the configuration change accordingly. As a result, the identifying references, or the content, or the characteristic features, metaphysical or otherwise, of the *kind* also change. That is, what is constituted and thus conceived is a different *kind* from the original one. The completion of such a process of reconfiguration is the substance of a scientific revolution through which theorists have changed their ontological commitment and thus the ontological character of the whole theory.

In addition, the structural construction and reconstruction of FE, although reliable, is fallible and subject to revisions. Thus, the attainment of objective knowledge at the level of underlying entities can only be realized through a historical process of negotiations among empirical investigators, theoretical deliberators and metaphysical interpreters. This character of our approach to FE is crucial to the realist conceptualization of the history of science, as we will see shortly.

Let us come back to the central question: Is it legitimate to claim the referential continuity of FE across radical theory-changes? For example, is it true that we refer to the "same" electron when the description moves from Thompson's theory to Rutherford's, Bohr's, Heisenberg's, and Dirac's theory? Since we don't have direct access to electrons, and the notion electron makes sense only within a particular theoretical context, many holisticists argue, surely the "sameness" of the electron cannot be justified when the electron is described in radically different theories.

This is true. From the perspective of SHASR, however, some kind of referential continuity of an entity can still be argued for by appealing to the notion of

[8] For the case of QCD, see Cao (2010, 2014a); for a more general discussion of the claim, see Cao (2014b).

reconfiguration discussed above.[9] The referential continuity based on, epistemically as well as ontologically, the reconfiguration may appear in three different ways. First, if some identifying-features-placing structural statements, such as the lightest mass and smallest negative charge in an atom in the case of electron, are retained in the new configuration, then no matter how radical changes have happened between theories across conceptual revolutions, such as those between Thompson's theory and Dirac's theory, it is justifiable to say that physicists are basically talking about the same electron.[10]

Second, it may happen that the expansion and reconfiguration of structural knowledge of an entity and other entities in the domain under investigations results in a change of the ontological status (primary or derivative) of the entity. In the case of strong interactions, for example, the pion in Yakawa's theory was a primary agent for the strong interaction; later it was relegated into a status of epiphenomenon in the quark model and QCD. But a change of status does not deny its existence and identity, and thus the referential continuity in this case cannot be denied.

Finally, the referential continuity may also be realized though a mechanism of ontological synthesis, which, different from the two ways mentioned above that can be accepted without too much reflection, is comprehensible only from the perspective of our new approach to entity. If there are two distinctive configurations of structural statements, each of which is responsible for constituting a distinctive entity, and if an empirically adequate combination of one (or more) constitutive structural statement(s) from the core subset of each configuration constitute a new core subset in an enlarged and/or reconfigured set of structural statements, then the new configuration with a new core subset may be responsible for constituting a new entity, which, if approved by nature, would be a case of ontological synthesis. A few examples will be given below.

It is worth noting that the notion of reconfiguration in constituting FE (a constructed natural kind term) has provided the theoretical resource to comprehend the rich structure of scientific developments (both in normal science and during scientific revolutions). In the history of science, fundamental science in its evolution frequently reshuffles and re-organizes the constitutive factors of its FE. In addition to simple discard or retention of FEs across theory changes, the notion of

[9] In the new configuration associated with the new entity (with new and different essence, thus a different entity from the old one) constituted thereby, the retained structural features from the old configuration retain their constitutive roles in the new context, as the extended notion of GE mentioned above suggests, although their places (at the core or periphery) and functions (identifying-features-placing or not) have changed.

[10] In an important sense, the classical electron in Thompson's theory and the quantum electron in Dirac's theory are different entities. In an even deeper sense, however, they, as manifestation of two different aspects at two different levels of the same electron in the causal-hierarchical structure of the noumenal world, refer to the same entity, or more precisely, refer to different aspects at different levels of the same noumenal electron. Thus realism defined in SHASR is not the naïve realism about unobservable entities or properties or mechanism, etc., but the realism of various manifestations of the causal-hierarchical structure of the noumenal world.

reconfiguration offers a mechanism for accommodating both the apparent ontological discontinuity (new FEs having replaced old ones) and a deep continuity in our knowledge of what *exists* in the world, in terms of factors, that constitute the old FE and partially constitute the new FE, existing before and after radical theory changes.

More interesting from the perspective of SHASR, however, is that the notion of reconfiguration also helps us to understand how a new fundamental theory is created. Generally speaking, reconfiguration underlying the emergent of a new FE is essentially a generalized version of ontological synthesis (OS). "Generalized" in the sense that what are synthesized are not necessarily those factors already having constituted an FE, but include some constituting factors which, although have not constituted any FE, are nevertheless genetically entrenched as hinted earlier.

Emergence of a new FE through ontological synthesis should be understood as an epistemic process through which another aspect or (perhaps deeper) level of reality is revealed. For example, as I have discussed elsewhere (Cao 2001, 2006), the revision of the ontological foundations of the general theory of relativity (GTR) and quantum field theory (QFT) can be viewed as an attempt at an ontological synthesis so that the combination of two structural features—one (the universal couplings) is constitutive of the gravitational field, and the other (the violent fluctuations) is constitutive of quantum entity—can be consistently adopted to constitute a new FE, the quantum gravitational field, which is violently fluctuating but is also universally coupled with all physical entities. Here the foundational constraints posed by predecessor theories (GTR and QFT) have to be taken seriously because these constraints have encoded all the knowledge we have acquired through the predecessor theories, and thus have provided us with the only epistemic access to the unobservable reality we intend to describe in the successor theory. Another well-know example is the emergence of quarks and gluons (the FEs of QCD) from synthesizing the structural constraints posed by the predecessor models (the parton model and the constituent quark model), namely the scaling law and notion of color.

Closely related with the epistemic emergence discussed above is the ontological emergence, which, however, has to be understood differently. Let me illustrate this subtle point with an example from quantum gravity. The quantum entity epistemically emerged from the classical entity under certain quantum constraints (that is, as the result of epistemic ontological synthesis) and the classical entity cannot be the same entity only behave differently on different energy scales. If we take this point seriously, we have to give up the attempt at actively quantizing some classical degrees of freedom when it is not appropriate, for example, in the case of gravity. This means that we have to take a quantum realist position, starting from something which is already quantum in nature. Then the difficult question is what this quantum entity is. One clue to the answer it this. In order to recover classic gravity, which is a hard constraint posed by the predecessor theory discussed above, it must share one feature with gravity, namely, it is universally coupled with all kinds of physical degrees of freedom including self-coupling, although it cannot be a metric-kind or connection-kind of entity. Let us call it quantum gravitational field.

Surely the recovery of classical limit, which serves as a consistency check for the construction of quantum gravity, such as the recovery of geometry and material degrees of freedom from the dynamical processes of the same underlying field, proposed, e.g., by geometrogenesis approach (one of popular models in quantum gravity), is a typical example of ontological emergency. The recovery or ontological emergency has to go through chains of phase transitions determined by the interactions between sub-systems within the quantum system described by the quantum theory of gravity. That is, as the result of heterogenerous emergence, they are qualitatively novel, different from what happens in the quantum gravity regime. It should be stressed that the heterogeneous emergence can be realized in different ways, from coarse graining to collective excitations—such as phonons in condensed matter physics or pions in the Nambu model of PCAC, (Cao 1991)—to more complicated process similar to the one in which hadrons emerged from quarks and gluons as suggested by QCD (Cao 2010).

A striking example in this regard is the case of the so-called double special relativity. In addition to the classical limits recovered in traditional way by having the quantum effect removed or letting the Planck constant h approach to zero, resulted in GTR, there is another kind of limit, call it special limit, reached by having the gravity effect removed or letting the gravitational constant G approach to zero. The result should be the conventional QFT system defined on a Minkowskian background spacetime. But the heterogeneousness of the emergence in the limiting process, in the case of quantum gravity, may manifest itself in an unexpected way as follows. If in the process of letting both h and G approach to zero, but keep their ratio G over h, G/h constant, for example equal to the Planck mass squared, then in addition to ordinary special relativity, we would obtain a deformed or double special relativity, or DSR, which has provided, in addition to traditional ones of black hole phenomenology, big bang physics, gravitational waves as suggested by the observation of the binary pulsars, another falsifiable prediction, namely the energy dependence of the speed of light. The observational test of this additional prediction, if confirmed, would give the relevant model for quantum gravity an impressive empirical support; or if falsified, would discredit the model that gives such a prediction (Amelino-Camelia 2010; Cao 2007).

It should be stressed that SHASR, while transcending both structuralism and historicism, is also categorically different from both traditional scientific realism and the historicized or Kuhnianized version of neo-Kantianism (HNK) advocated by Michael Friedman and others. The difference with the former centered around two issues: the conception of unobservable entities (FE included) and the conception of objectivity. Regarding the first issue, the intrinsic nature or the ontological content of FE in traditional realism is frequently conceived in non-structural terms, such as haecceity and substance; while in SHASR it is taken to be constituted exclusively by structural properties and relations the entity possesses, and constituted in a holistic way as we discussed above. When structural terms are evoked by traditional realists in conceiving an entity, the entity is conceived as a a member of a preexisted and fixed natural kind, while for SHASR a natural kind is not preexisted and fixed, but rather, it is historically constructed, revisable, subject to reconstruction time and again.

The ontological basis of objectivity, for traditional realism, is the existence of the objective world which is independent of human activities, and thus the objectivity of knowledge can only be defined in terms of correspondence with this objective reality. For SHASR, however, the notion of objectivity is not detached from human involvement, which is an illusion. Rather, it is conceived in terms of nature's resistance to any arbitrary human construction. This kind of resistance, according to SHASR, is the only ontological basis for objectivity. Take the case of QCD as an example (Cao 2010). The ingredients of hadrons were conceived in various ways. Along one line of conception, they were first conceived through certain set of structural knowledge to be partons, then partons were reconceived as quarks and gluons. With the importation of the structural constraints posed by the notion of color (originated from the constituent quark model) into the "current quark" picture, they were reconceived again to be colored quarks and gluons, which conception was approved by nature and accepted by the high energy physics community. Along another line of thought, the ingredients of hadrons were conceived to be integrally charged ones, which was not approved by nature and thus was discarded by the community. All these were the result of human construction in terms of structural knowledge, but the objectivity of some conceptions and constructions rather than others is warranted by nature's approving and disapproving responses.

The categorical difference with HNK can be best captured by focusing on their different attitudes toward the role of the noumenal world in knowledge production. As a Kuhniainized Kantinism, HNK shares all the defects of Kuhnian historicism, namely historical relativism and anti-realism, mainly because it shares the latter's inability to see the active role of the noumenal world in the regulation of knowledge production. According to HNK, objective knowledge is possible if it is constructed within a constitutive CF, and thus can make sense only within this CF. Yes, the advocates of HNK acknowledge, CF nowadays cannot be conceived as an a priori framework anymore, rather, it has to be historicized. But, they argue, it is still rooted in mental faculties, and cannot be otherwise, thus is completely detached from the noumenal world. Progress and rationality of science is realized Friedman stresses, through the regulation of science, not by a priori reason as the Kantans insisted, but by intersubjective consensus of a Habermasian style. What is strikingly clear here is that the noumenal world is completely irrelevant. The laudable intention is to embrace "Enlightenment rationality and normativity" (Friedman 2012); yet realism, scientific or metaphysical, is gone (surely realism is not what he wants), the cause for the change of intersubjective consensus becomes mysterious, and thus the predicaments of Kuhnian historicism persist.

In contrast, according to SHASR, the noumenal world is the ultimate arbiter for the truth of scientific knowledge. It functions in this regard by ways of responses and resistance discussed above, which has provided the ontological foundation of objectivity for science. More importantly, a CF, whose interactions with the noumenal world results in the appearance of a phenomenal world in which various scientific knowledge is constructed, has to adapt itself to the historical situation created by resistances when they occur. That is, CF is porously but not completely closed framework: the noumenal world has zigzag ways of getting its influence into it. This dialectics between CF and the noumenal world is the real cause for the historical change of CF. It regulates the development of science, and renders the progress and rationality of science intelligible.

In sum, SHASR is a realism about the noumenal world, not a realism about isolated unobservable entities, properties and mechanisms. According to SHASR, science produces and expands our objective knowledge about the noumenal world, whose various aspects and levels of rich and hierarchical kind-structure, such as those manifested in classical and quantum electrons discussed above, are historically, step by step, captured by science through the interactions between CF with experience on the one hand and between CF and the noumenal world on the other. The interactions of science with its cultural context, metaphysical scheme included, result in a world picture or worldview, which underlies and guides human actions and is subject to change with the changing situation created by human actions, scientific explorations included.

Concluding Remarks

SHASR, by adopting the notion of reconstitution of FE (thus CF) under the conditioning and regulation of the noumenal world (by ways of responding), is able (i) to escape from the structuralist no-FE trap and its unfortunate consequence of inability in offering an account of scientific revolutions, and (ii) to meet the challenges posed by Kuhnianism, rejecting successfully its historical relativism and anti-realism. Constitution and reconstitution are metaphysical categories, although they also have epistemological implications and will appear, epistemically as construction and reconstruction of FEs and CFs. In the metaphysical sense, reconstitution can be viewed as the philosophical foundation for emergency, which characterizes the qualitative transition of the old to the new, as the result of the internal dynamics of the old.

Emergency is ubiquitous. In the realm of human cognition, we see the emergency of science from commonsense; within science, we see new CF emerges from old one, namely the radical conceptual change or scientific revolution. According to Kuhn, the world also change: a new world emerges from the old one, and we are living in different worlds before and after a scientific revolution. It is true. But the world here refers to the phenomenal world, which is intimately related or even co-extensive with CF, not the noumenal world, which for Kuhn is ineffable, indescribable and undiscussable.

Does the noumenal world change? What is the relevance of emergency to the noumenal world? For all Kantians and Neo-Kantians, Kuhn and Friedman included, these are unintelligible and illegitimate questions. From the perspective of SHASR, however, the relevance of emergency to the noumenal world is indisputable and can be summarized as follows.

First, the existence of the noumenal world is manifested in the phenomenal world, which, as the result of human interactions with the former, is inseparable from the former, and thus makes it accessible to humans.

Second, the noumenal world is infinitely rich. Its richness is manifested in incessantly emerging new phenomenal world, and its rich kind-structure is gradually revealed by incessantly emerging new CFs in the history of science.

Third, from this perspective, scientific realism is a version of metaphysical realism about the noumenal world, and the historical development of science is only the incessant pursuit of this on-going realist project.

References

G. Amelino-Camelia, Doubly-special relativity: facts, Myths and some key open issues. Symmetry **2**, 230–271 (2010)

T.Y. Cao, Spontaneous breakdown of symmetry: its rediscovery and integration into quantum field theory, in *Historical Studies in the Physical and Biological Sciences*, ed. by L.M. Brown, vol. 21(2), pp. 211–235, 1991

T.Y. Cao, *Conceptual Developments of 20th Century Field Theories* (Cambridge University Press, Cambridge, 1997)

T.Y. Cao, *Conceptual Foundations of Quantum Field Theory* (Cambridge University Press, Cambridge, 1999)

T.Y. Cao, Prerequisites for a Consistent Framework of Quantum Gravity, in the Studies. Hist. Philos. Mod. Phys. **32**(2), 181–204 (2001)

T.Y. Cao, Can we dissolve physical entities into mathematical structures? Synthese **136**(1), 57–71 (2003a)

T.Y. Cao, What is ontological synthesis? A reply to simon saunders. Synthese **136**(1), 107–126 (2003b)

T.Y. Cao, Structural realism and quantum gravity, in *Structural Foundation of Quantum Gravity*, ed. by D. Rickles, S. French, J. Saatsi (Oxford University Press, 2006)

T.Y. Cao, Conceptual Issues in Quantum Gravity, an invited 50 minute talk delivered (on Aug 11, 2007), in *The 13th International Congress of Logic, Methodology and Philosophy of Science,* 9–15 Aug 2007, Beijing, China; unpubliahed, 2007

T.Y. Cao, *From Current Algebra to Quantum Chromodynamics—A Case for Structural Realism* (Cambridge University Press, 2010)

T.Y. Cao, Key steps toward the creation of QCD—Notes on the logic and history of the genesis of QCD, in *What We Would Like LHC to Give Us*, ed. by A. Zichichi (World Scientific, 2014a), pp. 139–153

T.Y. Cao, Incomplete, but real—A constructivist account of reference. Epistemol. Philos. Sci. **41** (3), 72–81 (2014)

T.Y. Cao, The hole argument and the nature of spacetime—a critical review from a constructivist perspective, in *Einstein, Tagore and the nature of Reality*, ed. by P. Ghose (Routledge, 2016), pp. 37–44

R. Carnap, *Logische Syntax der Sprache*. (English translation 1937, *The Logical* Syntax of Language. Kegan Paul), 1934

R. Carnap, Empirecism, semantics and ontology, in *Revue Internationale de Philosophie*, vol. 4, pp. 20–40, 1950. Also in Carnap (1956), 205–221

S. French, J. Ladyman, Remodelling structural realism: quantum physics and the metaphysics of structure. Synthese **136**, 31–56 (2003a)

S. French, J. Ladyman, Between platonism and phenomenalism: Reply to Cao. Synthese **136**, 73–78 (2003b)

M. Friedman, *Dynamics of Reason: The 1999 Kant Lectures at Stanford University* (CSLI Publications, 2001)

M. Friedman, Einstein, Kant, and the relativized a priori, in *Constituting Objectivity: Transcendental Perspectives on Modern Physics*, ed. by M. Bitbol, P. Kerszberg, J. Petitot (Springer, 2009)

M. Friedman, Kuhnian approach to the history and philosophy of science. Monist **93**, 497–517 (2010)

M. Friedman, Extending the dynamics of reason. Erkenntnis **75**, 431–444 (2011)

M. Friedman, Reconsidering the dynamics of reason. Stud. Hist. Philos. Sci. **43**, 47–53 (2012)

M. Friedman, Neo-Kantianism, scientific realism, and modern physics, in *Scientific Metaphysics*, ed. by D. Ross, J. Ladyman, H. Kincaid, (Oxford University Press, 2013)

M. Friedman, *A Post-Kuhnian Philosophy of Science. Spinoza Lectures*. (Van Gorcum, 2014)

T.S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, 1962)

T.S. Kuhn, *The Structure of Scientific Revolutions* (Second enlarged edition University of Chicago Press, 1970)

T.S. Kuhn, *The Road Since Structure* (University of Chicago Press, 2000)

M. Massimi (ed.), *Kant and Philosophy of Science Today* (Cambridge University Press, 2008)

H. Poincaré, *La science et l'hypothese* (Flammarion, Paris, 1902)

M. Redhead, The intelligibility of the universe, in *Philosophy at the New Millennium*, ed by A. O'Hear (Cambridge University Press, 2001), pp. 73–90

B. Russell, *The Analysis of Matter* (Kegan Paul, London, 1927)

M. Schlick, *General Theory of Knowledge* (trans. By A.E. Blumberg, H. Feigl, Springer, New York, 1918)

P. Simons, Particulars in Particular Clothing: Three Trope Theories of Substance. Res. **54**, 553–574 (1994)

J.D. Sneed, *The Logical Structure of Mathematical Physics* (Reidel, 1971)

W. Stegmuller, *The Structuralist Vierw of Theories* (Springer, 1979)

W.C. Wimsatt, Evolution, Entrenchnebt, and Innateness, in *Reductionism and the Development of Knowledge*, ed by T. Brown, L. Smith, Mahwah, 2003

J. Worrall, Structural realism: the best of both worlds? Dialectica **43**, 99–124 (1989)

J. Worrall, *Reason om Revolution: a study of Theory-Change in Science* (Oxford University Press, 2007)

Chapter 7
Quantum Mechanics, the Manifestation of the Territory, and the Evolution of Maps

Ulrich Mohrhoff

An Ancient Conundrum

Here is a problem that Scholastic philosophers have discussed for centuries. Imagine that in front of you there are two exactly similar objects. All their properties are the same, except that they are in different places. Because they are in different places, they are different things. But is the fact that they are in different places the *sole* reason they are different things? Or is there another reason? If you believe that there is another reason, you will look for it in vain, for if two things are different, it is their properties that are different, and right now we are assuming that the two objects have exactly the same properties, except that they are in different places. On the other hand, if you believe that the two objects in front of you are different objects for the *sole* reason that they are in different places, then what you really believe is that the objects in front of you are *one and the same* thing in two places, which sounds preposterous. The resolution of this dilemma had to wait for the advent of quantum mechanics.

Consider the following experiment. Initially two identical particles—that is, particles lacking properties by which they can be distinguished—are observed to be moving Northward and Southward, respectively. The next thing that is known about them—and the next thing that *can* be known about them under the experimental conditions envisaged—is that they are moving Eastward and Westward, respectively. The obvious question then is, "Which incoming particle is identical with which outgoing particle?" It turns out that neither of the two possible answers, illustrated in Fig. 7.1, is consistent with what quantum mechanics predicts. In other words, there is no correct answer to the question "Which is which?" What gives?

U. Mohrhoff (✉)
Sri Aurobindo International Centre of Education, Pondicherry, India
e-mail: ujm@auromail.net

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_7

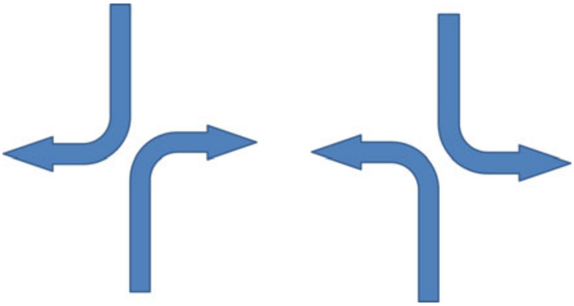


Fig. 7.1 Possible identities in a scattering experiment with incoming particles moving Northward and Southward and outgoing particles moving Eastward and Westward

Unanswerable questions tend to arise from false assumptions. In this particular case, the question implicitly assumes (falsely) that we are dealing with *two* things, while in reality we are dealing with *one and the same* thing observed twice. If the particles are one and the same thing, initially seen moving *both* Northward *and* Southward, and subsequently seen moving *both* Eastward *and* Westward, the question "Which is which?" can no longer be asked. This is how quantum mechanics resolves the dilemma of the Scholastic philosophers. The two objects they contemplated *are* the same thing in different places. Reality *is* preposterous.

What is more, there is no compelling reason to believe that the identity of the observed particles ceases when it ceases to have observable consequences owing to the presence of "identity tags"—properties by which they can be distinguished and re-identified. Nothing therefore stands in the way of the view that all particles—at any rate, all *fundamental* particles[1]—are identical in the strong sense of *numeric* identity.[2] What presents itself here with these properties and what presents itself there with those properties is one and the same "thing."

Fundamental particles are routinely described as pointlike. What is meant by this, however, is that they lack internal structure. By itself, lack of internal structure may be consistent with either a pointlike form or no form at all. There are, however, compelling reasons, both experimental and theoretical, why fundamental particles cannot be literally pointlike, but should instead be regarded as *formless*.

What, then, are the properties that a fundamental particle has, by itself, out of relation to anything else? It has none! To see this, consider a universe containing a single object. Would we be able to attribute to this object a position—to say where it is? Of course not, for we can only say where an object is *relative* to another object. Can we attribute to it a velocity or a momentum? Same answer, for we can only *compare* the velocities or momenta of different objects. Can we attribute to it a

[1] The particles presently considered fundamental are the leptons (which include the electron and the neutrinos) and the quarks (which "make up," among other things, the proton and the neutron).

[2] Numerically identical things are the same thing under different aspects. Thus the evening star and the morning stare are numerically identical—they are aspects of the planet Venus. In the same way "Barack Obama" and "the 44th President of the United States" refer to the same person. Similarly, "the particle moving Southward" and "the particle moving Northward" refer to the same object; they are the same object under two aspects.

mass? Negative again, for only the *ratios* of the masses of different objects are independent of our arbitrary measurement units and thus capable of representing objective properties. Nor can we attribute to it a charge, for charges characterize *interactions*. And so on. Hence, all that can be said about an existing fundamental particle by itself is that it exists.

Putting two and two together: what presents itself here with these properties and what presents itself there with those properties is (i) one and the same "thing" and (ii) something that, considered by itself, lacks properties. It is not *a* being but *Being* —that to which all existing properties owe their existence.

If every fundamental particle in existence is identically the same Being *and* formless, then the shapes of things resolve themselves into reflexive spatial relations, and physical space becomes the totality of spatial relations existing between Being and (the very same) Being. But if physical space only contains (in the proper, set-theoretic sense of containment) spatial relations and the forms they constitute, then Being itself is not contained in space. Rather, Being may be said to contain space, inasmuch as the relations that space contains are reflexive and, in this sense, internal to Being.[3]

Manifestation (According to Quantum Mechanics)

The key that unlocks the mysteries of the quantum world is the concept of manifestation. By entering into (or establishing) reflexive spatial relations, Being manifests both matter and space, for space is the totality of existing spatial relations, and matter is the corresponding (apparent) multitude of relata,[4] which physicists refer to as "fundamental particles."

We keep looking for the origin of the universe at the beginning of time, but this is an error of perspective. The origin of the universe is Being, which exists in an anterior relation to time, and the origination of the universe—its manifestation—is an atemporal transition from undifferentiated Being to the familiar spatially and temporally differentiated world. This transition takes place in stages. The first stage results in the (apparent) multitude of formless particles. The subsequent stages mark the emergence of form, albeit first as abstract forms that cannot yet be visualized. The forms of nucleons, nuclei, and atoms can only be mathematically described, as distributions over probability spaces of increasingly higher dimensions. Only at the penultimate stage do visualizable forms emerge, as atomic configurations of molecules.

[3] For detailed discussions of the interpretation of quantum mechanics invoked in this chapter see Mohrhoff (2013, 2014a, b, 2016, 2017).

[4] Because the relations are reflexive, the multiplicity of the relata is apparent rather than real. Does this mean that the material world is unreal, as some illusionistic philosophies assert? By no means, for the material world owes its existence to a multitude of reflexive *relations*, and these are real.

The question of how the general theoretical framework of contemporary physics can be a calculus of correlations between the possible outcomes of measurements—the notorious quantum measurement problem—becomes intelligible in this light. If quantum mechanics concerns the transition from the unity of Being to the multiplicity of the manifested world, then the question arises as to how the intermediate stages are to be described, and the answer is that whatever is not *as* differentiated as the manifested world, can only be described by assigning probabilities to the possible outcomes of measurements. Particles, atoms, and molecules, which mark the stages of this progressive realization of distinguishable objects and distinct regions of space, can only be described in terms of probability distributions over distinguishable objects or distinct regions of space.

Manifestation (According to Vedanta)

As we have seen, quantum mechanics allows us to infer the reality of an intrinsically undifferentiated Being, a Being that manifests the world by entering into spatial relations with itself. What quantum mechanics cannot tell us is how Being enters into spatial relations with itself, and why. For this we shall turn to what is arguably the most illuminating theory of manifestation available, which is part of the quintessential Indian philosophy known as *Vedanta* (Phillips 1995). What follows is based on the original formulations of Vedanta found in some of the Upanishads (Sri Aurobindo 2001, 2003) and on its contemporary development by Sri Aurobindo (Heehs 2008).

In the terminology of Vedanta, what is ultimately and solely real is called *Brahman*. Brahman relates to the world in essentially three ways: it is the substance (*sat*) that constitutes it, it is the consciousness (*chit*) that contains it, and it is an infinite Quality/Delight (*ānanda*) that expresses/experiences itself in the world. (Because *ānanda* transcends the dichotomy of subject and object, it is at once an infinite Quality and an infinite Delight or Bliss.) Brahman is that by which the world exists, it is the self or subject for which the world exists, and it is the reason why the world exists. It is *sachchidānanda* (*sat-chit-ānanda*).

Brahman can and does adopt a variety of poises of relation between subject and object, between self and world. In its primary poise, this relation is one of identity. Brahman considered as self is (i) coextensive with the content of Brahman considered as consciousness and (ii) identical with Brahman considered as substance. In a secondary poise, the one original self adopts a multitude of standpoints. Concentrating itself simultaneously in a multitude of individual forms, it identifies itself with each. Identified with an individual form, it views the content of its consciousness in perspective, from a particular location. It is in this poise that the dimensions of experiential space (viewer-centered depth and lateral extent) come into being. It is also here that the dichotomy between subject and object becomes a reality, for a self that is identified with an individual form cannot be one with the substance that constitutes all forms.

By a further departure from the original poise of relation between self and world, this multiple concentration of consciousness becomes exclusive. We all know the phenomenon of exclusive concentration, when consciousness is focused on a single object or task, while other occurrences are registered subconsciously, if at all. A similar phenomenon transforms individuals who are conscious of their essential identity into individuals who have lost sight of this identity and, as a consequence, have lost access to the *supramental* view of things. Their consciousness is *mental*, which not only means that it belongs to what appears to be a separate individual, but also that it perceives the world as a multitude of separate objects.

When carried further still, the multiple exclusive concentration of Brahman qua *chit* gives rise to a world whose inhabitants lack the ability to generate ideas—a world in which consciousness is reduced to its power of executing or realizing ideas, of giving them a material form. This power is the essence of what we call "life." Finally, when the multiple exclusive concentration of consciousness is carried to its furthest extreme, it gives rise to a world in which life itself is "involved" (rendered latent or dormant) in inanimate matter. And since the power of executing ideas is responsible for the existence of material forms, the result is a world of formless bearers of purely relational properties called "fundamental particles."

It is worth noting here that, beginning with Leibniz in the 17th Century, philosophers have argued that all physical properties are relational or extrinsic. This offers a way to circumvent the widely discussed "hard problem" of consciousness (Chalmers 1995), which is the problem of explaining how (quantitative) physical processes (in a brain) can give rise to experience, or to the sensory qualities that make up the content or perceptual consciousness. Arguably this too is a problem that arises from a false assumption—in this case the assumption that physical processes give rise to experience. If they don't, there remains the possibility of locating the evolutionary origin of consciousness among the intrinsic or non-relational properties of the relata which bear the relational properties. This possibility has been considered by Bertrand Russell (1927) and more recently by David Chalmers.

The problem with this approach is that it is hard to imagine how the consciousnesses of a myriad of particles can constitute something like the unified consciousness that we enjoy. But if not only all physical properties are relational but also all relational properties are ultimately reflexive—if, in other words, the particles are identical in the strong sense of numeric identity—then the concept of consciousness as an intrinsic aspect of the relata comes into its own. Consciousness is an intrinsic aspect of the relata because the relata are identically the same Being, and because Being, Vedantically conceived, relates to the world not only as an all-constituting substance but also as an all-containing consciousness.

Why the Laws of Physics Are just so

While the evolution of consciousness, and arguably the evolution of life as well,[5] is not a subject for physics, a proper theory of life and consciousness, such as the original Vedanta of the Upanishads, may well be able to tell us why the laws of physics have the form that they do.

In the article that featured his well-known cat paradox, Erwin Schrödinger (1935) noted that "Measurements on separated systems cannot directly influence each other—that would be magic." Three decades later, John Bell (1964) derived his famous inequality, whose violation by quantum mechanics proved that the magic was real. Measurements on separated systems *can* directly influence each other. The magic consists in the fact that such influences cannot be explained by any process continuous in space and time—neither in terms of something propagating from one measurement apparatus to the other nor by something propagating from a single event anterior to both measurements and affecting their outcomes.

But this only highlights the fact that quantum mechanics *never* tells us what (if anything) happens *between* measurements (except, possibly, other measurements), whether they are made on the same system at different times or on different systems at the same time. The theory only explains—via its conservation laws—why certain things will *not* happen. And this is exactly what one would expect if the force at work in the world were an infinite force operating under self-imposed constraints, such as the power by which Brahman manifests itself to itself, in various poises of relation between itself as object and itself as subject. In that case one would have no reason to be surprised (or dismayed) by the impossibility of explaining the correlations that quantum mechanics predicts. After all, it would be self-contradictory to explain the working of an infinite force by a physical mechanism or natural process. What would need explaining is why—to what end—this force works under the particular self-imposed constraints that it does.

To find out why it works under the particular constraints known to us as the laws of physics, we need to do three things. First we need to characterize "ordinary objects" as objects that

[5] A hundred years ago, it seemed obvious to many that life could not have emerged from utterly lifeless matter, just as today it seems obvious to many that consciousness could not have emerged from utterly unconscious matter. Yet today no one appears to seriously doubt that life did emerge from utterly lifeless matter; the seemingly insuperable "hard problem of life" simply dissolved. So why should it not be the same with the hard problem of consciousness, a hundred years from now? As Strawson (2006) has pointed out, one cannot draw such a parallel unless one considers life completely apart from conscious experience. If consciousness is essential to life—and Vedantically conceived, life is essentially the power that executes what consciousness conceives—then life cannot be reduced to physics (via chemistry) if consciousness cannot be reduced to physics (via neurobiology).

1. have spatial extent (they occupy finite volumes of space),
2. are sufficiently stable (they neither collapse nor explode the moment they are formed), and
3. are made (or manifested by means) of finite numbers of objects that lack spatial extent.

Then we need to investigate the conditions under which ordinary objects are possible; in other words, we need to ask what the existence of such objects entails. What we find is that it entails the validity of quantum mechanics. Lastly, we need to ask why ordinary objects are made (or manifested by means) of finite numbers of objects that lack spatial extent. And the answer to this question is that the stage for Brahman's adventure of evolution was set by carrying the multiple exclusive concentration of consciousness to its furthest extreme, for this resulted in an (apparent) multitude of fundamental particles which, being formless, lack spatial extent.

But if Brahman's intention was to set the stage for a thoroughly evolutionary manifestation of its inherent qualities and powers, we can expect an evolutionary origin for all but the simplest structures that are instrumental in the manifestation of forms, and then it can be shown that not only quantum mechanics (the general theoretical framework of contemporary physics) but also many aspects of the well-tested special laws of contemporary physics (the so-called standard model of particle physics and the general theory of relativity) are needed to set the stage for evolution (Mohrhoff 2002, 2009, 2011).

Supermind Versus Mind

If the physical world were accessible to our senses on all scales of length, it would be differentiated all the way down. Taking for granted that this is the case, classical physics allows us to model reality from the bottom-up, either by explaining wholes in terms of interacting parts or by associating physical properties directly with the points of space, as classical field theories do. Quantum theory's "explanatory arrow" points in the opposite direction. If in our minds we go on dividing a material object into distinct parts, we reach a point at which the "parts" cease to be distinct. The attempt to (conceptually) divide the physical world into distinct parts leads to numerically identical particles and thus back to undifferentiated Being. We might say that ultimately there is but one "thing," and this is everything.

By the same token, if in our minds we keep partitioning the physical world into distinct regions of space, we reach a point at which the distinctions we make between regions no longer correspond to anything in the physical world. The spatial differentiation of the physical world is therefore incomplete—it does not go all the way down. If we choose to think of space as a continuous expanse, rather than as a family of relations, we have to think of it as an intrinsically undifferentiated expanse. We might then say that ultimately there is only one place, and this is everywhere. And much the same goes for the temporal differentiation of the world.[6] It follows that quantum mechanics does not permit us to model the world from the bottom-up. The world is structured ("built") from the top down, by a differentiation of Being that does not "bottom out."

The difference between the world of classical physics (differentiated all the way down and built from the bottom up) and the world of quantum physics (incompletely differentiated and structured top-down) reflects a difference between mind and the creative self-knowledge native to Brahman's primary poise which, following Sri Aurobindo, I call *supermind*. The action of supermind is primarily qualitative and infinite and only secondarily quantitative and finite. Mind is essentially the agent of supermind's secondary, quantifying, and delimiting action. If mind is employed by supermind, as it is in reality, its tendency to divide space and things ad infinitum is checked. This is why there are limits to the objective reality of the distinctions we make between things and between regions of space. If, on the other hand, mind is separated from its supramental parent and left to run wild, as it is in us, it not only divides ad infinitum but also takes the resulting multiplicity for the original truth or fact. This is why we tend to construct reality from the bottom up, and why we find it so hard to make sense of our fundamental physical theory. By implying that the world is created top-down, by a differentiation that does not bottom out, quantum mechanics is trying to tell us that the original creative principle is supramental rather than mental.

There remains the question of why Brahman = *sachchidānanda* would involve its infinite creative delight and its omnipotent consciousness-force in formless particles and a seemingly mechanical action. *Sachchidānanda* being what it is, there is only one possible answer: for "fun" (*ānanda*). In the physical world, Brahman is "playing Houdini," enchaining itself as best it can, challenging itself to escape, to re-discover its true self and its powers, to affirm itself in conditions that appear to be its very opposite—nonbeing[7] rather than *sat*, inconscience rather than *chit*, insentience first and then pain of every kind rather than *ānanda*. In the words of Sri Aurobindo (2005, 426–427):

> a play of self-concealing and self-finding is one of the most strenuous joys that conscious being can give to itself, a play of extreme attractiveness. There is no greater pleasure for man himself than a victory which is in its very principle a conquest over difficulties, a victory in knowledge, a victory in power, a victory in creation over the impossibilities of creation.... There is an attraction in ignorance itself because it provides us with the joy of

[6] In the physical world, temporal differentiation supervenes on spatial differentiation, for the existence of temporal relations requires something that can change, and in the physical world only the spatial relations can change.

[7] How does Brahman create something that appears to lack being? Think of space. Seen from the aperspectival poise of the original creative consciousness, it is a self-extension of what is at once, indistinguishably, an undivided substance and a single self. It is *sachchidānanda* extending itself to make room for variations. Now look at the same thing from the perspectival poise of a consciousness that has lost sight of its single self and, consequently, of the undivided substance constituting the world. In this poise of relation between self and world, space presents itself as a void, an extended nothing, a nonbeing, which nonetheless somehow exists.

> discovery, the surprise of new and unforeseen creation…. If delight of existence be the secret of creation, this too is one delight of existence; it can be regarded as the reason or at least one reason of this apparently paradoxical and contrary Lila.

Lila is a term of Indian philosophy that describes the manifested world as the field for a joyful sporting game made possible by self-imposed limitations.

Evolution

What is meant here by "evolution" is neither descent with modification nor the Darwinian process postulated to explain this historical fact. Essentially, evolution is the gradual reversal of the multiple exclusive concentration of consciousness which culminated in the creation of matter. Because life came to be involved in matter, life can evolve in matter; because mind came to be involved in life, mind can evolve in living matter; and because supermind came to be involved in mind, supermind can evolve in mentally conscious matter. But evolution does not simply retrace the steps that led to the creation of matter, for if it had done so, particles would have acquired forms. Evolution proceeds by integrating (rather than re-absorbing) the lower principles into the higher. When life appears, what is essentially added to individual forms is the power of executing ideas, and when mind appears, what is essentially added is the power of generating ideas. What has yet to evolve is a consciousness that is not exclusively concentrated in the individual, a consciousness aware of the essential identity of all individuals, a consciousness no longer confined to the perspectival outlook of a localized individual but capable of integrating its perspectival outlook into the supramental "view from everywhere," a consciousness informed with the infinite Quality/Delight at the heart of reality and capable of throwing it into mutable forms of its own immutable substance.

At bottom, all we can rationally understand is what can be reduced to laws. If there is something that is inexplicable in terms of natural laws, we consider it random. Because evolution has aspects that cannot be explained in terms of natural laws, the rationalist is compelled to attribute the origin of species to random mutations, in addition to environmental selection pressures and biological processes that are intelligible in terms of natural laws.

Don't get me wrong. I am not an advocate of Intelligent Design. (Nor am I an advocate of stupid design.) The constraints under which a designer works are different from the constraints under which evolution works. If Brahman has the power to enter into reflexive relations and to subject them to physical laws, then it also has the power to modify these laws. If there are limits to this power, they are self-imposed. The difference is that a designer makes use of the physical laws without being able to change them, whereas evolution works through modifications of these laws.

The objection may be raised that modifications of the laws of physics have never been observed. But this is what we should expect. Given the Houdiniesque purpose of this evolutionary manifestation, it stands to reason that the range of possible modifications will be seriously limited—so limited that no presently feasible experiment can reveal statistically significant departures from what the physical laws predict.

The Force at work in the physical world has two aims to pursue. The first is to bring into play the creative powers of life and mind—the power of executing ideas and the power of generating ideas. Because it has to accomplish this through tightly constrained modifications of the physical laws, the evolution of life necessitates the creation of increasingly complex organisms, and the evolution of mind necessitates the creation of increasingly complex nervous systems. The second aim is to express, at any stage in the course of evolution, by whatever means available at that stage, the infinite Quality at the heart of reality. (How could the angiosperms not be the works of accomplished artists? What if not a frenzy of creative ecstasy could have produced the arthropods?).

It used to be said that qualities (like colors and sounds) are "nothing but" quantities (such as electromagnetic or acoustic frequencies). It would be much closer to the truth to say that quantities are nothing but means of manifesting qualities. And here I am not speaking merely of sensory qualities; I am also speaking of the transcendental qualities of beauty and goodness. The reason this is not obvious is that the dynamic link between quality and form is inaccessible to a consciousness whose characteristic activity is the formation of ideas. This link is accessible only to a consciousness whose characteristic activity is the development of quality into expressive ideas—a supramental consciousness that is directly aware of the qualities to which it gives expression. If our social world exhibits an appalling lack of the good and the beautiful, it is for two reasons. The first is that such a consciousness is yet to evolve, and the second is that it would not evolve if we were not duly appalled.

Brahman's power to modify its self-imposed constraints cannot be explained in terms of another self-imposed constraint. In other words, it cannot be reduced to laws, and therefore we have no way of knowing how it works. If the constraints were loosened, more would become possible while less would be comprehensible. If the constraints were removed, everything would become possible and nothing would remain comprehensible to our mental way of knowing. The evolution of supermind will remove the constraints. As you will remember, it was due to these constraints that the evolution of life had to depend on the creation of complex organisms, and that the evolution of mind had to depend on the creation of complex nervous systems. Once the constraints are removed, this complexity will have served its purpose. Matter will no longer offer any resistance to the executive force of life, nor will life offer any further resistance to the ideative faculty of mind. Fully integrated into the supramental dynamism, both life and mind will participate in the unhampered development of quality into fully expressive material forms.

All of this seems perfectly preposterous, to be sure, but here is why: we tend to conceive of the evolution of consciousness as an emergence of increasingly successful adaptations to, if not increasingly faithful representations of, a pre-existent world. And we tend to think of science as being in the business of devising more and more faithful models of such a world. From the Vedantic point of view, this is a serious mistake, for it blinds us to the possibilities of the future. The evolution of consciousness consists in the successive emergence of increasingly rich and complex ways in which Brahman presents itself to itself, not in the progressive uncovering of a pre-existent world.

It will be instructive to contrast our present consciousness structure (to borrow a term coined by the cultural historian and philosopher of evolution Jean Gebser) with the one that preceded it. Take the ancient notion that the world is enclosed in a sphere, with the fixed stars attached to its boundary, the firmament. We cannot but ask: what is beyond that sphere? Those who held this notion could not, because for them the third dimension of space—viewer-centered depth—did not at all have the reality it has for us. Lacking this dimension, the world they experienced was in an important sense two-dimensional. This is why they could not handle perspective in drawing and painting, and why they were unable to arrive at the subject-free stance which is a prerequisite of modern science, and which Thomas Nagel (1986) has called “the view from nowhere.” All this became possible with the consolidation, during the Renaissance, of what Gebser (1985) has called the mental structure of consciousness, which superseded what he termed the mythical structure. While the latter structure’s characteristic way of making sense of the world was through myths, the way we, at this point in history, attempt to make sense of the world is through science and rational philosophy.

As the mythical consciousness structure was superseded by the mental, so the mental structure will be superseded by a structure Gebser termed integral, and which he equated with the consciousness Sri Aurobindo termed supramental. And just as mythological thinking could not foresee the technological explosion made possible by science, so scientific thinking cannot foresee the radical consequences of the birth of a new world, brought about by the evolution of the integral structure.

Our very concepts of space, time, and matter are bound up with, are creations of our present, characteristically three-dimensional consciousness structure. It is not matter that has created consciousness; it was consciousness that has created matter, first by carrying its multiple exclusive concentration to the point of being reduced to an apparent multitude of formless particles, and again by mutating into our present consciousness structure, which is capable of integrating our location-bound perspectives into a subject-free world of three-dimensional objects. Ahead of us lies the evolution of a consciousness that transcends our time- and space-bound experience, a consciousness to which our theoretical dealings with the world will seem as dated as the mythical explanations of the pre-scientific era seem to us. To this characteristically four-dimensional consciousness matter will be transparent, revealing its ultimate constituent as well as the identity of the latter with the ultimate subject of all consciousness.

Map and Territory

Quantum theory requires us to distinguish between two kinds of measurable quantities and two corresponding domains—quantities belonging to a microscopic or quantum domain, which possess definite values only when they are measured, and quantities belonging to the macroscopic or classical domain, which possess values that are definite per se. Half of the crux of the aforementioned measurement problem lies in understanding the origin of the definiteness of these values. The other half concerns the statistical character of quantum mechanics, which I have already addressed, without mentioning that it too makes it necessary to distinguish between those two domains. The difficulty of understanding why we need this distinction has bedeviled the interpretation of quantum mechanics from the get-go. Yet it is readily understood once we recognize it as the distinction between the manifested world and what is instrumental in its manifestation.

Why, then, are the properties of the manifested world definite per se? The reason is that when we speak of the manifested world, we mean the world manifested *to us*. We mean the world that Brahman manifest to itself at the present stage of the evolution of consciousness. It is only in our experience that measurement pointers have definite position and, consequently, that measurements have definite outcomes. After countless ways have been tried to disprove this, it has become increasingly clear that the origin of definiteness—like that of so many other features of the "objective" world—lies in the nature of human conscious experience. Niels Bohr was right all along, insisting as he did that "in our description of nature the purpose is not to disclose the real essence of the phenomena but only to track down, so far as it is possible, relations between the manifold aspects of our experience" (1934, pp. 17–18).

But if all that science can do is track down relations between the manifold aspects of our experience, then what we call "nature" or "the physical world" is a construct of such relations—grammatical or logical relations like the relation between subject and predicate, to which we owe the concepts of substance and property, and spatiotemporal relations like the relation between here/now and there/then, to which we owe the concepts of causality and interaction—as Immanuel Kant (1781/1998) has shown. And then there can be no question of science mapping the territory of a pre-existent world. If there were such a world, we would have no concepts to describe it, as Bishop Berkeley (1710) has shown.

To what, then, can the metaphor of map and territory be applied? We tend to take a mind-constructed map for the territory in which evolution takes place. But if this "map" is exclusive to a particular structure of consciousness and thus limited in time, what could be the territory in which consciousness itself evolves? If we think of the different structures of consciousness as maps, what could be the territory in which one map is replaced by another? In a Vedantic context, the obvious answer is: the supramental consciousness in which "this apparently paradoxical and contrary Lila" takes place.

If there is a real world beyond the world constructed by our minds, it is not a world existing out of relation to consciousness altogether, but the world as it exists —as it is perceived and by being perceived created—in the primary poise of relation between Brahman qua subject and Brahman qua object. Sri Aurobindo (2005, p. 143) speaks of "a consciousness higher than Mind which should regard our past, present and future in one view, containing and not contained in them, not situated at a particular moment of Time for its point of prospection," a consciousness "not situated at any particular point of Space, but containing all points and regions in itself," and he adds that

> At certain moments we become aware of such an indivisible regard upholding by its immutable self-conscious unity the variations of the universe. But we must not now ask how the contents of Time and Space would present themselves there in their transcendent truth; for this our mind cannot conceive,—and it is even ready to deny to this Indivisible any possibility of knowing the world in any other way than that of our mind and senses.

This indivisible regard, upholding by its immutable self-conscious unity the variations of the universe, is the territory. And so it is when this indivisible regard manifests itself in a species of supramentally conscious beings that the maps will have become one with the territory.

References

J.S. Bell, On the Einstein Podolsky Rosen paradox. Physics **1**, 195–200; reprinted in Wheeler, Zurek, 403–408 (1983)

G. Berkeley, *A Treatise Concerning the Principles of Human Knowledge* (1710)

N. Bohr, *Atomic Theory and the Description of Nature* (Cambridge University Press, 1934)

D.J. Chalmers, The puzzle of conscious experience. Sci. Am. **273**, 80–86 (1995)

J. Gebser, *The Ever-Present Origin* (Ohio University Press, 1985)

P. Heehs, *The Lives of Sri Aurobindo* (Columbia University Press, 2008)

I. Kant, *Critique of Pure Reason*, trans. and ed. by P. Guyer, A.W. Wood (Cambridge University Press, 1781/1998)

U. Mohrhoff, Why the laws of physics are just so. Found. Phys. **32**(8), 1313–1324 (2002)

U. Mohrhoff, Quantum mechanics explained. Int. J. Quantum Inf. **7**(1), 435–458 (2009)

U. Mohrhoff, *The World According to Quantum Mechanics: Why the Laws of Physics Make Perfect Sense After All* (World Scientific, 2011)

U. Mohrhoff, Consciousness in the quantum world: an Indian perspective, in *Quantum Physics Meets the Philosophy of Mind: New Essays on the Mind-Body Relation in Quantum-Theoretical Perspective*, ed. by A. Corradini, U. Meixner (De Gruyter, 2013), pp. 85–97

U. Mohrhoff, Manifesting the quantum world. Found. Phys. **44**(6), 641–677 (2014a)

U. Mohrhoff, Quantum mechanics and the manifestation of the world. Quantum Stud. Math. Found. **1**(3–4), 195–202 (2014b)

U. Mohrhoff, The quantum mechanics of being and its manifestation. Cosmology **24** (2016). http://cosmology.com/ConsciousnessUniverse3.html

U. Mohrhoff, Quantum mechanics in a new light. Found. Sci. **22**(3), 517–537 (2017)

T. Nagel, *The View From Nowhere* (Oxford University Press, 1986)

S. Phillips, *Classical Indian Metaphysics* (Open Court, 1995)

B. Russell, *The Analysis of Matter* (Routledge, 1927)

E. Schrödinger, Die gegenwärtige Situation in the Quantenmechanik. Naturwissenschaften **23**, 807–912, 823–828, 844–849 (1935); trans. by J.D. Trimmer, The present situation in quantum mechanics, ed. by Wheeler, Zurek (1983), pp. 152–167

Sri Aurobindo, *Kena and Other Upanishads* (Sri Aurobindo Ashram Publication Department, 2001)

Sri Aurobindo, *Isha Upanishad* (Sri Aurobindo Ashram Publication Department, 2003)

Sri Aurobindo, *The Life Divine* (Sri Aurobindo Ashram Publication Department, 2005)

G. Strawson, Realistic monism: why physicalism entails panpsychism. J. Conscious. Stud. **13**(10–11), 3–31 (2006)

J.A. Wheeler, W.H. Zurek (eds.), *Quantum Theory and Measurement* (Princeton University Press, 1983)

Part II
Theoretical Physics

Chapter 8
How We Make Sense of the World: Information, Map-Making, and the Scientific Narrative

Marcelo Gleiser and Damian Sowinski

> *All philosophy is based on two things only: curiosity and poor eyesight.*
> *The trouble is, we want to know more than we can see.*
>
> —Bernard le Bovier de Fontenelle

Introduction

The physicist Werner Heisenberg, one of the founding fathers of quantum mechanics, put it clearly: *What we observe is not Nature, but Nature exposed to our method of questioning*. Excluding those lost in solipsistic confabulations, most people would agree that there is a world external to us, a world that we can apprehend, however indirectly and incompletely, through our sensorial experience, augmented (or not) by adequate instruments. Science's central goal, in a nutshell, is to construct meaning from what we perceive of the world. Given that what we can perceive of the world is contingent on how we look at it, and that the ways we have been looking at the world have changed with technological advances and consequent shifts in perspective, our constructed meaning of the world is a work in progress: ontologies change due to shifts in epistemic strategies. What the world is made of and how it operates are notions frequently revised.

Our relationship with the world is affected through the gathering and exchange of information. Experiences are events that promote information flow. This is true as we trek along a mountain path, as we communicate with another human, as we dream, or as we read gauges in an experiment. As information flows, our awareness state gets

M. Gleiser (✉)
Institute for Cross-Disciplinary Engagement, Dartmouth College, Hanover, NH 03755, USA
e-mail: mgleiser@dartmouth.edu

D. Sowinski
Department of Physics and Astronomy, Dartmouth College, Hanover, NH 03755, USA
e-mail: Damian.Sowinski@Dartmouth.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_8

updated. Senses gather outside information which is then processed in our brains, allowing us to evolve our awareness state. Thus informed, we make choices: take another step forward along the path; or don't, given that you reached a precipitous cliff. From this simple example, we see that the most informative experiences are those that cause sharper changes in our awareness state.

This conceptualization is not limited to humans, being applicable to any agent, animal or machine, capable of awareness, broadly defined here as the ability to sense the environment. For example, a thermostat installed in an air conditioner is an awareness-changing device: it gathers information (data) from the outside environment, sensing changes in temperature. Such changes, in turn, will affect the air conditioner's awareness state, causing it to adapt its functioning: if the temperature rises, work harder; if it cools, turn off. We call such simple agents *Map Walkers*. They respond to experiences by changing their awareness state but are incapable of self-awareness. They dwell on their perceived Map by updating their awareness states through the flow of their experiences. However, they cannot conceive a Map. Different Map Walkers experience different maps due to their different sensory apparatuses: a bacterium and an earthworm are both Map Walkers, although they move on very different Maps. Their experiences of reality are profoundly different.

What differs radically between a human and an air conditioner or an earthworm is the ability to contextualize experience subjectively. Humans are aware not only of their environment, but also of their own awareness: they are self-aware. A machine or simple animal can process but cannot discern and internalize the notion of surprise, but humans can. The most dramatic experiences often surprise us and cause sudden, discontinuous jumps in our awareness state. It is the awareness that we have of these abrupt changes that quantifies surprise. The more surprising the experience, the higher the jump. These are the experiences that carry the most information. Map Walkers will, of course, also respond to a sudden change in conditions if they are programmed—algorithmically or biologically—to do so. Case in point, a self-driving car must react quite quickly if it senses a child running after a ball in front of it. A bacterium driven by chemotaxis will change course if it senses a change in nutrient concentration. However, we cannot associate the notion of surprise to a preprogrammed response in a Map Walker because they are not aware of the changes in their awareness. Perhaps the essential difference between a self-aware and an aware agent is that a preprogrammed Map Walker cannot choose what response it will have for a given stimulus beyond some kind of optimization code. We can.[1]

[1] We are thus making a distinction between awareness and self-awareness. For our purposes, awareness is the property of connecting with both the external and internal environment (awareness states) through an exchange of information; self-awareness is the property of having subjective knowledge of one's own existence. Given the current lack of knowledge on the nature of consciousness, we consider it here as the black box that endows us (and other presumably self-aware animals or alien creatures) with this subjective capacity.

In an effort to conceptualize the role of information in our sense-making of the world, we will therefore focus here mostly on self-aware agents. We call these complex agents *Map Makers*. Note that Map Makers are also Map Walkers but clearly not vice-versa.

Note to Non-mathematical Readers: Sections "From Language to Belief" and "Information Hidden in the World" have some formulas, which we use to make our general statements more quantitatively precise. If you are not versed in math, don't worry. Skip the formulas. We tried hard to be as explanatory as possible in the text before and after the equations so that you can still capture the general sense of our ideas.

Beyond Shannon: Bringing on the Subjective

If information can be found in the change that is induced in an agent's ability to contextualize experiential data, then, at its heart, information is an epistemic, not ontic, quality: Information is not found in the world; it is found in interactions with the world. This is in contrast to the way in which information was introduced in Claude Shannon's *A Mathematical Theory of Communication*, the paper that gave birth to the field of information theory (Shannon 1948). It is important to recall that Shannon was interested, as his paper bluntly puts it, in communication: data sent from a source to a receiver over some channel. To Shannon, information is the reliable transmission of messages over that channel. A channel can relay messages perfectly at a rate no greater than its channel capacity, given the optimal coding scheme. Here we glimpse a thread that connects Shannon's view of information as being a reliable transmission of messages, and the epistemically subjective notion of information which we began to develop above: Shannon's result relies on the existence of an internal mechanism, an encoding.

Here's a useful analogy. Imagine someone pouring one liter of water into a funnel. The sender is the agent pouring the water, the one liter of water is the message, the funnel is the channel, and the receiver is on the other side of the funnel. If water is poured too quickly, it will overwhelm the funnel and leak from the top: the channel is not being used to its maximum capacity and some of the message will be lost. If water is poured too slowly, the operation will take longer than needed. To maximize the channel's capacity, the sender must pour water into the funnel at the same rate that it can flow from its narrow bottom. This is the optimal encoding for this operation and it depends on the sender's prior knowledge of the channel he is operating with. Only with that knowledge can the sender optimize his encoding.

Message transmission requires an alphabet. What Shannon discovered was that even in a noisy channel there is an optimal—*error-free*—transmission rate known as the channel capacity. An error-free transmission rate is referred to as having perfect fidelity. In the analogy above, not a drop of water is lost. So, to maximize the efficiency of a transmission, we need to know the alphabet and, in particular, the frequencies that different letters appear. This way, if a letter appears more frequently, it will be more efficient to encode it with fewer bits than one that appears only rarely. Think of a bit as the fundamental unit of storage, taking on only two possible values: on or off. In the English language, the letter e appears about once every 8 letters, while q appears once every 1000 letters (Sowinski 2016). It would be very wasteful to encode both with the same number of bits. Shannon's result then tells us that if we use the natural frequencies of letters in a language with an alphabet $\mathcal{A}$ with N symbols, $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$, each appearing with a frequency $p(a_i) = p_i$, then the optimal encoding scheme will use $\log \frac{1}{p_i}$ bits of information to encode the ith letter a_i of the alphabet. His magnum opus then relates the channel capacity to the average number of bits needed to store the entire alphabet, the Shannon Entropy:

$$S[\mathcal{A}] = -\sum_{i=1}^{N} p_i \log p_i \tag{8.1}$$

Here, again, we see subjectivity crawling into Shannon's results. Information, as measured by the optimal rate of message transmission, relies on an internal knowledge of the frequencies of letters in a specific language.

We thus see that in Shannon's approach to communication, a knowledge of the distribution of letters in a language is necessary. This knowledge is encoded in a probability distribution over the alphabet. For practical purposes, these distributions can be found by examining a large corpus of literature, and extracting the frequencies of each letter. This is a statement concerning *contextuality*: Without the context of the language being used, the maximal rate of message transmission is meaningless. Only Map Makers are capable of creating and encoding meaningful messages. Map Walkers are locked into their programs and do not search for meaning or experience doubt. For them, there is just doing, as their state of awareness is updated. An air conditioner may adjust its functioning due to an input from its thermostat, but its communication with its thermostat, their message exchange, is locked into a one-dimensional realm. A laptop will perform commands as programmed, but will not initiate its own encoding or willfully depart from its program. To be aware as defined above is not enough to be able to question an algorithmic command; to doubt and choose one needs to be self-aware.

Let us turn then to the context, namely the probability distribution of letters in the alphabet. Here the term probability relies on a frequentist interpretation: probabilities are the frequencies with which letters occur. The frequentist interpretation of probability is useful in this context, as it is in games of chance and any other repeatable experiences. However, it relies on a sort of unrealistic notion wherein the full corpus of all possible messages is used to define the probability with which letters occur. Realistic senders and receivers are unlikely to have this full corpus. They need to use probability distributions that are inferred from incomplete data. Map Makers are always partially ignorant to what they are mapping. If we consider science as a Map of physical reality, it follows that scientific theories will always be incomplete maps, given that the information we collect from physical reality (the territory) is inferred from incomplete data.

Probability distributions are inferred from data. This is important, for it is a statement about the epistemic state of the sender and the receiver. The connection here goes beyond probabilities of letters, and brings us to the Bayesian interpretation of probability. Unlike the frequentist point of view, the Bayesian perspective concerning probabilities is that they are not absolute quantities forced on thinking things by the external world, but epistemic measures of the strength of belief thinking creatures like ourselves have about the world (Jaynes 2003). This interpretation has great strength, as it opens the possibility of defining the probability of not just repeatable events, but of unique ones.

The connection between statements about the world, whether they concern repeatable or unique events, and epistemic states will be explored in the next section. We will use an idealized thinking entity, that we will call Idealized Epistemic Agent (IEA), to be defined below. Propositional logic will be developed to understand how an epistemic agent talks about the world (Enderton and Enderton 2001; Cox 1961). Following the work of Cox, probability theory will be derived from an epistemic theory about belief, and information will be shown to emerge from processes in which epistemic agents have experiences that alter their beliefs (Cox 1946).

Using IEAs as idealized Map Makers, the central thesis of the beginning of this chapter, that information is contained in experiences that change us, will be put on a solid quantitative foundation. We will derive Shannon's information measure on purely epistemic grounds, by considering how information is hidden from epistemic agents by the world and revealed (partially) through experience.

The transmission of messages is then the continuous conversation that a Map Maker has with the world; the latter sends messages, in the form of experiences, to the former. The results of Shannon's theory then become statements concerning the optimal rate at which Map Makers (epistemic agents) can extract information from the continuous stream of experience. In other words, they quantify how good a map

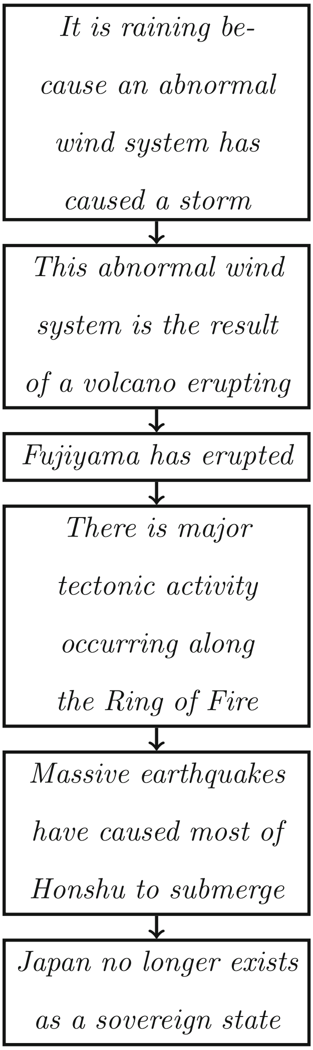
Map Makers can create. Unlike the receiver in the theorems, Map Makers do not have a fixed set of beliefs, a probability distribution. Their interaction with information will cause their beliefs to change: maps get updated.

Epistemic Agents as Idealized Map Makers

Given the obvious difficulties in defining human consciousness, in order to make conceptual progress on how we relate to the world we need to introduce an idealized Map Maker: an Idealized Epistemic Agent (IEA). Humans will be imperfect approximations to IEAs in ways that will be clarified in the following. Humans would be EAs, not IEAs. Jaynes referred to these inferential automatons as robots (Jaynes 2003), while Caticha as idealized rational agents (Caticha 2009). We prefer the term IEA, since it places emphasis on two aspects that are central to the following: thought and agency. The first of these is encapsulated by an EA having a web of consistent beliefs (Caticha 2008). The latter is found in an EA's ability to go beyond simply having experiences as do Map Walkers: EAs generate experiments that evolve their beliefs.

Idealized Epistemic Agents are entities that use an idealized, perfect, language to talk about and conceptualize the world around them. Unlike human beings, IEAs are infallible in the use of language. Their understanding of language is completely structural, and when they analyze the truth or falsity of a statement, they do so in the context of all possible statements that exist in the language. When a particular statement's truth value is established by an IEA, all statements that are logically connected to that statement instantly feel that update. An IEA cannot hold within itself contradictions, because all possible statements that could be said in the language are always present within the *mind* of the IEA. Humans are not so lucky; we do not have the cognitive resources that an IEA has. One may consider paradoxical statements, such as *this statement is a lie*. For now we must assume that these types of self referential statements do not exist within the language.

An IEA's ability to have access to the complete set of statements has profound consequences to the way they make inferences during experiences. For us, an experience does not change all our beliefs in an instantaneous fashion, but for an IEA this is not the case. Learning that it is raining has no effect on our belief that the Yen is the currency used in Japan. An IEA, however, does not see these statements as being completely independent. By pinning down the truth value of rain occurring at this very moment, the IEAs beliefs instantly respond, no matter how tenuously connected they may seem to us:



It is raining because an abnormal wind system has caused a storm

This abnormal wind system is the result of a volcano erupting

Fujiyama has erupted

There is major tectonic activity occurring along the Ring of Fire

Massive earthquakes have caused most of Honshu to submerge

Japan no longer exists as a sovereign state

(With apologies to our Japanese friends!) This chain is so implausible that the average human being wouldn’t register differences in belief in the links beyond the first. The effect of an experience to an IEA is a change, however seemingly insignificant, to its belief in everything else. It doesn’t matter how far-fetched a line of reasoning goes, all that matters to an IEA is the structure of language: An IEA responds to experience through an instantaneous updating of all of its beliefs.

Humans require time to process experience and to fully understand the implications it may have on our beliefs. Because processing takes time, experiences for us overlap with one another, blurring the borders between moments. Data from multiple experiences might get mixed up, leading to mistakes in our inference schemes. This is not the world that IEAs experience. Their ability to instantly process ensures that any experiences separated by finite time intervals, however small, will update the complete set of beliefs the IEAs have in the order in which they happen. It is important to keep this distinction in mind in what follows.

From Language to Belief

Language is that which is used to convey states of the world, both to ourselves, and to our IEAs. A language is thus a tool to allows an EA to describe not just the state of the world, but all possible states of the world, and therefore all possible worlds. A particular state in the world relies on pinning down the truthfulness of certain statements in the language, which is accomplished via a *truth assignment*. A language must also have the capacity to bring together statements to construct new statements, which is accomplished via **logical connectives**. A language is *rational* over the possible worlds on which its logical connectives are consistent. (For technical aspects of logical language construction consult (Sowinski 2016)). Possible worlds in which this does not hold are *irrational* with respect to the language: the language is incapable of consistently describing those worlds. Hence even though the machinery of a language can be used to talk about all possible worlds, only those worlds on which the language can say things clearly can truly be spoken of.

One mustn't confuse the use of the term language in what follows as a case of a natural language such as English, which will be used in examples. The language that IEAs use is, like them, an idealization. It is assumed that this language is sufficient at describing all the experiences that an IEA can have. Clearly, that's not the case for any human dialect.

Now, some definitions. A language is made of two kinds of statements, **atomic** and **compound**. As the name suggests, atomic statements are irreducible, that is, cannot be decomposed into simpler statements: *it is raining*, *the cat is alive*, and so forth. A compound statement could be, *it is raining AND the cat is alive.* A language thus consists of atomic statements and logical connectives. An epistemic state is a truth assignment on the set of atomic statements. The set of all epistemic states with all possible truth assignments is called the epistemic realm. (In the example above, the epistemic realm would include: *it is raining*; *it is not raining*; *the cat is alive*; *the cat is not alive*; *it is raining AND the cat is alive*; *it is raining AND the cat is not alive*; etc. It would not include *it is not raining AND cat.*) Given its experience of the world, an IEA would assign a truth value to each of these. The subset of the epistemic realm on which the language is consistent with its truth assignment is the realm of discourse. A language is rational on its realm of discourse.

Compound statements form the bulk of the statements that are used in common discussions between humans. If humans disagree on the truth value of even one statement, then they are not talking about the same world. Arguments and experiences may cause us to change truth values. We need to see how to incorporate these dynamics in IEAs. Furthermore, truth values of statements alone are not enough to capture the possibility that we may be ignorant of the truth or falsehood of a statement: I am not sure whether the statement *It will rain tomorrow* is true. These considerations lead us to another layer of epistemology, where we must leave behind absolutes, and consider possibilities.

Upon fixing a language, a realm of discourse is established concerning what an epistemic agent may now think about. There are many possible worlds in the realm of discourse. Many such worlds overlap for the same truth value of a particular statement, or set of statements. An epistemic agent prioritizes the value of the truth of statements via a belief function. More precisely, for every statement s, an IEA has a belief about the statement, $b(s|K)$. Here we have introduced the prior knowledge of the IEA through the variable K. The term $b(s|K)$ should be read as *the strength of belief than an epistemic agent with prior knowledge, K, has about the statement s.* In this sense, the epistemic agent prioritizes worlds through belief, and considers the ontological state of the world to coincide with the epistemic world which maximizes their belief. What an epistemic agent believes the world to be is that which it is. This is true for IEAs and for human epistemic agents. Since beliefs are meant to prioritize statements, they are an ordering imposed by the IEA.

Beliefs are transitive, in the sense that if the epistemic agent believes A more than B, and B more than C, then they must believe in A more than C. This structure restricts the possible objects that one can use to describe beliefs, which is very useful. The transitive property means that we can represent beliefs as real numbers, which we write as $b(s|K) \in \mathbb{R}$. Belief, then, can be represented mathematically as a function from the set of all statements to the real numbers, $\mathbb{R}$, given the background knowledge of the IEA. Ab initio this background knowledge pertains to the choice of logical connectives chosen by the IEA. As the IEA incorporates dynamics in the form of experience, background knowledge will grow to include these experiences.

Though one may at first think of belief as being binary—one either does or does not believe in a statement s—this is not the case. Beliefs are graded, and the binary nature of belief only comes about due to a self-imposed threshold. An IEA says they believe in a statement when their belief function is beyond this threshold. Thus, an IEA's belief in a statement could start off at any value. Experience may nudge this value by tiny amounts until finally the threshold is reached, and even then it is possible for the IEA to continue to strengthen their belief in a statement, all the while declaring that they believe. We therefore postulate that the belief function is a continuous function, the first of the Cox axioms (Cox 1946).

What about the range of the belief function? Can an EA have a belief of 7 in one statement, and $-\pi$ in another? Is there a maximum strenght of belief, or a minimum? To make things worse, it is unclear whether higher numbers correspond to stronger or weaker belief. In fact, every EA can have a different range of belief. All that matters to an IEA is the ordering of beliefs. This freedom in choosing a range needs to be remedied, since in the end we will want IEAs to be able to compare their beliefs; science is, after all, a dialogue not only between the scientist and reality, but between scientist and scientist. Changing the scale of belief is accomplished via regraduation. This is very similar to choosing a different temperature scale in thermodynamics. One chooses the scale in order to make things as simple as possible; scales are chosen on pragmatic grounds, as when you go to the doctor and there's a form to fill up that says *On a scale from* 0 *to* 10 *how's your level of pain today?* The scale restricts everyone to the same range. This is what regraduation does, it rescales the beliefs of all IEAs, while maintaining the ordering of belief, so as to make different IEAs share the same scale. Regraduation is intimately related with the logical connectives of the language, in that the freedom to choose a scale is constrained only by the demand that the belief function works coherently with the structure of the language. For example, consider a statement and its negation, $\{s, \neg s\}$. An IEAs strength of belief in one of these depends on its strength of belief in the other. We can write this relationship as $b(\neg s|K) = F(b(s|K))$, where F is shorthand for the relationship. Now the structure of the language kicks in. Since double negation cancels out (think *it is not not raining*), we have that $b(s|K) = b(\neg\neg s|K) = F(b(\neg s|K)) = F(F(b(s|K))$. This puts a severe constraint on the form of the relationship F, namely that:

$$F(F(x)) = x \tag{8.2}$$

This *functional equation* is quite famous, and is known as *Babbage's equation*. Babbage's work on the analytical engine laid the foundation for modern day computers, and the polymath worked on solving this equation as far back as 1815 (Dubbey 1978). Another such functional equation for the relationship $G(x, y)$, which is constrained by the logical connective *and*, reads:

$$G(G(x, y), z) = G(x, G(y, z)) \tag{8.3}$$

This relationship is called the *Associativity equation*, and its beautiful solution can be found in Aczél's comprehensive lectures on functional equations (Aczel 1966).

Using the results from the analysis of both (8.2) and (8.3), it can be shown that belief in a truthhood must take on the value 1 (Cox 1946; Jaynes 2003). Surprisingly, falsehood can either be represented as 0 or infinity. Since truth and falsity are typically represented with a 1 and 0, respectively, we make the choice of representing the belief in a falsehood (the least amount of belief that one can have in a statement) with a 0. This regraduation also creates a particularly recognizable relationship between beliefs constrained by logical connectives, namely that the sum of the belief in a statement and its negation must equal 1:

$$b(s|K) + b(\neg s|K) = 1. \tag{8.4}$$

A similar result is familiar from the theory of probability. In fact, the structural relationships between beliefs turn out to be identical to those of probabilities after regraduation, giving weight to the Bayesian interpretation of probability. Probabilities are not ontic aspects of reality, but epistemic strengths of belief within IEAs.

From Belief to Information

Our working premise is that information is that which changes belief (Caticha 2009). For an experience to be informative, the EA having the experience must be changed upon its conclusion: her awareness state gets updated. Let us take a look again at the process of having an experience. Consider an EA with a prior set of beliefs about different statements s based on her background knowledge K, $b(s|K)$. After the experience, she has a posterior set of beliefs where the experience e has been incorporated into her background knowledge, $b(s|e \wedge K)$. The symbol $\wedge$ represents the now expanded background knowledge incorporating the experience e to the previous background K. This represents an update of the EA's state of awareness. The two states (before e and after e) are related by the so-called likelihood, $\mathcal{L}$:

$$b(s|e \wedge K) = \mathcal{L}(e; s, K)b(s|K) \tag{8.5}$$

Note that the likelihood is not a belief, since it can be less than, greater than, or equal to 1. It is, however, a very important quantity epistemically. If, for a given statement s, the likelihood is greater than one, then it strengthens the EAs belief in that statement. If it is less than one it weakens it. It acts on the old belief distribution multiplicatively to create the new belief distribution.

This procedure can be generalized to an arbitrary number of sequential experiences. These experiences will update the IEAs beliefs. One can either consider the experiences sequentially, or treat them all simultaneously as a single experience (for Idealized EAs only, not for human EAs). The equivalence of these two is a statement of the *holistic nature of experience*: an IEA may partition her experiences in any way she chooses, but this does not affect her final belief.

The holistic nature of experience assumes an equivalence between a causal (sequential in time) series of experiences and an acausal—*all-together*—simultaneous piling up of experiences. Clearly, this points to what is missing in the theory, since if we are to assume that the beliefs of an EA require a physical substrate, then there must be some ontological cost to changing belief. Moving through a series of in-between beliefs will then not necessarily have the same cost as moving from the initial to final belief. For example, an EA which throws a die will experience a change in belief due to the outcome. They may then draw a card from a deck, which, too, will change their belief function. For the holistic nature of experience to hold, the resulting belief after both experiences must be the same in both sequential and simultaneous updating. In the case where both are performed sequentially, we would expect that the ontological substrate would have undergone more changes to get to its final state than if they were performed simultaneously. The excess change in sequential versus simultaneous updating would be the source for this ontological cost difference, which may not be negligible. Indeed, any updating operation involves a thermodynamic cost with a resulting entropy increase. The holistic nature of experience, as an idealization, assumes the same heat loss in both cases. Despite this shortcoming, it will provide us with the conceptual basis to relate Shannon's entropic formula to the information hidden in the world as experienced by an IEA. The key comes from our working premise: Information is that which changes belief. We have seen that experience changes belief from prior to posterior via the likelihood function. It is to the likelihood function, then, that we must turn our gaze if it is information that we are trying to understand.

Information Hidden in the World

Assuming the holistic nature of experience, we can also consider how independent experiences affect the likelihood. Independent events are ones whose outcomes do not affect each other. For example, throwing a die and drawing a card are independent actions. The order of the events is irrelevant. Picking a dessert and eating dinner can be dependent experiences, since their order influences one another. For independent experiences, the likelihood is the same whatever their order. Within this framework, we posit that the information contained in an experience e, about a statement s, to an IEA with background knowledge K, is a function f of the likelihood,

$$I_s(e|K) = f[\mathcal{L}(e; s, K)]. \tag{8.6}$$

We can quickly say a few things from this general statement. First, since an uninformative experience doesn't change belief, it must contain no information. Mathematically: $f(1) = 0$. Second, as the likelihood changes by an infinitesimal amount, we do not expect the amount of information contained in the experience to change discontinuously, so the function f must be continuous. Lastly, the information gathered from independent events must reflect the commutativity of their temporal ordering: $f(xy) = f(x) + f(y)$. It can be shown that the only function that satisfies these properties is the logarithm, allowing us to write

$$I_s(e|K) = A \log[\mathcal{L}(e; s, K)], \tag{8.7}$$

where A is an arbitrary constant.

Since the likelihood is the ratio of prior and posterior beliefs, we can use the properties of logarithms to rewrite the information as:

$$I_s(e|K) = A \log b(s|e \wedge K) - A \log b(s|K), \tag{8.8}$$

expressing the information gained by the IEA due to experience e as the change between final and initial states motivates the definition of **hidden information**:

$$h(s|K) = -A \log b(s|K). \tag{8.9}$$

Why call this hidden information? Equation 8.8 tells us that information is the negative change in h. So if the information contained in an experience is positive, then h must have decreased; similarly, if the information contained in the experience is negative, then h must have increased. This is simply a relationship between revealed and hidden information. If that which is revealed increases, then that which was hidden has decreased, and vice-versa. Furthermore, since we would like the total information hidden by the world to be positive, this means that A is a positive constant. The arbitrariness of the constant can be absorbed into the arbitrariness of the base of the logarithm being used. We therefore write

$$\begin{aligned} h(s|K) &= -\log b(s|K) \\ &\Downarrow \\ I_s(e|K) &= -\Delta h \\ &= h(s|K) - h(s|e \wedge K) \end{aligned}$$

When an IEA has an experience e about a statement s there is a change in hidden information. This holds for EAs as well. *The fact that the map can never be completely faithful to the territory is, within this framework, an expression of the fact that the hidden information can never be zero; the world will always hide information from an EA.*

There is a nuance in the above that must be addressed. An experience changes not only the belief in a statement, but also belief in the statement's negation. Consider the pair: *It is raining* and *it is not raining*. If you look outside, your experience of the weather will make you update your belief in both of these simultaneously. The total hidden information should depend on both $h(s|K)$ and $h(\neg s|K)$. Our prior is that we don't know if it's raining or not. We need to think about how both of these contribute to the total. If the EA has a strong belief in s ("I'm almost certain it's raining"), then there is very little hidden information in the world concerning s. On the other hand, if the EA has a very low belief in $\neg s$, there is a lot of information hidden in the world concerning $\neg s$. These statements seem contradictory. Is there a little or a lot of hidden information in the world concerning the pair $\{s, \neg s\}$?

To settle this question, we might consider the sum of hidden informations. Writing $b = b(s|K)$ for brevity,

$$\begin{aligned} h(s|K) + h(\neg s|K) &= -\log b - \log(1-b) \\ &= -\log b(1-b). \end{aligned} \tag{8.10}$$

This function is symmetric around $b = \frac{1}{2}$ and gets larger as we move away from this value, diverging as $b \to 1$ or $b \to 0$. So, as the EA becomes more certain in either s or $\neg s$, the sum of hidden informations increases indefinitely. This is absurd! We must remedy it to properly take into account the ignorance of our EA. Apparently the total hidden information in the world is not just the mere sum of hidden informations. How then is the total measured?

Up to this point, we have been considering the pair $\{s, \neg s\}$, and want a measure of total hidden information H conditioned on knowledge K, which has the property that as the corresponding belief pair becomes more polarized, the measure should get smaller. (The more the EA knows about the weather, the more she knows whether it's raining or not. Her belief in one of the two options gets strengthened.) When the gap between the two beliefs approaches zero, it is clear that the amount of hidden information should maximize: very sensibly, the state with maximal hidden information corresponds to the maximally ignorant epistemic state where the EA has no preference whatsoever in believing in the truth or falsehood of a statement: *I have no clue whether it's raining or not.* These arguments can be generalized to many mutually exclusive statements, which we label as s_i, so that $b_i = b(s_i|K)$ for brevity.

What other properties is H endowed with? Since it depends on the epistemic state of the EA, it should depend on the beliefs of the EA. We now assume that the total hidden information can be factored into a piece that depends explicitly on prior knowledge and another piece, the entropy, that depends implicitly on prior knowledge. The simplest possible candidate for the explicitly dependent function is belief itself. We then have

$$H[\{s_i\}|K] = b(K)S[\{s_i\}|K]. \tag{8.11}$$

By considering how information is related to questioning, it can be shown (Sowinski 2016) that the implicit piece must be an expectation value over statements:

$$\begin{aligned} S[\{s_i\}|K] &= \sum_i b(s_i|K)h(s_i|K) \\ &= -\sum_i b_i \log b_i \end{aligned} \tag{8.12}$$

Formally, this is the same formula that Shannon used to define the entropy of an alphabet. The differences here are twofold. First, belief is playing the role of probability. This is not surprising since we saw earlier that beliefs are structurally the same as probabilities, allowing us to use the Bayesian interpretation of probability. Secondly, statements about the world are playing the role of the alphabet. The messages of experience are transcribed into language, which then forms the core of what belief is about. We do not believe in an experience, since an experience is something that just IS. We believe in statements about the world that those experiences inform us about.

For an epistemic agent to be truly honest with their beliefs, not allowing herself to be held back by anything other than the information contained in experience, the EA must strive to always have beliefs that maximize the hidden information. An IEA, of course, does this automatically. For humans, it is not always so easy to rid oneself of preconceived notions that have no experiential support. Interestingly, by postulating this *method of maximal entropy* (MaxENT), we connect the epistemic ideal of intellectual honesty with the Second Law of Thermodynamics. The connection between entropy in the epistemic sense developed here, and the thermodynamic sense of Gibbs, led to the resolution of an age old problem in statistical mechanics known as *Maxwell's Demon* (Parrondo et al. 2015; Sagawa 2012).

Making Sense of the World: The Relevance of Scale

Confronted with sensory data, an EA is continuously updating her beliefs about the world. The inferences made are a result of applying the Bayesian formalism on experience to generate posteriors, which are then the input for the next round of experience. The flux of belief is a signature of there being information in the experiences; the experiences are relevant to the agent and his state of awareness is updated. As experiences cause certain beliefs to polarize, the amount of hidden information decreases, corresponding to an increase in the certainty that the agent has about statements.

An EA in the world will come to believe that certain things happen more regularly than others and that certain objects will appear more frequently in their day to day than others. We are pretty sure that on our daily commute to the office we will not have the pleasure of seeing a Triceratops. We are, however, pretty certain that we will come across some cars, maybe a bus, and quite a few people. Events that occur on familiar scales tend to leave us with very polarized beliefs concerning them. They don't carry a lot of information. Events that happen on scales much smaller or larger leave us with a sense of surprise. These scales may be spatial or temporal. Indeed, routine events do little to change belief; the rarer the event the more impactful it will be. This was quantified in the previous section as we equated hidden information with Shannon's entropy.

Let's finesse this further by considering our daily interactions with objects which we perceive on length scales close to our own. Imagine entering a room for the very first time. The walls are painted in a slightly off red color, and there is a desk with a computer. There are shelves along the walls filled with books, and posters on the walls. Next to the corner closest to us we notice a small spider, dangling from a silken thread, its legs moving frantically in ascent. We turn and look more closely at the shelves, noticing books about physics and philosophy, and when we focus back to the desk we become interested in papers on it which seem to be covered with scribbles. Closer inspection reveals equations, and as we touch the papers, a pencil rolls from underneath them to the edge of the desk, and falls off the side. We pick it up, put it back on the desk, and then stand back, panning our gaze over the room as a whole. We imagine the academic that must call this room her home, and all the questions we'd like to ask her about the books and the papers. We're standing now close to the corner next to the entrance, when a memory tickles us, and we turn back to see the spider. It is no longer there. We panic just a bit, brushing ourselves in case the spider jumped on our body in the brief moments that we were taking stock of the office. Several brushes and a quick inspection leaves us confident that the spider is most likely not on us, but as we look around frantically we cannot find it. Realizing we're being foolish, we smile and look back at the shelves to see if any of the books might be of interest. We note a few and remember them. If we ever see the owner of the office we will have to ask her about those books. We turn and go from whence we came.

What can we learn from this story? There are a few things that seem obvious, but should be stated explicitly. When we first entered the office, we were attentive to the existence of the shelves and the desk, and to the colors of the walls. The largest things in the room caught our attention, and created the skeletal structure of the model in our minds of the room we were in. Once our beliefs were polarized about the general shape of the room and its coarsest features, we began to pay more attention to smaller things: The types of books and papers on the desk; the motion of the spider alerting us to its presence even though it is much smaller than the scales we are currently investigating. After taking note of it we turned to the equations which, too, were small relative to the other objects in the room. Why do we do things in this order? Why do we investigate starting at the coarsest scales first, and then move to finer scales? How can instability, as exemplified by the unexpected presence of the spider, derail us from this general mode of investigation?

We should look at the informational narrative of the above process, given our framework. We began with quite a bit of hidden information when we entered the room. Experience quickly remedied that. But why? For an answer, we turn to the concept of *coarse-graining*. A coarse-grained distribution has less hidden information in it, in general, than a fine distribution. (*Look, it's a group of 100 people.*) Pinpointing a value in such a distribution seems to not decrease the amount of hidden information by much. (*Look, it's one of these people.*) However, the act of conditioning a fine-grained distribution based on the information from a coarse-grained experience will significantly decrease the hidden information in the fine distribution. (*Look, people are wearing shirts of different colors and are grouped according to color.*) Thus, the act of minimizing hidden information by forcing experiences through actions (looking around the room, gazing at what is on the shelves and on the desk), can be accomplished more efficiently by pinning down coarse-grained features first, and using them to throw away irrelevant fine features. The philosopher Evan Thompson referred to this relation with the world as autopoiesis, seen as the *dynamic co-emergence of an interiority and exteriority* (Thompson 2007).

Of course, the spider was not a coarse-grained object in the room, so noticing it seems to be at odds with the interpretation that gaining information in the most efficient manner should progress from coarser to finer scales. What made something at a finer scale more relevant to us? Well, first off it was in motion. Since the rest of the office was in a static state, the best way to proceed in reducing hidden information would be the process of conditioning on coarser experiences. Things that move capture our attention because of the instability that change introduces to the static world. Abrupt change leads to an abrupt update in the state of awareness.

Ignoring the particulars of different objects that are the same size, what is a good measure for the amount of hidden information to an EA which has just been introduced into a room that contains objects, or distributions of objects, at many different scales? Correlations between the scales and positions of objects will need to be considered, since sizes and orientations contribute to the overall information that the EAs want to discover about the general shape of the space that they're in. Scales at which correlations are the largest will cause the EA's belief about objects at those scales to polarize the fastest. Smaller and larger scales relative to this correlation scale will contain much more hidden information. In making sense of the world, the unusual is what makes the difference.

We are thus led to consider a measure that can capture spatial correlations, such that less correlated scales imply more hidden information. This quantity is called Configurational Entropy (CE), introduced by Gleiser and Stamatopoulos in 2012, and inspired by Shannon's information entropy (Gleiser and Stamatopoulos 2012). Essentially, the CE measures the spatial correlation of objects at different length scales. It can be used very effectively to detect patterns that jump out of the background at a certain distance scale (a certain size). If all we see (the *message*) is a noisy mess, the CE will be large: there are no perceived correlations at any scale and hidden information is large. If, on the other hand, there are patterns in space at certain distances, the CE will be smaller and so will be hidden information.

A possible analogy is George Seurat's pointillist paintings. Looked at too closely, all we see are colored points with no discernible pattern. Take a step or two back, and patterns begin to emerge until a picture forms in our minds, a scene in a park with people, parasols, trees, and animals. Although this is not the proper place to go into the technical details of CE, we can state that it offers a measure of spatial complexity in the physical world based on the concept of hidden information and, thus, on the relation between an EA and the world she perceives through experience. In the next section, we sketch the foundational aspects of this relation, which we will present in more detail in a future publication. For phenomenological applications of CE, look at Refs. (Gleiser and Sowinski 2013, 2015; Gleiser and Jiang 2015; Sowinski and Gleiser 2017).

Psychophysical Foundations of Configurational Entropy

For most of human history, there were 1025 stars in the Heavens (Pratt 2015). These pinpricks of light were all that the human eye could see and write about, and these thousand stars, together with the wanderers (*planetos*) and the occasional shooting star, were the sum total of human knowledge about the census of the sky. The stars were subdivided into six classes based on how bright they were, with $m = 1$ being the brightest, and $m = 6$ the faintest. It wasn't until the invention of the telescope in 1608 that this catalogue began to increase in size, with dimmer and dimmer stars being discovered. Initially, the magnitude scale appeared to be increase linearly with brightness based on the human eye's ability to register light. However, when more sophisticated methods were brought to bare on the field of photometry, it turned out that magnitudes were *logarithmically* related to the amount of light being received: stars differing by one magnitude point were twice as bright, those differing by two magnitudes four times, three magnitudes eight times, and so on.

Logarithms are very important in the realm of perception. For two stimuli to register as being different, the senses must perceive them with a *just noticeable difference* (JND), a threshold for perception. Stimuli that do not create a JND to the senses are perceived as being the same. Consider, if you will, two rulers placed before you.

One of them is a decimeter in length, while the other a meter. Imagine that both are instantaneously increased in length by one centimeter. Perceiving a change in the smaller ruler will be obvious. However, an EA may or not perceive the change in the larger ruler. Had one of the rulers been a kilometer in length, it is certain hat a change by one centimeter would not register in the EA's senses.

In 1860, Ernst Weber proposed a quantitative relation for the change in perception, $\Delta\mathcal{P}$, equating it to both a change in stimulus, ΔS, and the stimulus, S (Fechner 1860):

$$\Delta\mathcal{P} \propto \frac{\Delta S}{S}. \tag{8.13}$$

The solution (the change in perception) is a logarithm. In that same year, Gustav Fechner tested the relation experimentally. Since then, it's been known as the Weber-Fechner relation (WFR). Equating the experience e, with the change in perception $\Delta\mathcal{P}$, we can write

$$e(S) \propto \log\left(\frac{S}{S_0}\right), \tag{8.14}$$

where S_0 is a baseline stimulus used to calibrate the relationship.

The WFR begins to break down at the boundary of sense perception, though it holds for each of the senses. In particular, we can use it in the context of scale perception as a way to constrain the belief distribution that an EA has about its environment.[2]

In the case of spatial perception, we propose that the stimulus is related to the *two-point correlation function*, a mathematical quantity that describes how pairs of points are correlated in a spatial environment, peaking at scales typical of most objects or physical properties in that environment. For an example of a physical property, consider the temperature at different points in a room. If there is little change in temperature from point to point, the two-point correlation will be large, given that most points have similar temperatures and are thus highly-correlated. If, instead, the temperature fluctuates randomly from point to point, the two-point correlation will be very small. More precisely, the two-point correlation function gives the relative power at different scales, which, in our formulation, is related to the strength of the stimulus at different scales (Fig. 8.1). (In the example of the temperature in a room, the scales will be related to the different sizes of the volumes in the room that have the same temperature. The biggest volumes with the same temperature will have the most power in the two-point correlation function.)

Constraining the hidden information for an IEA by the mean experience of spatial scales using the WFR, one then finds that an IEA should have a belief distribution over spatial scales that is a power law of the power spectrum, coined the modal fraction in Ref. (Gleiser and Stamatopoulos 2012). As shown in that work, the modal fraction is the key ingredient of the Configuration Entropy, introduced in section

[2] With this relation, we can write that the experience for an EA with prior knowledge K and belief $b(k|K)$ of a given scale (size) k is $e(k) = b(k|K)e(S(k))$, so that the average experience is the sum over all scales k, $\langle e\rangle = \sum_k b(k|K)e(S(k))$.

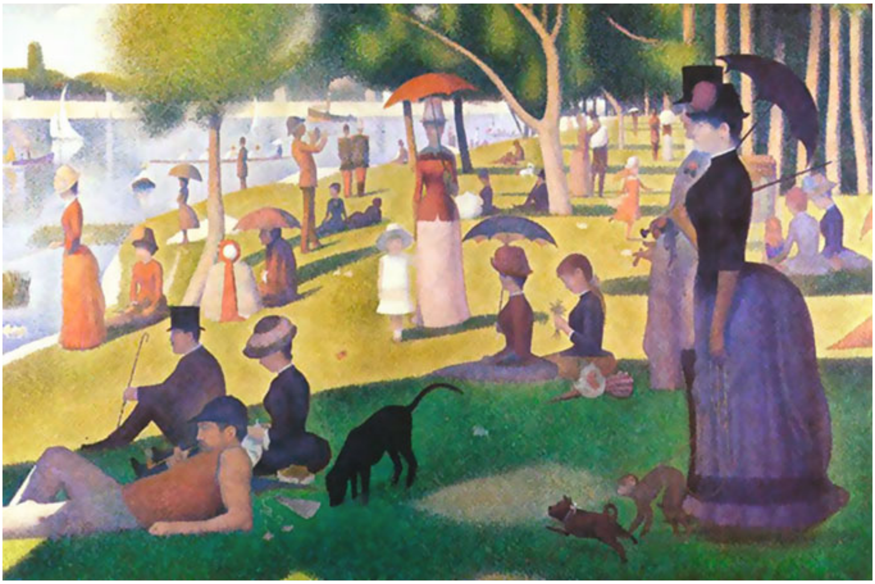


Fig. 8.1 Georges Seurat's classic pointillist painting, *A Sunday Afternoon on the Island of La Grand Jatte*. Helen Birch Bartlett Memorial Collection, The Art Institute of Chicago

"Making Sense of the World: The Relevance of Scale". Essentially, the modal fraction $f(\mathbf{k})$ gives the relative probability of a given spatial scale over all others. If a spatial scale is prominent, it will dominate the modal fraction, which can have a value between 0 and 1.[3] Using this belief distribution, the Configurational Entropy (CE) becomes a quantitative measure of how much information in spatial-complexity the external world hides from an IEA: a world of sameness hides little, while a world rich in spatial patterns at different scales hides a lot. In this way, the CE offers a quantitative measure for an IEA's different experiences. These experiences are then used by an IEA to construct a map of its perceived physical reality.

Concluding Remarks

As sentient beings, humans are forever locked within their limited perception of physical reality. One may conjecture of an "ultimate reality," what we could call the perfectly complete ontological Territory, but such entity is certainly out of reach,

[3] Mathematically, the modal fraction is proportional to the *power spectrum*, $f(\mathbf{k}) \propto \mathcal{P}(k)$ and so is directly related to the two-point correlation function. (For the mathematically savvy, the power spectrum is the Fourier Transform of the two-point correlation function.) The Configurational Entropy is the hidden information of the MaxEnt distribution under the constraint of average experience k, $S_{CE} = -\sum_k f(\mathbf{k}) \log[f(\mathbf{k})]$.

even for Idealized Epistemic Agents. At best, we can collect partial information about the Territory as we confront the world with our prior knowledge, working to decrease, through experience and its decoding, the amount of hidden information. The Maps we construct are the products of such a process, always works in progress. Science is one of such Maps, but certainly not the only one. As we discussed, what we sense of reality is, in the parlance of information theory, the message. As we experience the world through different stimuli and consequently update our state of awareness, we decrease the amount of hidden information. In our framework, information is that which changes belief.

When applied to how an EA senses and moves in the world, the information updating depends on spatial perception. We have presented a formalism to describe how this process works by matching a quantity called Configurational Entropy to the hidden information. One can think of the Configurational Entropy as a measure of the spatial complexity of an objects and how it relates to other objects nearby. As an EA enters a new environment and begins to sense it, she searches for spatial correlations among objects. Stimuli that promote the most change in the EA's state of awareness are the ones that carry the most information. These tend to be the stimuli that depart most strongly from the average experience. According to the Weber-Fechner relation, such change grows with the logarithm of the intensity of the stimulus. Using this expression, we were able to show that, under certain assumptions, the hidden information is given by the Configurational Entropy, computed from the spatial correlation between different objects in the room. The more varied the room, the greater the stimulus, the richer the experience, and the more hidden information. As the EA keeps on exploring and updating her belief, different experiences will decrease the amount of hidden information.

Acknowledgements This work was supported by the Institute for Cross-Disciplinary Engagement at Dartmouth College (http://ice.dartmouth.edu) by a John Templeton Foundation grant.

Appendix A: Outline of Derivation of Configurational Entropy

Consider a stimulus S. This can be anything from sound to light to touch. Denote an EA's experience of the perception of that stimulus as $e(S)$. Now consider changing the stimulus by some small amount, $S \to S + \delta S$. The empirical Weber-Fechner relation states that a change in perception is proportional to a change in the stimulus, but inversely proportional to the stimulus itself:

$$\delta e = \eta \frac{\delta S}{S}, \tag{A1}$$

where η is some empirically determined constant of proportionality. This implies that the experience is proportional to the logarithm of some power of the stimulus,

$$e(S) = e_0 + \log S^\eta. \tag{A2}$$

Here e_0 is a constant of integration. It is reasonable that when the stimulus reaches some minimum value, S_0, then one can no longer perceive it, thus

$$e(S) = \log \left(\frac{S}{S_0} \right)^\eta. \tag{A3}$$

For the experience of spatial scale, the stimulus is proportional to the power spectrum

$$S(k) \propto \mathcal{P}(k), \tag{A4}$$

where $k = \frac{\pi}{L}$ is a wave-mode at some inverse length scale, L. Note that the smallest perceptible scale will correspond to some maximum wave-mode, k_*, so that $S_0 \propto \mathcal{P}(k_*)$. An EA will have some belief distribution over the importance of scales, $b(k|K)$, based on its background knowledge K. The EA's mean experience of scales will then just be:

$$\begin{aligned} \langle e \rangle &= \sum_k b(k|K) e(S(k)) \\ &= \eta \sum_k b(k|K) \log \frac{\mathcal{P}(k)}{\mathcal{P}(k_*)}. \end{aligned} \tag{A5}$$

What distribution should the EA have over scales that is as unbiased as possible considering nothing but mean experience? To find an answer, we apply the MaxEnt method (Parrondo et al. 2015).

The total hidden information in spatial scales for the EA can be expressed by the functional:

$$\begin{aligned} H[\{b\}; \alpha, \beta, \gamma] = &- \sum_k b(k) \log b(k) + \alpha \left(1 - \sum_k b(k) \right) \\ &+ \beta \left(\langle e \rangle - \eta \sum_k b(k) \log \frac{\mathcal{P}(k)}{\mathcal{P}(k_*)} \right) + \gamma(\cdots), \end{aligned} \tag{A6}$$

where α, β, and γ are Lagrange multipliers that enforce constraints. The first term is the entropy of the distribution. It tries to drive the belief distribution towards uniformity (ignorance). The second term is the constraint enforcing that the belief distribution be normalized. The third term is the constraint enforcing mean experience. The last term is all the constraints that impose the EA's background knowledge, K. Since we are only worried in how mean experience of scale constrains belief, we can set γ to zero. Varying with respect to the Lagrange multipliers simply reproduces the constraints:

$$\frac{\delta H}{\delta \alpha} = 0 \Rightarrow \sum_k b(k) = 1; \tag{A7}$$

$$\frac{\delta H}{\delta \beta} = 0 \Rightarrow \eta \sum_k b(k) \log \frac{\mathcal{P}(k)}{\mathcal{P}(k_*)} = \langle e \rangle. \tag{A8}$$

Varying with respect to the belief distribution, we find

$$\frac{\delta H}{\delta b(k)} = -\log b(k) - 1 - \alpha - \beta\eta \log \frac{\mathcal{P}(k)}{\mathcal{P}(k_*)}.$$

Setting this variation to 0,

$$b(k) = e^{-(1+\alpha)} \left(\frac{\mathcal{P}(k)}{\mathcal{P}(k_*)} \right)^{-\beta\eta}. \tag{A9}$$

Imposing the normalization constraint we obtain,

$$b(k) = \frac{\mathcal{P}(k)^{-\beta\eta}}{\sum_{k'} \mathcal{P}(k')^{-\beta\eta}}. \tag{A10}$$

Choosing $\beta\eta = -1$ (Sowinski and Gleiser 2016), we obtain a special case for which the belief distribution is the modal fraction, $f(k)$:

$$b(k) = \frac{\mathcal{P}(k)}{\sum_k' \mathcal{P}(k')} = f(k). \tag{A11}$$

Plugging this, and the constraints, into the hidden information functional gives us the expression for the Configurational Entropy in terms of the modal fraction from Ref. (Gleiser and Stamatopoulos 2012):

$$S_C = H[\{b\}; \cdots] = -\sum_k f(k) \log f(k). \tag{A12}$$

This results establishes a quantitative link between the psychophysical foundations of spatial perception and the configurational entropy as a measure of hidden information for an EA.

References

J. Aczél, *Lectures on Functional Equations and Their Applications*, Mathematics in science and engineering (Academic Press, 1966)

A. Caticha, ArXiv e-prints **0808**, 0012 (2008)

A. Caticha, in *AIP Conference Proceedings* vol. 1193, (2009), p. 60. arXiv:0908.3212 [physics.data-an]

R.T. Cox, Am. J. Phys. **14**, 1 (1946)

R.T. Cox, *The Algebra of Probable Inference* (Johns Hopkins University Press, 1961)

J.M. Dubbey, *The Mathematical Work of Charles Babbage* (Cambridge University Press, 1978)

H. Enderton, H.B. Enderton, *A Mathematical Introduction to Logic* (Elsevier Science, 2001)

G. Fechner, *Elemente der Psychophysik*, Elemente der Psychophysik No. v. 1 (Breitkopf und Härtel, 1860)

M. Gleiser, N. Jiang, Phys. Rev. D **92**, 044046 (2015)

M. Gleiser, D. Sowinski, Phys. Lett. B **727**, 272 (2013)

M. Gleiser, D. Sowinski, Phys. Lett. B **747**, 125 (2015)

M. Gleiser, N. Stamatopoulos, Phys. Lett. B **713**, 304 (2012)

E.T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003)

J.M.R. Parrondo, J.M. Horowitz, T. Sagawa, Nat. Phys. **11**, 131 (2015)

J.P. Pratt, The Ptolemy star catalogue, www.JohnPratt.com

T. Sagawa, Prog. Theor. Phys. **127**, 1 (2012)

C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948)

D. Sowinski, Blog: atoms and void, entropic compression and english (2016), https://atomsandvoid.wordpress.com/2016/04/08/entropic-compression-and-english

D. Sowinski, *Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy* (Dartmouth College, Department of Physics and Astronomy, 2016)

D. Sowinski, M. Gleiser, *The Psychophysical Foundations of Configurational Entropy* (in progress)

D. Sowinski, M. Gleiser, J. Stat. Phys. **167**, 1221 (2017)

E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind* (Harvard University Press, 2007), p. 79

Chapter 9
Theories of Knowledge and Theories of Everything

David H. Wolpert

Introduction

Suppose we are given a “theory of everything”, and want to perform an experimental test of one of its claims. For example, that test could involve our turning a knob on an apparatus to a particular setting, observing the resultant value of a sensor, etc. Tautologically, if we do not *know* what setting we turned the knob to, or do not *know* the sensor value that was observed, then we have not run the experimental test of the theory. Evidently then, the phenomenon of our coming to know the values of some physical variables (in our example, the variables are the setting of the knob and the resultant sensor value) is central to any test we could do of a theory of everything. In other words, it is central to any physical significance we could ascribe to the theory. As a result, if a theory of everything is to truly be a theory of *everything*, it needs to include a formal description of what it means for us to know the value of a physical variable—even a variable as mundane as the setting of a knob on an experimental apparatus.

The point is not that any theory of everything must cover the processes governing any experimental apparatus used to test the theory (though that is certainly true). Rather the point is that the theory must also include a formal description of what it means for us to “know” the values of some of the physical variables specifying the state of that apparatus. Unfortunately, reflecting a long-standing aversion in physics to theories with any subjective aspects, typical theories of everything have nothing to say about what it means to know the value of a physical variable. Accordingly, they are incomplete, stopping just shy of covering all that they need to cover to give a complete description of the universe.

In this paper I review a minimal formalization of what it means to know the value of one or more physical variables. This formalization provides one possible way of expanding theories of everything so that they are truly complete. However as I show in this paper, this formalization has physical consequences, imposing a priori restrictions on the possible properties of any theory of everything.

D. H. Wolpert (✉)
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
e-mail: david.h.wolpert@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_9

I begin by noting that there are four common ways that an agent can come to have knowledge concerning a physical variable: by observation, by memory, by control, or by (correct) prediction. As I describe below, these four ways of acquiring knowledge share a common mathematical structure. This structure possesses a set of properties that restrict the possible form of any theory of everything. For example, this structure means that no universe can contain two independent "agents" (e.g., two observers) that both know all physical facts in that universe. This is true independent of the precise physical laws that govern that universe. Tongue firmly in cheek, one might describe this is a "monotheism theorem", since in a formal sense, it means that although there might be one agent in a universe who knows everything about that universe (depending on the details of the laws of that universe) there cannot be two.

The same analysis that leads to the monotheism theorem also proves that Laplace's famous demon is logically flawed. It does not matter whether the universe is finite, classical and non-chaotic, or whether Laplace's demon has super-Tuning computing abilities. There are always physical variables whose value the demon cannot predict ahead of time.

In addition, it may be that a particular universe contains an agent who, at one particular moment, knows all facts that will every hold in that universe. However it turns out that no universe can contain an agent who *at more than one moment* knows all facts that will ever hold in that universe. Again tongue firmly in cheek, this could be described as an "intelligent design theorem".

These results all concern subsets of an entire universe, e.g., one or two IDs embedded in a larger universe. However in keeping with the theme of this book—"The map and the territory"—in the last section I expand the scope to view an entire universe through the lens of ID theory. The idea is to *define* a "universe", with whatever associated laws of physics, to be a set of physical systems and IDs (e.g., a set of scientists), where the IDs can have knowledge concerning those physical systems and/or one another. Adopting this approach, in the last section I use the theory of IDs to derive impossibility results concerning the nature of the entire universe.

Although not presented here, it is worth noting that other recent work has substantially extended the theory of IDs (Wolpert 2017). In particular, we now have an understanding of some of the more elementary connections between the theory of IDs and the theory of Turing Machines (Chaitin 2004; Lloyd 1990, 2006; Schmidhuber 2000; Zenil 2012; Zurek 1989a, b, 1990; Zuse 1969). To date, this work has concentrated on analyzing the properties of an ID version of universal Turing machines and of an ID version of Kolmogorov complexity. In particular it is shown in (Wolpert 2017) that the ID versions of those two quantities obey many of the familiar results of Turing machine theory (e.g., the invariance theorem of TM theory)—but not all of them.

In (Wolpert 2017) it is also shown how to extend the theory of IDs to the case where there is a probability distribution over the states of the universe, so that no knowledge is ever 100% guaranteed to be true. In particular, that paper derives a result concerning the products of probabilities of error of two separate IDs, a result which is formally similar to the Heisenberg uncertainty principle.

Finally, (Wolpert 2017) also introduces a strengthened form of IDs, which provides a full epistemic logic. That paper contains a preliminary analysis of the relationship between these strengthened IDs and conventional epistemic logics, in particular Kripke structures and Aumann structures (Scott Aaronson 2013; Aumann 1999; Aumann and Brandenburger 1995; Robert 1976; Binmore and Brandenburger 1988; Fagin 2004; Fudenberg and Tirole 1991; Parikh 1987; Zalta et al. 2003). In particular, we say that the property of *logical omniscience* holds in a give epistemic logic if under that logic, any agent who knows a set of propositions A must know all of the logical implications of A. As an example, under the property of logical omniscience, if someone knows the axioms underlying number theory, then they must know all theorems of number theory—and so know the answers to many of the major open problems in contemporary mathematics. The conventional epistemic logics like Kripke structures and Aumann structures obey logical omniscience. However logical omniscience need not hold for the strengthened version of IDs. So IDs avoid what is perhaps the major problem plaguing conventional work in epistemic logic.

There are no proofs in this paper. All proofs of the results that not explicitly referenced can be found in (Wolpert 2008).

Inference Devices

In this section I review the elementary properties of inference devices, mathematical structures that are shared by the processes of observation, prediction, recall and control (Binder 2008; Wolpert 2001, 2008, 2010). These results are proven by extending Epimenides' paradox to apply to novel scenarios. Results relying on more sophisticated mathematics, some of them new, are presented in (Wolpert 2017).

Observation, Prediction, Recall and Control of the Physical World

I begin with two examples that motivate the formal definition of inference devices. The first is an example of an agent making a correct observation about the current state of some physical variable.

Example 1 Consider an agent who claims to be able to observe $S(t_2)$, the value of some physical variable at time t_2. If the agent's claim is correct, then for any question of the form "Does $S(t_2) = L$?", the agent is able to consider that question at some $t_1 < t_2$, observe $S(t_2)$, and then at some $t_3 > t_2$ provide the answer "yes" if

$S(t_2) = L$, and the answer "no" otherwise. In other words, she can correctly pose any such binary question to himself at t_1, and correctly say what the answer is at t_3.[1]

To formalize this, let U refer to a set of possible histories of an entire universe across all time, where each $u \in U$ has the following properties:

(i) u is consistent with the laws of physics,
(ii) In u, the agent is alive and of sound mind throughout the time interval $[t_1, t_3]$, and the system S exists at the time t_2,
(iii) In u, at time t_1 the agent considers some L-indexed question q of the form "Does $S(t_2) = L$?",
(iv) In u, the agent observes $S(t_2)$,
(v) In u, at time t_3 the agent uses that observation to provide her (binary) answer to q, and believes that answer to be correct.[2]

The agent's claim is that for any question q of the form "Does $S(t_2) = L$?", the laws of physics imply that for all u in the subset of U where at t_1 the agent considers q, it must be that the agent provides the correct answer to q at t_3. Any prior knowledge concerning the history that the agent relies on to make this claim is embodied in the set U.

The value $S(t_2)$ is a function of the actual history of the entire universe, $u \in U$. Write that function as $\Gamma(u)$, with image $\Gamma(U)$. Similarly, the question the agent has in her brain at t_1, together with the time t_1 state of any observation apparatus she will use, is a function of u. Write that function as $X(u)$. Finally, the binary answer the agent provides at t_3 is a function of the state of her brain at t_3, and therefore it too is a function of u. Write that binary-valued function as $Y(u)$. Note that since U embodies the laws of physics, in particular it embodies all neurological processes in the agent (e.g., her asking and answering questions), all physical characteristics of S, etc.

So as far as this observation is concerned, the agent is just a pair of functions (X, Y), both with the domain U defined above, where Y has the range $\{-1, 1\}$. A necessary condition for us to say that the agent can "observe $S(t_2)$" is that for any $\gamma \in \Gamma(U)$, there is some associated X value x such that for all $u \in U$, so long as $X(u) = x$, it follows that $Y(u) = 1$ iff $\Gamma(u) = \gamma$.

I now present an example of an agent making a correct prediction about the future state of some physical variable.

Example 2 Now consider an agent who claims to be able to predict $S(t_3)$, the value of some physical variable at time t_3. If the agent's claim is correct, then for any

[1] It may be that the agent has to use some appropriate observation apparatus to do this; in that case we can just expand the definition of the "agent" to include that apparatus. Similarly, it may be that the agent has to configure that apparatus appropriately at t_1. In this case, just expand our definition of the agent's "considering the appropriate question" to mean configuring the apparatus appropriately, in addition to the cognitive event of her considering that question.

[2] This means in particular that the agent does not lie, does not believe she was distracted from the question during $[t_1, t_3]$.

question of the form "Does $S(t_3) = L$?", the agent is able to consider that question at some time $t_1 < t_3$, and produce an answer at some time $t_2 \in (t_1, t_3)$, where the answer is "yes" if $S(t_3) = L$ and "no" otherwise. So loosely speaking, if the agent's claim is correct, then for any L, by their considering the appropriate question at t_1, they can generate the correct answer to any question of the form "Does $S(t_3) = L$?" at $t_2 < t_3$.[3]

To formalize this, let U refer to a set of possible histories of an entire universe across all time, where each $u \in U$ has the following properties:

(i) u is consistent with the laws of physics,
(ii) In u, the agent exists throughout the interval $[t_1, t_2]$, and the system S exists at t_3,
(iii) In u, at t_1 the agent considers some question q of the form "Does $S(t_3) = L$?",
(iv) In u, at t_2 the agent provides his (binary) answer to q and believes that answer to be correct.[4]

The agent's claim is that for any question q of the form "Does $S(t_3) = L$?", the laws of physics imply that for all u in the restricted set U such that at t_1 the agent considers q, it must be that the agent provides the correct answer to q at t_2.

The value $S(t_3)$ is a function of the actual history of the entire universe, $u \in U$. Write that function as $\Gamma(u)$, with image $\Gamma(U)$. Similarly, the question the agent considers at t_1 is a function of the state of his brain at t_1, and therefore is also a function of u. Write that function as $X(u)$. Finally, the binary answer the agent provides at t_2 is a function of the state of his brain at t_2, and therefore it too is a function of u. Write that function as $Y(u)$.

So as far as this prediction is concerned, the agent is just a pair of functions (X, Y), both with the domain U defined above, where Y has the range $\{-1, 1\}$. The agent can indeed predict $S(t_3)$ if for the space defined above U, for any $\gamma \in \Gamma(U)$, there is some associated X value x such that, no matter what precise history $u \in U$ we are in, due to the laws of physics, if $X(u) = x$ then the associated $Y(u)$ equals 1 iff $\Gamma(u) = \gamma$.

Evidently there is a mathematical structure, in the form of functions X and Y, that is shared by agents who do observation and those who do prediction. As formalized below, I refer to any such pair (X, Y) as an "inference device". Say that for some function Γ, for any $\gamma \in \Gamma(U)$, there is some associated X value x such that, no matter what precise history $u \in U$ we are in, due to the laws of physics, if $X(u) = x$ then the associated $Y(u)$ equals 1 iff $\Gamma(u) = \gamma$. Then I will say that the device (X, Y) "infers" Γ.

[3] It may be that the agent has to use some appropriate prediction computer to do this; in that case we can just expand the definition of the "agent" to include that computer. Similarly, it may be that the agent has to program that computer appropriately at t_1. In this case, just expand our definition of the agent's "considering the appropriate question" to mean programming the computer appropriately, in addition to the cognitive event of his considering that question.

[4] This means in particular that the agent does not believe he was distracted from the question during $[t_1, t_2]$.

See (Wolpert 2008) for a more detailed elaboration of the processes of observation and prediction in terms of inference devices. Arguably to fully formalize each of these phenomena there should be additional structure beyond that defining inference devices (See App. B. of (Wolpert 2008)). Some such additional structure is investigated below, in the discussion of "physical knowledge".

It is also shown in (Wolpert 2008) that a system that remembers the past is an inference device. (Intuitively, memory is just retrodiction, i.e., it is using current data to predict the state of non-current data, but rather than have the non-current data concern the future, in memory it concerns the past.) Wolpert (2008) also shows that a device that controls a physical variable is an inference device. All of this analysis holds even if what is observed/predicted/remembered/controlled is not the answer to a binary question of the form, "Does $S(t) = L$?", but instead an answer to question of the form, "is $S(t)$ more property A than it is property B?" or of the form, "is $S(t)$ more property A than $S'(t)$ is?"

In the sequel I will sometimes consider situations involving multiple inference devices, (X_1, Y_1), (X_2, Y_2), . . ., with associated domains $U_1, U_2, \ldots$. For example, I will consider scenarios where agents try to observe one another. In such situations, when referring to "U", I implicitly mean $\cap_i U_i$, implicitly restrict the domain of all X_i, Y_i to U, and implicitly assume that the codomain of each such restricted Y_i is binary.

Notation and Terminology

To formalize the preceding considerations, I first fix some notation. I will take the set of binary numbers $\mathbb{B}$ to equal $\{-1, 1\}$. For any function Γ with domain U, I will write the image of U under Γ as $\Gamma(U)$. I will also sometimes abuse this notation with a sort of "set-valued function" shorthand, and so for example write $\Gamma(V) = 1$ for some $V \subset U$ iff $\Gamma(u) = 1\ \forall u \in V$. On the other hand, for the special case where the function over U is a measure, I use conventional shorthand from measure theory. For example, if P is a probability distribution over U and $V \subset U$, I write $P(V)$ as shorthand for $\sum_{u \in V} P(u)$.

For any function Γ with domain U that I will consider, I implicitly assume that the entire set $\Gamma(U)$ contains at least two distinct elements. For any (potentially infinite) set R, $|R|$ is the cardinality of R.

Given a function Γ with domain U, I write the partition of U given by Γ^{-1} as $\overline{\Gamma}$, i.e.,

$$\overline{\Gamma} \equiv \{\{u : \Gamma(u) = \gamma\} : \gamma \in \Gamma(U)\} \tag{1}$$

I say that two functions Γ_1 and Γ_2 with the same domain U are **(functionally) equivalent** iff the inverse functions Γ_1^{-1} and Γ_2^{-1} induce the same partitions of U, i.e., iff $\overline{\Gamma_1} = \overline{\Gamma_2}$.

Recall that a partition A over a space U is a *refinement* of a partition B over U iff every $a \in A$ is a subset of some $b \in B$. If A is a refinement of B, then for every $b \in B$ there is an $a \in A$ that is a subset of b. Two partitions A and B are refinements of each other iff $A = B$. Say a partition A is finite and a refinement of a partition B. Then $|A| = |B|$ iff $A = B$. For any two functions A and B with domain U, I will say that "A refines B" if $\overline{A}$ is a refinement of $\overline{B}$. Similarly, for any $R \subset U$ and function A, I will say that "R refines A" (or "A is refined by R") if R is a subset of some element of $\overline{A}$.

I write the characteristic function of any set $R \subseteq U$ as

$$\mathcal{X}_R(u) = 1 \Leftrightarrow u \in R \tag{2}$$

As shorthand I will sometimes treat functions as equivalent to one of the values in their image. So for example expressions like "$\Gamma_1 = \Gamma_2 \Rightarrow \Gamma_3 = 1$" means "$\forall u \in U$ such that $\Gamma_1(u) = \Gamma_2(u)$, $\Gamma_3(u) = 1$".

I define a **probe** of any variable V to be a function parametrized by a $v \in V$ of the form

$$\delta_v(v') = \begin{cases} 1 & \text{if } v = v' \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

$\forall v' \in V$. Given a function Γ with domain U I sometimes write $\delta_\gamma(\Gamma)$ as shorthand for the function $u \in U \to \delta_\gamma(\Gamma(u))$. When I don't want to specify the subscript γ of a probe, I sometimes generically write δ. I write $\mathcal{P}(\Gamma)$ to indicate the set of all probes over $\Gamma(U)$.

Weak Inference

I now review some results that place severe restrictions on what a physical agent can predict and be guaranteed to be correct. To begin, I formalize the concept of an "inference device" introduced in the previous subsection.

Definition 1 An **(inference) device** over a set U is a pair of functions (X, Y), both with domain U. Y is called the **conclusion** function of the device, and is surjective onto $\mathbb{B}$. X is called the **setup** function of the device.

Given some function Γ with domain U and some $\gamma \in \Gamma(U)$, we are interested in setting up a device so that it is assured of correctly answering whether $\Gamma(u) = \gamma$ for the actual universe u. Motivated by the examples above, I will formalize this with the condition that $Y(u) = 1$ iff $\Gamma(u) = \gamma$ for all u that are consistent with some associated setup value x of the device, i.e., such that $X(u) = x$ for some x. If this condition holds, then setting up the device to have setup value x guarantees that the device will make the correct conclusion concerning whether $\Gamma(u) = \gamma$. (Hence the terms "setup function" and "conclusion function" in Definition 1).

We can now formalize inference:

Definition 2 Let Γ be a function over U such that $|\Gamma(U)| \geq 2$. A device $\mathcal{D}$ **(weakly) infers** Γ iff $\forall \gamma \in \Gamma(U)$, $\exists x \in X(U)$ such that $\forall u \in U, X(u) = x \Rightarrow Y(u) = \delta_\gamma(\Gamma(u))$.

If $\mathcal{D}$ infers Γ, I write $\mathcal{D} > \Gamma$. I say that a device $\mathcal{D}$ infers a set of functions if it infers every function in that set.

A stripped-down example of weak inference is given in the following table, which provides functions $X(u)$, $Y(u)$ and $\Gamma(u)$ for all u in a space U that has only three elements:

u	$X(u)$	$Y(u)$	$\Gamma(u)$
a	1	1	1
b	2	-1	1
c	1	-1	2

In this example, $\Gamma(U) = \{1, 2\}$, so we are concerned with two probes, δ_1 and δ_2. Setting $X(u) = 2$ means that $u = b$, which in turn means that $\Gamma(u) = 1$ and $Y(u) = -1$. So setting $X(u) = 2$ guarantees that $\delta_2(\Gamma(u)) = Y(u)$ (which in this case equals -1, the answer 'no'). So the setup value $x = 2$ ensures that the ID correctly answers the binary question, "does $\Gamma(u) = 2$"? Similarly, setting $X(u) = 1$ guarantees that $\delta_1(\Gamma(u)) = Y(u)$, so that it ensures that the ID correctly answers the binary question, "does $\Gamma(u) = 1$"?

This example shows that weak inference can hold even if $X(u) = x$ doesn't fix a unique value for $Y(u)$. Such non-uniqueness is typical when the device is being used for observation. Setting up a device to observe a variable outside of that device restricts the set of possible universes; only those u are allowed that are consistent with the observation device being set up that way to make the desired observation. But typically just setting up an observation device to observe what value a variable has doesn't uniquely fix the value of that variable.

As discussed in App. B of (Wolpert 2008), the definition of weak inference is very unrestrictive. For example, a device $\mathcal{D}$ is 'given credit' for correctly answering probe $\delta(\Gamma(u))$ if there is *any* $x \in X(U)$ such that $X(u) = x \Rightarrow Y(u) = \delta(\Gamma(u))$. In particular, $\mathcal{D}$ is given credit even if the binary question we would intuitively associate with x is not whether $\Gamma(u) = \gamma$, but some other question. In essence, the device receives credit even if it gets the right answer by accident.

Unless specified otherwise, a device written as "$\mathcal{D}_i$" for any integer i is implicitly presumed to have domain U, with setup function X_i and conclusion function Y_i (and similarly for no subscript). Similarly, unless specified otherwise, expressions like "$\min_{x_i}$" mean $\min_{x_i \in X_i(U)}$. I also say that a device $\mathcal{D}_1$ infers a device $\mathcal{D}_2$ iff $\mathcal{D}_1 > Y_2$, i.e., $\mathcal{D}_1$ infers $\mathcal{D}_2$ if it can infer what $\mathcal{D}_2$ will conclude. In general inference among devices is non-transitive (see (Wolpert 2008) for an example).

The Two Laplace's Demon Theorems

> An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough ... nothing would be uncertain and the future just like the past would be present before its eyes.
>
> —Pierre Simon Laplace, "A Philosophical Essay on Probabilities"

There are limitations on the ability of any device to weakly infer functions. Perhaps the most trivial is the following:

Proposition 1 *For any device $\mathcal{D}$, there is a function that $\mathcal{D}$ does not infer.*

Proof Choose Γ to be the function Y, so that the device is trying to infer itself. Then choose the negation probe $\delta(y \in \mathbb{B}) = -y$ to see that such inference is impossible. (Also see (Wolpert 2008)). □

It is interesting to consider the implications of Proposition 1 for the case where the inference is prediction, as in Example 2. Depending on how precisely one interprets Laplace, Proposition 1 means that he was wrong in his claim about the ability of an "intellect" to make accurate predictions: even if the universe were a giant clock, it could not contain an intellect that could reliably predict the universe's future state before it occurred.[5] More precisely, for all Γ as in Proposition 1, there could be an intellect $\mathcal{D}$ that can infer Γ. However Proposition 1 tells us that for any fixed intellect, there must exist a Γ that the intellect cannot infer. (See Fig. 9.1.) The "intellect" Laplace refers to is commonly called Laplace's "demon", so I sometimes refer to Proposition 1 as the "first (Laplace's) demon theorem".

One might think that Laplace could circumvent the first demon theorem by simply constructing a second demon, specifically designed to infer the Γ that thwarts his first demon. Continuing in this way, one might think that Laplace could construct a set of demons that, among them, could infer any function Γ. Then he could construct an "overseer demon" that would choose among those demons, based on the function Γ

[5] Similar conclusions have been reached previously (MacKay 1960; Popper 1988). However in addition to being limited to the inference process of prediction, that earlier work is quite informal. It is no surprise than that some claims in that earlier work are refuted by well-established results in engineering. For example, the claim in (MacKay 1960) that "a prediction concerning the narrator's future ... cannot ... account for the effect of the narrator's learning that prediction" is just not true; it is refuted by adaptive control theory in general and by Bellman's equations in particular. Similarly, it is straightforward to see that statements (A3), (A4), and the notion of "structurally identical predictors" in (Popper 1988) have no formal meaning.

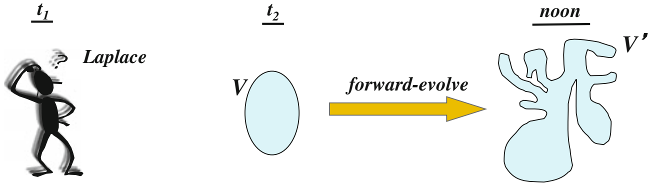


Fig. 9.1 The time t_1 is less than t_2, which in turn is less than noon. V is the set of all time-t_2 universes where Laplace is thinking the answer "yes" in response to the t_1 question Laplace heard—whatever that question was. V' is V evolved forward to noon. At t_1, we ask Laplace, "will universe be outside V' at noon?" It is impossible for Laplace to answer correctly, no matter what his computational capabilities are, what the laws of the universe are, etc.

that needs to be inferred. However this is not possible. To see this, simply redefine the device $\mathcal{D}$ in Proposition 1 to be the combination of Laplace with all of his demons.

These limitations on prediction hold even if the number of possible states of the universe is countable (or even finite), or if the inference device has super-Turning capabilities. It holds even if the current formulation of physics is wrong; it does not rely on chaotic dynamics, physical limitations like the speed of light, or quantum mechanical limitations.

Note as well that in Example 2's model of a prediction system the actual values of the times of the various events are not specified. So in particular the impossibility result of Proposition 1 still applies to that example even if $t_3 < t_2$—in which case the time when the agent provides the prediction is *after* the event they are predicting. Moreover, consider the variant of Example 2 where the agent programs a computer to do the prediction, as discussed in Footnote 3 in that example. In this variant, the program that is input to the prediction computer could even contain the future value that the agent wants to predict. Proposition 1 would still mean that the conclusion that the agent using the computer comes to after reading the computer's output cannot be guaranteed to be correct.

Proposition 1 tells us that any inference device $\mathcal{D}$ can be "thwarted" by an associated function. However it does not forbid the possibility of some second device that can infer that function that thwarts $\mathcal{D}$. To analyze issues of this sort, and more generally to analyze the inference relationships within sets of multiple functions and multiple devices, we start with the following definition:

Definition 3 Two devices (X_1, Y_1) and (X_2, Y_2) are **(setup) distinguishable** iff $\forall x_1, x_2, \exists u \in U$ such that $X_1(u) = x_1, X_2(u) = x_2$.

No device is distinguishable from itself. Distinguishability is symmetric, but non-transitive in general (and obviously not reflexive).

Having two devices be distinguishable means that no matter how the first device is set up, it is always possible to set up the second one in an arbitrary fashion; the setting up of the first device does not preclude any options for setting up the second one. Intuitively, if two devices are not distinguishable, then the setup function of one

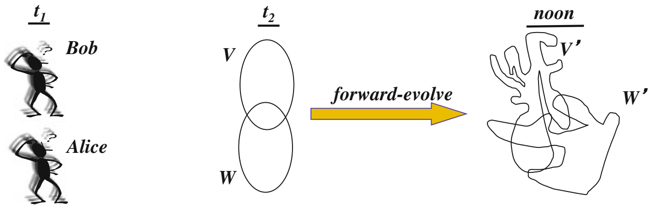


Fig. 9.2 The time t_1 is less than t_2, which in turn is less than noon. V is the set of all time-t_2 universes where Bob is thinking the answer "yes" in response to the t_1 question Bob heard—whatever that question was. W is the set of all time-t_2 universes where Alice is thinking the answer "yes" in response to the t_1 question Alice heard—whatever that question was. V' is V evolved forward to noon, and W' is W evolved forward to noon. At t_1, we ask Bob, "will the universe be in W' at noon?" (in other words, "was Alice thinking 'yes' at t_2?"). At that time we also ask Alice, "will the universe be outside of V' at noon?" (in other words, "was Bob *not* thinking 'yes' at t_2?"). It is impossible for both Bob and Alice to answer correctly, no matter what their computational capabilities are, what the laws of the universe are, etc.

of the devices is partially "controlled" by the setup function of the other one. In such a situation, they are not two fully separate, independent devices.

Proposition 2 *No two distinguishable devices* (X, Y) *and* (X', Y') *can weakly infer each other.*[6]

See Fig. 9.2 for an illustration of Proposition 2, for two IDs called "Bob" and "Alice".

This second Laplace's demon theorem establishes that a whole class of functions cannot be inferred by $\mathcal{D}$ (namely the conclusion functions of devices that are distinguishable from $\mathcal{D}$ and also can infer $\mathcal{D}$). More generally, let $\mathcal{S}$ be a set of devices, all of which are distinguishable from one another. Then the second demon theorem says that there can be at most one device in $\mathcal{S}$ that can infer all other devices in $\mathcal{S}$. It is important to note that the distinguishability condition is crucial to the second demon theorem; mutual weak inference can occur between non-distinguishable devices.

In (Barrow 2011) Barrow speculated whether "only computable patterns are instantiated in physical reality". There "computable" is defined in the sense of Turing machine theory. However we can also consider the term as meaning "can be evaluated by a real world computer". If so, then his question is answered—in the negative—by the Laplace demon theorems.

By combining the two demon theorems it is possible to establish the following:

Corollary 3 *Consider a pair of devices* $\mathcal{D} = (X, Y)$ *and* $\mathcal{D}' = (X', Y')$ *that are distinguishable from one another and whose conclusion functions are inequivalent. Say that* $\mathcal{D}'$ *weakly infers* $\mathcal{D}$*. Then there are at least three inequivalent surjective binary functions* Γ *that* $\mathcal{D}$ *does not infer.*

[6] In fact we can strengthen this result: If (X', Y') can weakly infer the distinguishable device (X, Y), then (X, Y) can infer neither of the two binary-valued functions equivalent to Y'. I will call Proposition 2 the "second (Laplace's) demon theorem".

In particular, Corollary 3 means that if any device in a set of distinguishable devices with inequivalent conclusion functions is sufficiently powerful to infer all the others, then each of those others must fail to infer at least three inequivalent functions.

Strong Inference—Inference of Entire Functions

As considered in computer science theory, a computer is an entire map taking an arbitrary "input" physical variable $\Gamma_1(u)$ to an "output" physical variable $\Gamma_2(u)$ (Hopcroft et al., 2000). It is concerned with saying how the value of $\Gamma_2(u)$ would change if the value of $\Gamma_1(u)$ changed. So it is concerned with two separate physical variables. In contrast, weak inference is only concerned with the value of a single physical variable. So we cannot really say that a device "infers a computer" in any sense if we only use the weak inference concept analyzed above. In this subsection we extend the theory of inference devices to include inference of entire functions. In addition to allowing us to analyze inference of computers, this lays the groundwork for the analysis in (Wolpert 2017) the relation between inference and algorithmic information theory.

To begin, suppose we have a function f that arises in the physical universe, in the sense that there is some S that is a function of U, along with some T, and S refines T, so that for all $s \in S(u)$, $f(s) = T(S^{-1}(s))$ is single-valued. We want to define what it means for a device to be able to emulate the entire mapping taking any $s \in S(U)$ to the associated value $T(S^{-1}(s))$.

One way to do this is to strengthen the concept of weak inference, so that for any desired input value $s \in S(U)$, the ID in question can simultaneously infer the output value $f(s)$ *while also forcing the input to have the value* s. In other words, for any pair $(s \in S(U), t \in T(U))$, by appropriate choice of $x \in X(U)$ the ID (X, Y) simultaneously answers the probe δ_t correctly (as in the concept of weak inference) *and* forces $S(u) = s$. In this way, when the ID "answers δ_t correctly", it is answering whether $f(s) = t$ correctly, for the precise s that it is setting. By being able to do this for all $s \in S(U)$, the ID can emulate the function f.

Extending this concept from single-valued functions f to multifunctions results in the following definition:

Definition 4 Let S and T be functions both defined over U. A device (X, Y) **strongly infers** (S, T) iff $\forall\ \delta \in \mathcal{P}(T)$ and all $s \in S(U)$, $\exists\ x$ such that $X(u) = x \Rightarrow \{S(u) = s, Y(u) = \delta(T(u))\}$.

If (X, Y) strongly infers (S, T) we write $(X, Y) \gg (S, T)$.

By considering the special case where $T(U) = \mathbb{B}$, we can formalize what it means for one device to emulate another device:

Definition 5 A device (X_1, Y_1) **strongly infers** a device (X_2, Y_2) iff $\forall\ \delta \in \mathcal{P}(Y_2)$ and all x_2, $\exists\ x_1$ such that $X_1 = x_1 \Rightarrow X_2 = x_2, Y_1 = \delta(Y_2)$.

See App. B in (Wolpert 2008) for a discussion of how unrestrictive Definition 5 is.

Example 3 Suppose $\mathcal{D}_2$ is a device that (for example) can be used to make predictions about the future state of the weather. Let Γ be the set of future weather states that the device can predict, and let X_2 be the set of possible current meteorological conditions. So if this device can in fact infer the future state of the weather, then for any question δ_γ of whether the future weather will have value γ, there is some current condition x_2 such that if $\mathcal{D}_2$ is set up with that x_2, it correctly answers whether the associated future state of the weather will be γ. On the other hand, if $\mathcal{D}_2 \not> \Gamma$, then there is some such question of the form, "will the future weather be γ?" such that for *no* input to the device of the current meteorological conditions will the device necessarily produce an answer y_2 to the question that is correct.

One way for us to be able to conclude that some device $\mathcal{D}' = (X', Y')$ can "emulate" this behavior of $\mathcal{D}_2$ is to set up $\mathcal{D}_2$ with an arbitrary value x_2, and confirm that $\mathcal{D}'$ can infer the associated value of Y_2. So we require that for all x_2, and all $\delta \in \mathcal{P}(Y_2)$, $\exists x'$ such that if $X_2 = x_2$ and $X' = x'$, then $Y = \delta(Y_2)$.

Now define a new device $\mathcal{D}_1$, with its setup function defined by $X_1(u) = (X'(u), X_2(u))$ and its conclusion function equal to Y'. Then our condition for confirming that $\mathcal{D}'$ can emulate $\mathcal{D}_2$ gets replaced by the condition that for all x_2, and all $\delta \in \mathcal{P}(Y_2)$, $\exists x_1$ such that if $X_1 = x_1$, then $X_2 = x_2$ and $Y = \delta(Y_2)$. This is precisely the definition of strong inference.

Say we have a Turing machine (TM) T_1 that can emulate another TM, T_2 (e.g., T_1 could be a universal Turing machine (UTM), able to emulate any other TM). Such "emulation" means that T_1 can perform any particular calculation that T_2 can. The analogous relationship holds for IDs, if we translate "emulate" to "strongly infer", and translate "perform a particular calculation" to "weakly infer". In addition, like UTM-style emulation (but unlike weak inference), strong inference is transitive. These results are formalized as follows:

Proposition 4 *Let $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$ be a set of inference devices over U and Γ a function over U. Then:*

(i) $\mathcal{D}_1 \gg \mathcal{D}_2$ *and* $\mathcal{D}_2 > \Gamma \Rightarrow \mathcal{D}_1 > \Gamma$.

(ii) $\mathcal{D}_1 \gg \mathcal{D}_2$ *and* $\mathcal{D}_2 \gg \mathcal{D}_3 \Rightarrow \mathcal{D}_1 \gg \mathcal{D}_3$.

In addition, strong inference implies weak inference, i.e., $\mathcal{D}_1 \gg \mathcal{D}_2 \Rightarrow \mathcal{D}_1 > \mathcal{D}_2$.

Most of the properties of weak inference have analogs for strong inference:

Proposition 5 *Let $\mathcal{D}_1$ be a device over U.*

(i) *There is a device $\mathcal{D}_2$ such that $\mathcal{D}_1 \not\gg \mathcal{D}_2$.*

(ii) *Say that $\forall\ x_1$, $|X_1^{-1}(x_1)| > 2$. Then there is a device $\mathcal{D}_2$ such that $\mathcal{D}_2 \gg \mathcal{D}_1$.*

Strong inference also obeys a restriction that is analogous to Proposition 2, except that there is no requirement of setup-distinguishability:

Proposition 6 *No two devices can strongly infer each other.*

Recall that there are entire functions that are not computable by any TM, in the sense that no TM can correctly compute the value of that function for every input to that function. On the other hand, trivially, any single output value of a function *can* be computed by some TM (just choose the TM that prints that value and then halts). The analogous distinction holds for inference devices:

Proposition 7 *Let U be any countable space with at least two elements.*

(i) *For any function Γ over U such that $|\Gamma(U)| \geq 3$ there is a device $\mathcal{D}$ that weakly infers Γ;*

(ii) *There is a function (S, T) over U that is not strongly inferred by any device.*

Proof Let $X(u)$ be the identity function (so that each $u \in U$ has its own, unique value x). Choose $Y(u)$ to equal 1 for exactly one u, $\bar{u}$. Then for the probe $\delta_{\Gamma(\bar{u})}$ we can choose $x = X(\bar{u})$, so that the device correctly answers 'yes' to the question of whether $\Gamma(u) = \Gamma(\bar{u})$. For any other probe δ_γ, note that since $|\Gamma(U)| \geq 3$, there must be a $u' \in U$ such that $\Gamma(u') \neq \gamma$. Moreover, by construction $Y(u') = -1$. So if we choose x to be $X(u')$, then the device correctly answers 'no' to the question of whether $\Gamma(u') = \gamma$. This proves the first claim.

To prove the second claim, choose both $S(u) = u$ and $T(u) = u$ for all u, so that $|S(U)| = |T(U)| = |U|$. So by the first requirement for some device (X, Y) to strongly infer (S, T), it must be that for any s, there is a value of X, $x(s)$, such that $X(u) = x(s) \Rightarrow S(u) = s$. This means that $x(s)$ must be a single-valued function, for each s choosing a unique (x which in turn choose a unique) u. This means that $Y(X^{-1}(x(s))$ must equal 1, in order for the device to correctly answer 'yes' to the probe of whether $T(u) = \delta_{T(S^{-1}(s))}$. However since this is true for all $s \in S(U)$, it is true for all $u \in U$. So $Y(U)$ is a singleton, contradicting the requirement that the conclusion function of any device be binary-valued. □

Modeling the Physical Universe in Terms of Inference Devices

I now expand the scope of the discussion to allow sets of many inference devices and/or many functions to be inferred. Some of the philosophical implications of the ensuing results are then discussed in the next subsection.

Formalization of Physical Reality Involving Inference Devices

Define a **reality** as a pair $(U; \{F_\phi\})$ where the space U is the **domain** of the reality, and $\{F_\phi\}$ is a (perhaps uncountable) non-empty set of functions all having domain U. We are particularly interested in **device realities** in which some of the functions are binary-valued, and we wish to pair each of those functions uniquely with some of the other functions. Such realities can be written as the triple $(U; \{(X_\alpha, Y_\alpha)\}; \{\Gamma_\beta\}) \equiv (U; \{\mathcal{D}_\alpha\}; \{\Gamma_\beta\})$ where $\{\mathcal{D}_\alpha\}$ is a set of devices over U and $\{\Gamma_\beta\}$ a set of functions over U.

Define a **universal device** as any device in a reality that can strongly infer all other devices and weakly infer all functions in that reality. Proposition 6 means that no reality can contain more than one universal device.

For simplicity, assume the index set ϕ is countable, with elements $\phi_1, \phi_2, \ldots$. It is useful to define the **reduced form** of a reality $(U; \{F_\phi\})$ as the image of the function $u \to (F_{\phi_1}(u), F_{\phi_2}(u), \ldots)$. In particular, the reduced form of a device reality is the set of all tuples $([x_1, y_1], [x_2, y_2], \ldots; \gamma_1, \gamma_2, \ldots)$ for which $\exists\, u \in U$ such that simultaneously $X_1(u) = x_1, Y_1(u) = y_1, X_2(u) = x_2, Y_2(u) = y_2, \ldots; \Gamma_1(u) = \gamma_1, \Gamma_2(u) = \gamma_2, \ldots$. By working with reduced forms of realities, we dispense with the need to explicitly discuss U entirely.[7]

Example 4 Take U to be the set of all possible histories of a universe across all time that are consistent with the laws of physics. So each u is a specification of a trajectory of the state of the entire universe through all time. The laws of physics are then embodied in restrictions on U. For example, if one wants to consider a universe in which the laws of physics are time-reversible and deterministic, then we require that no two distinct members of U can intersect. Similarly, properties like time-translation invariance can be imposed on U, as can more elaborate laws involving physical constants.

Next, have $\{\Gamma_\beta\}$ be a set of physical characteristics of the universe, each characteristic perhaps defined in terms of the values of one or more physical variables at multiple locations and/or multiple times. Finally, have $\{\mathcal{D}_\alpha\}$ be all prediction/observation systems concerning the universe that all scientists might ever be involved in.

In this example the laws of physics are embodied in U. The implications of those laws for the relationships among the agent devices $\{\mathcal{D}_\alpha\}$ and the other characteristics of the universe $\{\Gamma_\beta\}$ is embodied in the reduced form of the reality. Viewing the universe this way, it is the $u \in U$, specifying the universe's state for all time, that has "physical meaning". The reduced form instead is a logical implication of the laws of the universe. In particular, our universe's u picks out the tuple given by the Cartesian product $[\bigtimes_\alpha \mathcal{D}_\alpha(u)] \times [\bigtimes_\beta \Gamma_\beta(u)]$ from the reduced form of the reality.

As an alternative we can view the reduced form of the reality as encapsulating the "physical meaning" of the universe. In this alternative u does not have any physical meaning. It is only the relationships among the inferences about u that one might want to make and the devices with which to try to make those inferences that has physical meaning. One could completely change the space U and the functions defined over it, but if the associated reduced form of the reality does not change, then there is no way that the devices in that reality, when considering the functions in that reality, can tell that they are now defined over a different U. In this view, the laws of physics i.e., a choice for the set U, are simply a calculational shortcut for encapsulating patterns

[7] This means that all of the non-stochastic analysis of the previous sections can be reduced to satisfiability statements concerning sets of categorial variables. For example, the fact that a device cannot weakly infer itself is equivalent to the statement that there is no countable space X with at least two elements and associated set of pairs $\mathcal{U} = \{(x_i, y_i)\}$ where all $y_i \in \mathbb{B}$, such that for both probes δ of y_i, there is some value $x' \in X$ such that in all pairs $(x', y) \in \mathcal{U}$, $y = \delta(y)$.

in the reduced form of the reality. It is a particular instantiation of those patterns that has physical meaning, not some particular element $u \in U$.

See (Tegmark 2008) for another perspective on the relationship between physical reality and mathematical structures.

Given a reality $(U; \{(X_1, Y_1), (X_2, Y_2), \ldots\})$, we say that a pair of devices in it are **pairwise (setup) distinguishable** if they are distinguishable. We say that the reality as a whole is **mutually (setup) distinguishable** iff $\forall\ x_1 \in X_1(U), x_2 \in X_2(U), \ldots\ \exists\ u \in U$ s.t. $X_1(u) = x_1, X_2(u) = x_2, \ldots$.

Proposition 8 (**i**) *There exist realities* $(U; \mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3)$ *where each pair of devices is pairwise setup distinguishable and* $\mathcal{D}_1 > \mathcal{D}_2 > \mathcal{D}_3 > \mathcal{D}_1$.

(**ii**) *There exists no reality* $(U; \{\mathcal{D}_i : i \in \mathcal{N} \subseteq \mathbb{N}\})$ *where the devices are mutually distinguishable and for some integer n,* $\mathcal{D}_1 > \mathcal{D}_2 > \ldots > \mathcal{D}_n > \mathcal{D}_1$.

(**iii**) *There exists no reality* $(U; \{\mathcal{D}_i : i \in \mathcal{N} \subseteq \mathbb{N}\})$ *where for some integer n,* $\mathcal{D}_1 \gg \mathcal{D}_2 \gg \ldots \gg \mathcal{D}_n \gg \mathcal{D}_1$.

There are many ways to view a reality with a countable set of devices $\{\mathcal{D}_i\}$ as a graph, for example by having each node be a device while the edges between the nodes concern distinguishability of the associated devices, or concern whether one weakly infers the other, etc. In particular, given a countable reality, define an associated directed graph by identifying each device with a separate node in the graph, and by identifying each relationship of the form $\mathcal{D}_i \gg \mathcal{D}_j$ with a directed edge going from node i to node j. We call this the **strong inference graph** of the reality.

Proposition 7(ii) means that no reality with $|U| > 3$ can have a universal device if the reality contains all functions defined over U. Suppose that this is not the case, so that the reality may contain a universal device. Proposition 6 means that such a universal device must be a root node of the strong inference graph of the reality and that there cannot be any other root node. In addition, by Proposition 4(ii), we know that every node in a reality's strong inference graph with successor nodes has edges that lead directly to every one of those successor nodes (whether or not there is a universal device in the reality). By Proposition 8(iii) we also know that a reality's strong inference graph is acyclic.

Note that even if a device $\mathcal{D}_1$ can strongly infer all other devices $\mathcal{D}_{i>1}$ in a reality, it may not be able to infer them *simultaneously* (strongly or weakly). For example, define $\Gamma : u \rightarrow (Y_2(u), Y_3(u), \ldots)$. Then the fact that $\mathcal{D}_1$ is a universal device does not mean that $\forall \delta \in \mathcal{P}(\Gamma)\ \exists\ x_1 : Y_1 = \delta(\Gamma)$. See the discussion in)(Wolpert 2001) on "omniscient devices" for more on this point.

We now define what it means for two devices to operate in an identical manner:

Definition 6 Let U and $\hat{U}$ be two (perhaps identical) sets. Let $\mathcal{D}_1$ be a device in a reality with domain U. Let R_1 be the relation between X_1 and Y_1 specified by the reduced form of that reality, i.e., $x_1 R_1 y_1$ iff the pair (x_1, y_1) occurs in some tuple in the reduced form of the reality. Similarly let R_2 be the relation between X_2 and Y_2 for some separate device $\mathcal{D}_2$ in the reduced form of a reality having domain $\hat{U}$.

Then we say that $\mathcal{D}_1$ **mimics** $\mathcal{D}_2$ iff there is an injection, $\rho_X : X_2(\hat{U}) \rightarrow X_1(U)$ and a bijection $\rho_Y : Y_2(\hat{U}) \leftrightarrow Y_1(U)$, such that for $\forall x_2, y_2, x_2 R_2 y_2 \Leftrightarrow \rho_X(x_2) R_1 \rho_Y(y_2)$. If both $\mathcal{D}_1$ mimics $\mathcal{D}_2$ and vice-versa, we say that $\mathcal{D}_1$ and $\mathcal{D}_2$ are **copies** of each other.

Intuitively, when expressed as devices, two physical systems are copies if they follow the same inference algorithm with ρ_X and ρ_Y translating between those systems. In particular, say a reality contains two separate physical computers that are inference devices, both being used for prediction. If those devices are copies of each other, then they form the same conclusion for the same value of their setup function, i.e., they perform the same computation for the same input.

The requirement in Definition 6 that ρ_Y be surjective simply reflects the fact that since we're considering devices, $Y_1(U) = Y_2(U) = \mathbb{B}$. Note that because ρ_X in Definition 6 need not be surjective, there can be a device in U that mimics multiple devices in $\hat{U}$. The relation of one device mimicking another is reflexive and transitive. The relation of two devices being copies is an equivalence relation.

Say that an inference device $\mathcal{D}_2$ is being used for observation and $\mathcal{D}_1$ mimics $\mathcal{D}_2$. The fact that $\mathcal{D}_1$ mimics $\mathcal{D}_2$ does not imply that $\mathcal{D}_1$ can emulate the observation that $\mathcal{D}_2$ makes of some outside function Γ. The mimicry property only relates $\mathcal{D}_1$ and $\mathcal{D}_2$, with no concern for third relationships with any third function. (This is why for one device to "emulate" another is defined in terms of strong inference rather than in terms of mimicry.)

Proposition 9 *Let $\mathcal{D}_1$ be a copy of $\mathcal{D}_2$ and both exist in the same reality.*

(i) *It is possible that $\mathcal{D}_1$ and $\mathcal{D}_2$ are distinguishable and $\mathcal{D}_1 > \mathcal{D}_2$, even for finite $X_1(U), X_2(U)$.*

(ii) *It is possible that $\mathcal{D}_1 \gg \mathcal{D}_2$, but only if $X_1(U)$ and $X_2(U)$ are both infinite.*

Implications—How the Map Forces Itself on the Territory

The subject of this book is "The map and the territory". One way to interpret that phrase is that the "map" means a body of mathematics describing a universe, and the "territory" means that universe itself. One might suppose that a priori, it could be that the territory is not described by *any* map. Or that if it is described by a map, *which* map is in many ways arbitrary; many maps would work.

The results reviewed in this paper tell us that this is not the case. So long as we suppose that the universe (the "territory") allows for agents like us, then the theory of IDs (which is part of the "map") applies to the universe. In this sense, the map *forces itself upon the territory*, by imposing constraints on the possible properties of the territory.

As an illustration, return to the first case described in Example 4, where U is a set of laws of physics (i.e., the set of all histories consistent with a set of such laws). The results above provide general restrictions that must relate any devices in such a universe, regardless of the detailed nature of the laws of that universe. In particular, these results would have to be obeyed by all universes in a multiverse (Aguirre and Tegmark 2005; Carr 2007; Smolin 2002). Accordingly, it is interesting to consider these results from an informal philosophical perspective. Say we have a device $\mathcal{D}$ in a reality that is distinguishable from the set of all the other devices in the reality. Such a device can be viewed as having "free will", in that the way the other devices are set up does not restrict how $\mathcal{D}$ can be set up. Under this interpretation, Proposition 2 means that if two devices both have free will, then they cannot predict/recall/observe each other with guaranteed complete accuracy. A reality can have at most one of its devices that has free will and can predict/recall/observe/control the other devices in that reality with guaranteed complete accuracy.[8]

Proposition 6 then goes further and considers devices that can emulate each other. It shows that independent of concerns of free will, no two devices can unerringly emulate each other. (In other words, no reality can have more than one universal device). As mentioned above, somewhat tongue in cheek, taken together, these results could be called a "monotheism theorem".

Proposition 9 tells us that if there is a universal device in some reality, then it must be infinite (have infinite $X(U)$) if there are other devices in the reality that are copies of it. Now the time-translation of a physical device is a copy of that device.[9] Therefore any physical device that is *ever* universal must be infinite. In addition, the impossibility of multiple universal devices in a reality means that if any physical device is universal, it can only be so at one moment in time. (Its time-translation cannot be universal). Again somewhat tongue in cheek, taken together this second set of results could be called an "intelligent design theorem".

In addition to the questions addressed by the monotheism and intelligent design theorems, there are many other semi-philosophical questions one can ask of the

[8] There are other ways to interpret the vague term "free will". For example, Lloyd has argued that humans have "free will" in the sense that under the assumption that they are computationally universal, then due to the Halting theorem they cannot predict their own future conclusions ahead of time (Lloyd 2012). The fact that an ID cannot even weakly infer itself has analogous implications that hold under a broader range of assumptions concerning human computational capability, e.g., under the assumption that humans are not even computationally universal, or at the opposite extreme, under the assumption that they have super-Turing reasoning capability.

[9] Formally, say that the states of some physical system S at a particular time t and shortly thereafter at $t + \delta$ are identified as the setup and conclusion values of a device $\mathcal{D}$. In other words, $\mathcal{D}$ is given by the functions $(X(u), Y(u)) \triangleq (S(u_t), S(u_{t+\delta}))$. In addition, let R_S be the relation between X and Y specified by the reduced form of the reality containing the system. Say that the time-translation of $\mathcal{D}$, given by the two functions $S(u_{t'})$ and $S(u_{t'+\delta})$, also obeys the relation R_S. Then the pair of functions $(X_2(u), Y_2(u)) \triangleq (S(u_{t'}), S(u_{t'+\delta}))$ is another device that is copy of $\mathcal{D}$. So for example, the same physical computer at two separate pairs of moments is two separate devices, devices that are copies of each other, assuming they have the same set of allowed computations.

form “Can there be a reality with the following properties?”. By formulating them in terms of reduced realities, such questions can usually be reduced to a constraint satisfaction problem. In this sense, the theory of IDs allows us to reduce many of the questions that have animated philosophy since time immemorial into simple constraint satisfaction problems.

Acknowledgements I would like to thank Philippe Binder and especially Walt Read for many useful discussions. This work was supported by the Santa Fe Institute, Grant No. FQXi-RHl3-1349 from the FQXi foundation, and the Silicon Valley Foundation.

References

S. Aaronson, Why philosophers should care about computational complexity. Computability Turing Gödel Church Beyond 261–327 (2013)

A. Aguirre and M. Tegmark, Multiple universes, cosmic coincidences, and other dark matters, hep-th/0409072 (2005)

R.J. Aumann, Interactive epistemology ii: probability. Int. J. Game Theory **28**, 301–314 (1999)

R.J. Aumann, A. Brandenburger, Epistemic conditions for nash equilibrium. Econometrica **63**(5), 1161–1180 (1995)

R.J. Aumann, Agreeing to disagree. Ann. Stat. **4**(6), 1236–1239 (1976)

J.D. Barrow, Godel and physics. *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (2011), p. 255

P. Binder, Theories of almost everything. Nature **455**, 884–885 (2008)

K. Binmore, A. Brandenburger, Common knowledge and game theory. ST/ICERD Discuss. Pap. 88/167, London School of Economics (1988)

B. Carr (ed.), *Universe or Multiverse?* (Cambridge University Press, 2007)

G.J. Chaitin, *Algorithmic Information Theory*, vol. 1 (Cambridge University Press, 2004)

R. Fagin, J.Y. Halpern, Y. Moses, M. Vardi, *Reasoning about Knowledge* (MIT press, 2004)

D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, MA, 1991)

J.E. Hopcroft, R. Motwani, U. Rotwani, *JD: Introduction to Automata Theory, Languages and Computability* (2000)

S. Lloyd, Valuable information, *Complexity, Entropy and the Physics of Information* ed. by W.H. Zurek (1990), pp. 193–197

S. Lloyd, *Programming the Universe: a Quantum Computer Scientist Takes on the Cosmos* (Vintage, 2006)

S. Lloyd, A turing test for free will. Phil. Trans. R. Soc. A **370**(1971), 3597–3610 (2012)

D.M. MacKay, On the logical indeterminacy of a free choice. Mind, New Series **69**(273), 31–40 (1960)

R. Parikh, Knowledge and the problem of logical omniscience. ISMIS **87**, 432–439 (1987)

K. Popper, The impossibility of self-prediction, *The Open Universe: From the Postscript to the Logic of Scientific Discovery* (Routledge, 1988), p. 68

J. Schmidhuber, Algorithmic Theor. Everything. arXiv:quant-ph/0011122 (2000)

L. Smolin, *The Life of the Cosmos* (Weidenfeld and Nicolson, 2002)

M. Tegmark, The mathematical universe. Found. Phys. **38**(2), 101–150 (2008)

D.H. Wolpert, Computational capabilities of physical systems. Phys. Rev. E **65**, 016128 (2001)

D.H. Wolpert, Physical limits of inference. Phys. D **237**, 1257–1281 (2008). More recent version at http://arxiv.org/abs/0708.1362

D.H. Wolpert, Inference concerning physical systems, in *CiE Proceedings on Programs, Proofs, and Processes*, (2010), pp. 438–447

D.H. Wolpert, Constraints on physical reality arising from a formalization of knowledge. arXiv:1711.03499 [physics.hist-ph] (2017)

E.N. Zalta et al., *Stanford Encyclopedia of Philosophy* (2003)

H. Zenil, *A Computable Universe: Understanding and Exploring Nature as Computation*, (World Scientific Publishing Co., Inc., 2012)

W.H. Zurek, Algorithmic randomness and physical entropy. Phys. Rev. A **40**, 4731–4751 (1989a)

W.H. Zurek, Thermodynamic cost of computation, algorithmic complexity and the information metric. Nature **341**, 119–124 (1989b)

W.H. Zurek (ed.), *Complexity, Entropy and the Physics of Information* (Addison-Wesley, 1990)

K. Zuse, *Rechnender Raum (Calculating Space)* (1969)

Chapter 10
Substantivalism and Relationism as Bad Cartography: Why Spatial Ontology Needs a Better Map

Edward Slowik

In the philosophy of space and time, the "territory" is, of course, the ontology of space and time (i.e., its nature or being), whereas the principal "map" is the substantivalism (or absolutism) versus relationism dichotomy. Providing a precise definition of these opposing viewpoints is quite difficult, but a fairly straightforward reading is that substantivalists reckon space (or spacetime) to be an entity of some sort that can exist independently of material entities, whereas relationists contend that space (or spacetime) is a mere relation among material things, and is thus neither an entity nor independent of matter.[1] There are numerous difficulties with this modern or "standard dichotomy", as we will designate the debate between substantivalists and relationists, but our investigation will focus on two specific issues as a means of examining and developing alternative ontological conceptions of space that go beyond the limitations imposed by the standard dichotomy. First, while Newton and Leibniz are often upheld as the progenitors of, respectively, substantivalism and relationism, their own work in the natural philosophy of space often contradicts the central tenets of the modern dichotomy. A brief survey of Newton and Leibniz' respective views in the section, "The Spatial Ontology of Newton and Leibniz", will demonstrate this point. Second, while the modern substantivalist-relationist dichotomy is to some extent functional within the setting of Newtonian mechanics, it has proved extremely problematic when transferred to the setting of modern field theories, in particular, general relativity, but also as regards the more recent quantum gravity hypotheses. In the section, "Substantivalism and Relationism in General Relativity", the currently

[1] Throughout the remainder of our investigation, unless otherwise noted, the terms "space" and "spatial" will be used to designate both the standard 3-dimensional space as well as the 4-dimensional spacetime conception in modern relativity physics.

E. Slowik (✉)
Department of Philosophy, Winona State University, Winona, MN, USA
e-mail: eslowik@winona.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_10

accepted substantivalist and relationist interpretations of general relativity will reveal these difficulties.

The Spatial Ontology of Newton and Leibniz

Newton's concept of absolute space is the commonly accepted forerunner to modern substantivalism: "absolute space, of its own nature without reference to anything external, always remains homogeneous and immovable" (N 64).[2] But, is Newton's absolute space a substance (i.e., as in "substantivalism")? In the seventeenth century, a substance was usually defined as an independently existing thing, such that it could exist in the absence of all other substances—but that is not how Newton conceives the ontological status of space. In his unpublished work, *De gravitatione*, which most likely predates his most important scientific work, the *Principia*, Newton insists that space "has its own manner of existing which is proper to it and which fits neither substance nor accident [i.e., property]" (N 21). Space "is not a substance … because it is not absolute in itself, but is as it were an emanative effect of God and an affection of every kind of being; on the other hand, because it is not among the proper affections that denote substance, namely actions, such as thoughts in the mind and motions in body" (N 21). He adds that space is not a substance since it cannot "act upon things, yet everyone tacitly understands this of substance" (N 21), but neither is it an accident of body, "since we can clearly conceive extension existing without any subject, as when we imagine spaces outside the world or places empty of any body whatsoever" (N 22). Rather, as an "emanative effect of God", space would seem to be a property of God, and not an independently existing substance at all. In a later portion of this same work, Newton would seem to further advance this "space as God's property" view by conceiving a hypothetical world wherein God's own spatial extension is directly endowed with bodily properties, such as impenetrability or color, without requiring an underlying corporeal substance to house these accidents: "If [God] should exercise this power, and cause some space projecting above the earth, like a mountain or any other body, to be impervious to bodies and thus stop or reflect light and all impinging things, it seems impossible that we should not consider this space really to be a body from the evidence of our senses" (N 27-28). If we accept this hypothesis, then Newton contends that "we can define bodies as *determined quantities of extension which omnipresent God endows with certain conditions*" (28); the "conditions" being, first, that these determined quantities are mobile, second, that they can bring about perceptions in minds, and third, that two or more cannot coincide. By this process,

[2] The following abbreviations will be used for various cited works: AG = Leibniz (1989); NE = Leibniz (1996), referenced with book, chapter, section; N = Newton (2004); L = Leibniz and Clarke Leibniz (2000) (Leibniz letter), referenced with letter number and section; C = Leibniz and Clarke Leibniz (2000) (Clarke letter), referenced with letter number and section; Lm = Leibniz (1969).

Newton argues that these bundles of quantities can exactly replicate our everyday experience of material bodies without need of Descartes' conception of material substance, or the Scholastic notion of prime matter (N 27-31).

Oddly enough, the "space as God's property" view also seems applicable to Leibniz, although it includes a number of important differences, namely, that God grounds the existence of space but, unlike Newton, God's substance is not in space. Leibniz contends that space's "truth and reality are grounded in God, like all eternal truths", and that "space is an order [of situations] but that God is the source" (NE II. xiii.17). Yet, in contrast to Newton's hypothesis of a spatially extended God, Leibniz rebuffs the notion that "God discerns what passes in the world by being present to the things", rather, God discerns things "by the dependence on him of the continuation of their existence, which may be said to involve a continual production of them" (L.V.85). As he states in the *New Essay*, "God is not present to things by situation but by essence; his presence is manifested by his immediate operation" (L.III.12). All of these themes are nicely encapsulated in the following passage:

> Where space is in question, we must attribute immensity to God, and this also gives parts and order to his immediate operations. He is the source of possibilities and of existents alike, the one by his essence and the other by his will. So that space like time derives its reality only from him, and he can fill up the void whenever he pleases. It is in this way that he is omnipresent. (NE II.xv.2)

One of the differences between Newton and Leibniz' respective conceptions of space that is relevant to this discussion—but which has been overlooked by the substantivalism versus relationism debate—is the distinction between platonism and nominalism, where "platonists" hold that abstract objects (such as concepts, numbers, geometric structure, etc.) have an independent existence apart from the material things that exemplify those abstract objects, while "nominalists" reckon that abstract objects only exist in material things. In short, one of the components of Leibniz' theory of space that has been curiously overlooked by commentators is his tendency to equate spatial absolutism (substantivalism) with platonism, an outlook manifest in his assertion that "there is no need to postulate two extensions, one abstract (for space) and the other concrete (for body)" (NE II.iv.5). Spatial absolutists, conversely, do posit two extensions, one for space and one for body—but so do platonists: extension (space) as an abstract object that exists independently of all material instantiations, and extension (space) as a property of matter that constitutes the instantiation of that abstract object. Unlike the substantivalist/relationist dichotomy, consequently, the platonism/nominalism dichotomy does offer a way of securing a consistent interpretation of important features of Leibniz' natural philosophy of space; e.g., for why space is not an "absolute reality" (L.V.62), or as he more carefully puts it, "an absolute being" (L.III.5), *but* does expresses "real truths" (L.V.47). In essence, to regard space as an "absolute reality" or "absolute being" is "to postulate two extensions, one abstract (for space) and the other concrete (for body)", hence "space as absolute reality/being" can be interpreted as positing space as an entity that functions like a platonist's abstract object, i.e., space as a really existing (geometric) object that exists in addition to the existence of spatially-extended bodies.[3] On the other hand, to insist that "abstract space is that order of situations when they are conceived as being possible. Space is therefore something merely ideal" (L.V.104), while simultaneously claiming that space involves "real truths", is to employ a type of explanation that correlates with nominalism in the precise sense that it rejects space's status as an independently existing abstract object.

Furthermore, a nominalist about space can reject a relational account of space, i.e., one that claims that space is only the relations among bodies—and Leibniz does appear to undermine spatial relationism on many occasions. Consider the following account of "place" in the *New Essay* (1703), delivered by Leibniz' spokesman, Theophilus:

> 'Place' is either *particular*, as considered in relation to this or that body, or *universal;* the latter is related to everything, and in terms of it all changes of every body whatsoever are taken into account. If there were nothing fixed in the universe, the place of each thing would still be determined by reasoning, if there were a means of keeping a record of all the changes or if the memory of a created being were adequate to retain them—as the Arabs are said to play chess on horseback by memory. However, what we cannot grasp is nevertheless determinate in the truth of things. (NE II.xiii.8)

While it may seem innocent enough at first glance, the above quotation undercuts the prospects for any body-centered interpretation of Leibniz' concept of space, i.e., modern relationism. In the discussion that directly precedes this passage, Philalethes, expressing the empiricist's conception, contends that "same place" is relative to different contexts, and can thus be applied, for instance, to a chess-board in a ship: "The chess-board, we also say, is in the *same place* … if it remains in the same part of the cabin, though, perhaps, the ship which it is in [has set sail]" (NE II. xiii.8). Leibniz responds by referring to Philalethes' genuinely body-based relational conception as "particular" place, "in relation to this or that body", and he then goes on to contrast this idea with a "universal" notion of place which "is related to everything, and in terms of it all changes of every body whatsoever are taken into account". Leibniz' claim that "if there were nothing fixed in the universe, the place of each thing would still be determined by reasoning" is deeply antithetical to relationism, needless to say, and mimics Newton's use of absolute place in the *Principia* (N 66). Put simply, a relationist must define the notion of, for instance, "same place" by means of a material reference frame: they cannot, as does Leibniz, countenance the possibility that there may be no fixed material frames at all while simultaneously insisting that "same place" is still "determinate in the truth of things"—determinate with respect to what? Leibniz' claim implies that there is something else besides material existents, i.e., his universal place, that records all bodily changes of place, a conclusion that runs counter to all versions of

[3] In the correspondence with Des Bosses, Leibniz states that "I hold every absolute to be substantial [a substance]" (Lm 608). Hence, "space as absolute being" is identical to "space as substance" as well.

relationism. Of course, Leibniz would deny that universal place is an independent entity that can exist apart from matter; rather, it is simply an internal feature of some sort in bodies (and, ultimately, monads) that, presumably, allows a reconstruction of the prior places that bodies had occupied. While this last inference would seem correct, it nonetheless still falls afoul of relationist doctrine since any record or memory of a universal place within matter—a record by means of which "all changes of every body whatsoever are taken into account"—is akin to absolute place or space, with the only difference being the reinterpretation of absolute place as an internal feature of each body.

If Leibniz' ideas about space appear to be hostile to relationism, a natural defense might be to invoke his allegedly pure theory of relational motion, e.g., "[i]f we consider only what motion contains precisely and formally, that is, change of place, motion is not something entirely real, and when several bodies change position among themselves, it is not possible to determine, merely from a consideration of these changes, to which body we should attribute motion or rest" (AG 51). Thus, a central tenet of Leibniz' conception of motion is consistent with relationism, but there are other aspects of Leibniz' account that directly undermine relational motion. In particular, Leibniz conceives force, or the cause of motion, as breaking the symmetry of relational motion, so that bodies can now be assigned individual speeds (e.g., "body A is at rest, and body B is in motion", whereas a relationist can only claim that there is a speed/velocity difference between the bodies): after the above quotation, he adds "[b]ut the force or proximate cause of these changes is something more real, and there is sufficient basis to attribute it to one body more than another" (AG 51). So, once again, a closer look at the details of Leibniz' natural philosophy raises severe obstacles for a relationist interpretation.

In response, the devotee of the standard dichotomy might claim that Newton and Leibniz' spatial ontologies are simply bad examples of the substantivalism and relationism, but that mere historical fact does not diminish the significance of the standard dichotomy. While this defense has some merit, a close examination of the specific details of many, if not most, spatial ontologies in the Early Modern period (as well as most other periods) reveals a complexity that is often equal to Newton and Leibniz' respective conceptions, even if the details are different—and, more importantly, the complexity of these other spatial ontologies is not captured by the substantivalist versus relationist dichotomy either. Hence, given these numerous alternative conceptions that diverge from the standard dichotomy, the most reasonable course of action for the aspiring spatial ontologist should be to augment or change that dichotomy, or to jettison it altogether. For example, the distinction between platonism and nominalism as regards spatial structure is related to the standard dichotomy, but it does not align with it, and this difference helps to explain those aspects of Leibniz' spatial ontology that resemble absolute space (since a nominalist needs not be a relationist, as noted above). Likewise, as we have seen, both Newton and Leibniz posit a space that is much like a property of God, whether as an internal feature (Newton) or as a property that "emerges" in some fashion from a deeper non-spatial level of ontology (Leibniz). Yet, the standard dichotomy does not address these additional aspects of spatial ontology, and this helps to explain its inadequacies, both in terms of historical accuracy but also as regards its application to modern theories in physics (some of these problems will be examined in the next section). In particular, the grounding relationship between a foundational and a derived or emergent level of ontology, a feature that typifies the vast majority of seventeenth century spatial theories, is closely analogous to the grounding relationship postulated between the microlevel structures and processes and the resulting macrolevel phenomena in numerous current quantum gravity hypotheses (see, Slowik (2013, 2016), for more on this theme). Given a more accurate account of seventeenth century spatial theories, entirely new avenues can be opened up for the evaluation of all spatial ontologies, both in historical and metaphysical terms, since a host of additional tools will be available for the analysis of spatial ontologies.

Substantivalism and Relationism in General Relativity

In assessing the deficiencies of the substantivalist versus relationist dichotomy, the rise of twentieth century field physics plays a central role. In physics, a “field” is a physical quantity (usually represented by a number or geometric object, such as a vector or tensor) that is spread out over a space such that each point in the space has a value. The reason a physical field is problematic for the standard dichotomy is that both the substantivalists and relationists can make a persuasive case that their spatial ontology is the natural setting for a field: for the substantivalists, a field requires space (or spacetime) points for the assignment of the field quantity, and so they can claim that an independently existing space is a necessary prerequisite for assigning these quantities; for the relationist, on the other hand, a field is a physical *thing* that is spatially extended, and thus the spatial points of the field are really just material points or parts of that physical thing (and so an independently existing space is not required). Pre-twentieth century Newtonian mechanics, in contrast, provides a better framework for distinguishing between a substantivalist and relationist ontology since the Newtonian physical world mainly consists of bodies and empty space, e.g., a lone body in an otherwise empty universe cannot move for a relationist (since motion is a relation among bodies under relationism), but it can move for a substantivalist (since motion is a relation between a body and space under substantivalism). In contrast, it is difficult to characterize any scenario or event in a physical field that would clearly distinguish a substantivalist from a relationist interpretation of that situation (more on this below). In what follows, the focus of the analysis will be General Relativity (hereafter, GR), the theory that both replaced Newtonian gravitational theory as the best account of the large scale structure of space and time and which has also served as the battleground for the contemporary substantivalist versus relationist debate.

The Hole Argument and Sophisticated Substantivalism

Many recent investigations of spacetime structure can be traced to Earman and Norton's (1987) "hole" argument against the contemporary position dubbed "manifold substantivalism", the latter being the substantivalist hypothesis that that the substance of space/spacetime is the topological, differentiable manifold of points, M, which underlies all other spacetime structures. Informally, the manifold in a theory like GR furnishes the topology and continuity of spacetime, and hence presupposes that its basic elements, i.e., points, possess an identity apart from all of the higher level geometric structures that are mapped on to the manifold, such as vector or matter fields. In short, M allows position but lacks a concept of distance. While the exact details will be discussed below, a host of substantivalist-inclined commentators have instead opted for metrical structure as the basis of substantivalism, where the metric in GR, g, includes the geometry of spacetime, such as distance, curvature, and volume, as well as causal structure. Put briefly, the hole argument concludes that substantivalism, in its modern GR setting, leads to an unacceptable form of indeterminism. By shifting the metric and matter (stress-energy) fields, g, and, T, respectively, on the manifold of points, M, with the latter representing the substantivalist basis of spacetime, one can obtain different instances of the theory that satisfy the field equations of GR but which are observationally identical: i.e., some of these different instantiations of GR, which we can label, (M, $\breve{g}$, $\check{T}$), possess different geometries in various parts of the spacetime, namely, the "hole", despite the fact that they are observationally identical to other instantiations of GR, say, (M, g, T), which do not possess a different geometry in the hole. (More carefully, the new model is acquired from the old one via a "hole diffeomorphism", h: $\mathrm{h}(g) = \breve{g}$, $\mathrm{h}(T) = \check{T}$, and the mapping is the identity transformation outside of the hole, but a non-identity mapping inside the hole.) Overall, the substantivalist will not be able to determine the trajectory of a particle within the hole despite the observationally identical nature of the two worlds, (M, g, T) and (M, $\breve{g}$, $\check{T}$), i.e., the spacetime with, and without, the transformation, and hence the ability to uniquely determine the paths of particles is undermined.

One of the more influential substantivalist solutions to the hole argument is to reject a straightforward realist interpretation of the individuality of the points that comprise the manifold M. "A preferable alternative [to manifold substantivalism] is to strip primitive identity from space-time points: call this view *metric field substantivalism*. The focus of this view is on the metric tensor as the real representor of space-time in GTR" (Hoefer 1996, 24). Since the identity of the points of M are fixed by the metric g alone, any transformation of g, i.e., $\breve{g}$, does not result in the points of M possessing different g-values; rather, $\breve{g}$ simply gives back the very same spacetime points (since, to put it another way, a shift in g also shifts the identity criteria of the points of M along with it). In evaluating the various structural-role solutions to the hole argument (besides Hoefer (1996), there have been many similar hypotheses offered by others), the accusation can be made that metric field substantivalism bears a suspiciously close resemblance to relationism. In Belot and Earman (2001, 228), these structural-role constructions are somewhat pejoratively labeled "sophisticated substantivalism", with the charge being that the "substantivalists are helping themselves to a position most naturally associated with relationism" (Belot 2000, 588–589)—the rationale behind this estimate is that the identification of a host of observationally indistinguishable models with a single state-of-affairs is the very heart of relationism; e.g., Leibniz's rejection of the possibility that identical worlds could nonetheless possess a different position or speed in absolute space alone (L.III.5).

A few additional worries concerning metric field substantivalism should be recorded at this point. First, as Hoefer himself admits, the role of *M* in GR cannot be completely dismissed—namely, "it represents the continuity of space-time and the global topology" (Hoefer 1996, 24)—thus an alternative form of substantivalism, namely, "manifold plus metric substantivalism" (which Hoefer does not prefer), would seem to represent the mathematical structure of GR more faithfully. Here, the necessity of all of these mathematical structures for the entire spacetime is the key point. Hoefer provides two reasons for singling out the *g*-field (1996, 24): (i), that the "empirically useful" work of GR is "primarily" carried by the metric (such as inertial structure); and (ii), that a metric field without a global topology is possible "for at least small patches of space-time", although a manifold *M* without a metric cannot capture even a portion of spacetime. Yet, the problem with (i) is that *M* provides a great deal of useful empirical work as well, such as the dimension and global topology of the spacetime. If these aspects of the spacetime were missing or different, it would, needless to say, make a vast difference in our experience of that world. As for (ii), a small patch of spacetime is an incomplete and inadequate representation of the overall spacetime (global topology being a major concern, once again), hence it does little to support the notion that *g* alone is somehow more privileged than *M* alone. Another problem with metric field substantivalism is that by singling out a particular mathematical structure as "the real representor of space-time", despite the fact that the structure in question is only one of an interrelated set of such mathematical structures, sounds suspiciously like a form of ontological conventionalism—that is, one structure would seem to have been arbitrarily singled out as ontologically privileged even though the other structures are necessarily required for the overall spacetime. More carefully, Hoefer seems to be arbitrarily privileging a single structure, the *g*-field, from within the set of interrelated structures of GR's spacetime, (*M*, *g*, *T*), to serve as the ontological foundation of (*M*, *g*, *T*)—even though one could have as easily adopted, and offered plausible arguments for, the structure *M* to serve as the privileged ontological basis. Earman (1989), for example, develops arguments that favor the identification of *M* with substantivalism: e.g., "fields are not properties of an undressed set of space-time points but rather properties of the manifold *M*, which implies that fields are properties jointly of the points and their topological and differential properties" (1989, 201).

Sophisticated Relationism

Returning to the charge that sophisticated substantivalism steals a page from the relationist playbook, what is ironic about this allegation is that until recently they have usually gone the other way, with the substantivalists accusing many of the latest relational hypotheses of an illicit use of absolutist/substantivalist spatiotemporal structure. The allegedly non-relational structures can be classified broadly as those that pertain to space (topology, metric, etc.) and those that pertain to motion (in particular, acceleration and its accompanying force-effects). As regards space, a relationist who is confined to existing material bodies alone may not be capable of constructing the full range of spatial structures that are freely available on the absolutist/substantivalist picture. To take a simple example, if congruence (equal distance) relations are founded upon the measurements conducted on actually existing bodies, can the relationist consistently invoke the congruence of different regions of empty space? (Clarke raises this same objection against Leibniz in their famous correspondence on the nature of space; C.IV.41.) In order to meet this challenge, many relationists have adopted a modal or dispositional account of spatial structures which extends beyond the existing material bodies to cover all "physically possible configurations", e.g., a relationist can now invoke "possible bodies" in order to obtain a measure of spatial congruence in empty spaces. Since the use of possible bodies is often taken as a primitive notion by these "sophisticated relationists" (as we will dub this view), the substantivalist can thus claim that these relationists have illegitimately utilized an absolute structure that transcends the actual material relations among bodies. It has long been recognized that a physics limited to the mere relational motion of bodies, based on a strict brand of spatial relationism, faces serious difficulties in trying to capture the full content of modern physical theories. In GR, furthermore, the formalism of the theory makes it meaningful to determine if a lone body rotates in an otherwise empty universe, or whether the whole material content of the universe rotates in unison, but such possibilities are excluded on a strict form of relationism since there are no material bodies or frameworks from which to measure that single body's (or all bodies') rotation. Hence, there is natural motivation on the part of relationists to move away from a strict form of relationism (that only relies on existing bodies) to a version of relationism that allows hypothetical bodies or spacetime structures that go beyond strict forms of relationism. An example of a spacetime structure that violates strict relationism is the affine structure, ∇, which determines the inertial paths of bodies (i.e., ∇ is the structure that can determine whether a lone body in an empty universe accelerates, etc.). A potential rejoinder to the substantivalist's charge that relationism is usurping substantivalist spacetime structures is to reject the strict version of relationism, and simply hold that the spatial structures needed to make sense of motion and its effects do not supervene on some underlying, independent entity called "substantival space (spacetime)". The rejection of the strict brand of spatial relationism (that relies only on existing bodies) by these sophisticated relationisms allows the relationist to freely adopt any spatial structure required to explicate dynamical behavior, such as the affine structure, ∇, just as long as it is acknowledged that these structures are instantiated by material bodies or fields in some fashion. Yet, by embracing these richer structures, sophisticated relationism still remains open to the charge that it is an "instrumentalist rip-off" of substantivalism, which is the allegation leveled by Earman (1989, 128) at those spacetime ontologies that reject substantivalism but still employ non-relationist, substantivalist-leaning spacetime structures, such as ∇.

In the context of GR, however, the most plausible form of sophisticated relationism is "metric field relationism". As revealed above, Hoefer views the metric as representing substantival space, but Einstein judged the metric field as more closely resembling Descartes' theory of space—and Descartes' conception of space is normally categorized as relationist:

> If we imagine the gravitational field, *i.e.* the functions g_{ik}, to be removed, there does not remain a space … but absolutely *nothing* … There is no such thing as an empty space, *i.e.* a space without field. Space-time does not claim existence on its own, but only as a structural quality of the field.
>
> Thus Descartes was not so far from the truth when he believed he must exclude the existence of an empty space … It requires the idea of the field as the representative of reality, in combination with the general principle of relativity, to show the true kernel of Descartes' idea; there exists no space "empty of field". (Einstein 1961, 155–156)

In GR, The metric g provides the geometry of spacetime, but g also incorporates the inertial-gravitational field, thus explaining Einstein's nomenclature (but "metric field" is the default choice). By holding that all fields, including the metric/gravitational field, g_{ik}, are physical fields, the vacuum solutions to GR (which are void of matter) no longer correspond to empty spacetimes, thus eliminating a major obstacle to relationism. According to the standard matter-based conception of relationism, a spacetime entirely void of matter corresponds to either a meaningless state-of-affairs or a universe without space (spacetime). Yet, the vacuum solutions to the field equations in GR (where the stress-energy field, T, is 0) are a meaningful, as well as possible, state-of-affairs. Hence, if a relationist deems the metric field to be a physical field, then this maneuver automatically renders the vacuum solution of GR to be relationally meaningful since there is no spacetime, to use Einstein's phrase, "empty of [metric] field" (even in a matter-less vacuum): on Dieks' succinct explanation of this version of relationism, "[t]he metric field, a physical system of the same kind as [particle physics], is the bearer of the geometrical space-time properties" (Dieks 2001, 14).

A substantivalist might reject this interpretation of the metric field, g, and strive to formulate some ontological criterion for differentiating substantival and non-substantival aspects of GR. Unfortunately, given the peculiar complexities of the field concept in physics, as well as the argument that this particular field can produce measurable physical effects, it becomes difficult to imagine how such a criterion could avoid arbitrary ontological stipulations that beg the question against the relationist. As put forth in Earman and Norton (1987), the gravitational waves that can propagate through the empty spacetime solutions of GR are associated with the physical/material content of the spacetime, and not the underlying substantival space, since "in principle [the wave's] energy could be collected and converted into other types of energy, such as heat or light energy or even massive particles" (1987, 519).

Hoefer (2000) challenges Earman and Norton's conclusion by insisting that it depends on a well-defined conception of the stress-energy carried by the gravitational field, which, he adds, is not actually sanctioned by GR. Specifically, the problem arises because the term that represents the stress-energy of the gravitational field, t^{ab}, is a pseudo-tensor, where "its non-tensorial nature means that there is no well-defined, intrinsic 'amount of stuff' present at any given point" (2000, 193). Yet, Hoefer's response is not without its own set of difficulties, for it seems to entail that the energy lost by a gravity wave source is a real loss, and is not simply somewhere else in space; likewise, the energy gained by a gravity wave detector is a real gain, in apparent *ex nihilo* fashion. As Hoefer admits, "such a perspective seems to strain our general cause-effect intuitions by positing a cause-effect relationship without an intermediary carrier" (2000, 196).[4] In short, Hoefer's interpretation would forfeit energy-momentum conservation in GR, which is counter-intuitive given the long-standing presumption that this conservation law is a cornerstone of modern physics. On these grounds, if one is forced to choose between the rejection of the conservation of energy-momentum in GR versus a non-localizable conception of the stress-energy associated with the gravitational field, then the latter seems a much more preferable alternative. Friedman has likewise strived to uphold a substantivalist interpretation of the metric field, *g*, arguing that "it accords with the general-relativistic practice of not counting the gravitational energy induced by *g* as a component of the total energy, and it allows us to preserve a measure of continuity between general relativity and our previous space-time theories" (Friedman 1983, 223). But the practice of not counting the gravitational energy does not undermine the argument put forward by Earman and

[4] For the committed relationist, the vacuum solutions to GR without gravity waves might seem to undermine the suggestion made by Earman and Norton, since the absence of gravity waves means that there is no energy to convert into particles, etc. However, since the spacetime would still possess the capacity to generate waves, it is unclear that there is any real difference in this case for those who side with Einstein's hypothesis that the *g* field is a physical field. A relationist could also accept Harré's suggestion that the vacuum solutions to the field equations of GR have "no reasonable physical interpretations" (Harré 1986, 131), but this seems a rather unwarranted strategy since, as explained above, the vacuum solutions are mathematically meaningful. Some authors have attempted to account for the gravitational field, *g*, using the motions of point particles alone (Vasallo and Esfeld 2016), but the recent discovery of gravitational waves over the past few years completely undermines this approach. In the LIGO measurements from 2016, two black holes combined but lost roughly 3 solar masses that were converted into gravity waves. Hence, gravity waves exist, and so the claim by these authors, that an empty universe with gravity waves is merely "mathematical surplus structure" (Vasallo and Esfeld 2016, 104) is untenable. But there is an even more problematic issue at stake: Are Vasallo and Esfeld committed to the view that gravity waves exist if there is a particle somewhere in the universe that can be influenced by these waves, but that those same gravity waves just disappear if that particle is removed? This would be a highly implausible interpretation, needless to say.

Norton above, i.e., that gravitational energy can be converted into other forms of energy, and is thus real. Likewise, the alleged continuity between GR's metric field and our previous theory of the large scale structure of space, Newtonian gravitation theory, is precisely the point at issue, and so one cannot simply assume it. While the basic metric and manifold structure of *g* is, of course, closely related to the Euclidean geometry of Newtonian theory, the gravitational field aspect of *g* is decidedly not. Following Einstein's lead, Rovelli argues that *g* is much closer to matter than a spatiotemporal backdrop: "In general relativity, the metric/ gravitational field has acquired most, if not all, the attributes that have characterized matter (as opposed to spacetime) from Descartes to Feynman: it satisfies differential equations, it carries energy and momentum, and, in Leibnizian terms, *it can act and also be acted upon*, and so on" (Rovelli 1997, 193).

A substantivalist could also argue that the proper or most defensible conception of relationism in the context of GR is the Machian account, which in its most robust form, dubbed "Mach-heavy" in Huggett and Hoefer (2015), dictates that the spatial geometry and inertial structure of the spacetime, *g*, must be determined by the material content of that spacetime, i.e., the stress-energy, *T*. Yet, Mach-heavy is just one interpretation of relationism in a GR setting; and, if Einstein's views are deemed a deciding factor, then he had long abandoned the Mach-heavy conception by the time that he posited his Cartesian interpretation of the theory (as opposed to the Newton interpretation) in the quotation provided above.[5] Consequently, the proper relationist account of GR is as indeterminate as the proper relationist interpretation of Newtonian theory.

On the whole, the predicament posed by these conflicting substantivalist and relationist interpretations of *g* has led some to question the relevance of the standard dichotomy (substantivalism versus relationism) in the context of GR: if *g* can be coherently viewed as supporting either relationism or substantivalism, then what remains of the standard dichotomy that is useful in analyzing the ontology of GR (see, e.g., Rynasiewicz 1996)? Put differently: Is there really any meaningful difference between calling GR's metric field either a unique substance (substantivalism) or a unique physical entity (relationism)? Likewise, while the hole argument has motivated the development of sophisticated versions of substantivalism, which

[5] "Mach-lite" is the standard anti-substantivalist rejection of substantival space for a relationally acceptable alternative, such as the fixed stars, or, better yet, the center-of-mass reference frame of the world, which Mach stipulates must not accelerate (Mach 1960, 287). The most significant problem for Mach-heavy is the fact that the boundary conditions of GR's field equations are not totally determined by *T*, but have to be specified with respect to a choice of *g* as well. Hence, Mach-heavy seems inconsistent with the mathematical structure of GR. Brown (2017) attempts to defend a substantivalist interpretation of the metric in GR by invoking the universe's expansion: "If the universe expands but there is no material object expanding and there is no rearrangement of material objects relative to one another, then something non-material expands. This something is obviously space" (2017, 86). But, as Rovelli makes clear in the quotation above, "the metric/ gravitational field has acquired most, if not all, the attributes that have characterized matter (as opposed to spacetime)", so the claim that the metric, *g*, is non-material is simply untenable (and it also begs the question against the metric field relationist)

resemble relationism, the difficulties associated with strict forms of relationism have prompted the development of sophisticated forms of relationism, which resembles substantivalism. Given these developments, it is hard not to conclude that the present state of the substantivalist versus relationist debate is rather muddled.

Conclusion

Finally, it should be briefly noted that the metaphysical quagmire just outlined in the section "Substantivalism and Relationism in General Relativity" has also ensnarled philosophers concerned with the ontology of quantum gravity theories (QG), an assortment of strategies whose goal is to connect the physics at the micro-realm of quantum mechanics (QM) with the large-scale structure of space and time in GR. In short, the general consensus would seem to be that sophisticated versions of both substantivalism and relationism are equally consistent, or equally problematic, interpretations of QG (e.g., Rickles 2005; Earman 2006), a conclusion that is apparently reflected in the rival appropriations of an important QG hypothesis, loop quantum gravity (LQG), for either Leibnizian relationism or Newtonian substantivalism. For example, a Leibnizian lineage for LQG has been put forward by Smolin (2006, 200–203), among many others. Yet, in Dainton (2010), it is argued that the ontology of LQG "seems as substantival as any conception", prompting Dainton to ask, "What could be less Leibnizian?", despite the fact that LQG is "very different from Newton's absolute space" (405-406). As discussed in the section "The Spatial Ontology of Newton and Leibniz", however, the claim that Leibniz is a relationist, and Newton a substantivalist, is controversial as well, since both support ontological positions that are extremely difficult to reconcile with those particular spatial ontologies. Hence, given the confusion and uncertainties involved with applying the standard substantivalist versus relationist dichotomy, whether in the seventeenth century or the twenty-first, it is long past time to search for new methodological and taxonomic tools—better "maps"—for the assessment of the "terrain" of spatial ontology.

References

G. Belot, Geometry and motion. Br. J. Philos. Sci. **51**, 561–595 (2000)

G. Belot, J. Earman, Pre-socratic quantum gravity, in *Physics meets Philosophy at the Planck Scale*, ed. by C. Callender, N. Huggett (Cambridge University Press, Cambridge, 2001), pp. 213–255

J.R. Brown, Why spacetime has a life of its own, in *Space*, *Time*, *and the Limits of Human Understanding*, ed. by S. Wuppuluri, G. Ghirardi (Springer, Cham, Switzerland, 2017), pp. 77–86

B. Dainton, *Time and Space*, 2nd edn. (McGill-Queen's University Press, Montreal, 2010)

D. Dieks, Space-time relationism in newtonian and relativistic physics. Int. Stud. Philos. Sci. **15**, 5–17 (2001)

J. Earman, *World Enough and Space-Time* (MIT Press, Cambridge, 1989)

J. Earman, The implications of general covariance for the ontology and ideology of spacetime, in *The Ontology of Spacetime*, vol. 1, ed. by D. Dieks (Elsevier, Amsterdam, 2006), pp. 3–23

J. Earman, J. Norton, What price space-time substantivalism? the hole story. Br. J. Philos. Sci. **38**, 515–525 (1987)

A. Einstein, Relativity and the problem of space, in *Relativity: The Special and the General Theory* (Crown Publishers, New York, 1961)

M. Friedman, *Foundations of Space-Time Theories* (Princeton University Press, Princeton, 1983)

R. Harré, *Varieties of Realism* (Blackwell Publishers, Oxford, 1986)

C. Hoefer, J. Smylie, The metaphysics of space-time substantivalism. J. Philos. **93**(1), 5–27 (1996)

C. Hoefer, Energy conservation in GTR. Stud. Hist. Philos. Mod. Phys. **31**, 187–199 (2000)

N. Huggett, C. Hoefer, Absolute and relational theories of space and motion, in *The Stanford Encyclopedia of Philosophy*, ed. by E. Zalta (2015). http://plato.stanford.edu/archives/spr2015/entries/spacetime-theories

G.W. Leibniz, in *Leibniz: Philosophical Letters and Papers*, 2nd edn., ed. and trans. by L.E. Loemker (Kluwer, Dordrecht, 1969)

G.W. Leibniz, in *Leibniz: Philosophical Essays*, ed. and trans. by R. Ariew, D. Garber (Hackett, Indianapolis, 1989)

G.W. Leibniz, in *New Essay on Human Understanding*, ed. and trans. by P. Remnant, J. Bennett (Cambridge University Press, Cambridge, 1996)

G.W. Leibniz, S. Clarke, in *Leibniz and Clarke Correspondence*, ed. and trans. by R. Ariew (Hackett, Indianapolis, 2000)

E. Mach, in *The Science of Mechanics*, 6th edn., tans. by T. McCormack (Open Court, London, 1960)

I. Newton, in *Philosophical Writings*, trans. and ed. by A. Janiak, C. Johnson (Cambridge University Press, Cambridge, 2004)

D. Rickles, A new spin on the hole argument. Stud. Hist. Philos. Mod. Phys. **36**, 415–434 (2005)

C. Rovelli, Halfway through the woods: contemporary research on space and time, in *The Cosmos of Science*, ed. by J. Earman, J. Norton (University of Pittsburgh Press, Pittsburgh, 1997), pp. 180–224

R. Rynasiewicz, Absolute versus relational space-time: an outmoded debate? J. Philos. **XCIII**, 279–306 (1996)

E. Slowik, The deep metaphysics of quantum gravity: the seventeenth century legacy and an alternative ontology beyond substantivalism and relationism. Stud. Hist. Philos. Mod. Phys. **44** (2013), 490–499 (2013)

E. Slowik, *The Deep Metaphysics of Space* (Springer, Cham, Switzerland, 2016)

L. Smolin, The case for background independence, in *The Structural Foundations of Quantum Gravity*, ed. by D. Rickles, S. French, J. Saatsi (Oxford University Press, Oxford, 2006), pp. 196–239

A. Vasallo, M. Esfeld, Leibnizian relationalism for general relativity physics. Stud. Hist. Philos. Mod. Phys. **55**, 101–107 (2016)

Chapter 11
Force in Physics and in Metaphysics: A Brief History

Barry Dainton

The OED provides us with a characteristically concise and illuminating definition of *force*: "An influence tending to change the motion of a body or produce motion or stress in a stationary body. The magnitude of such an influence is often calculated by multiplying the mass of the body and its acceleration." That there are forces in precisely this sense at work in our world may seem so obvious that only the most radical of sceptics would dream of denying it. As every child soon discovers, moving a heavy object (such as a brick) requires more effort—and hence more force—than moving a light one (such as a feather). Getting a bicycle to move requires more than merely sitting on a saddle: one has to apply force to the pedals, the greater the force exerted, the greater the speed. Needless to say, there are many other instances of forces at work. The force of a strong wind can almost blow one over. Many children find magnets fascinating because of the way they can exert an influence—when attracting some paperclips, say—though empty space, seemingly *directly*, almost by magic. Later on we learn that Newton's gravitational force is responsible for keeping us bound to the Earth's surface, and responsible too for keeping the planets in orbit around the sun. During our school careers, many of us will also have been taught (even if we later forget it) Newton's second law, $F = ma$, which encapsulates the relationship between mass, acceleration and force referred to in the OED definition.

As will already be clear, there are two interrelated notions of force what we need to distinguish.

First, there is the very general notion of a force as *something which makes something else happen*, where the "somethings" in question are physical objects or events. When a hammer knocks in a nail into a piece of wood, the hammer not only exerts a force on the nail (assuming for the moment that forces do exist), it also *causes* the nail to move. Since force is an instance of the more general notion of a cause, and "cause" is itself a controversial and contested concept, the metaphysical controversies surrounding the latter will naturally extend to the former.

B. Dainton (✉)
Department of Philosophy, University of Liverpool, Liverpool, England
e-mail: Bdainton@liverpool.ac.uk

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_11

The influential 18th century philosophical critiques of causation in all its forms by Berkeley, Hume—and Kant's response to these—had an impact on physics and metaphysics which endures to this day.

Second, there is the narrower and more specific conception of force as it features in one or other scientific theories. The accounts of *what makes a body move* to be found in the Democritus, Aristotle, Kepler, Newton, Maxwell and today's quantum mechanics and relativity theories differ in profound and interesting ways—as do the understandings of what "a material body" might be. Since different physical theories often posit different types of force, proponents of competing theories will inevitably have different views regarding the forces that are actually operative in nature. There are also disagreements over the kind of force *it makes sense* to think might exist. One particularly important and controversial instance is "action at a distance" forces. Forces of this kind, if they were to exist, operate directly across a spatial interval without any intermediaries. When a magnet picks up a metal paperclip it seems to be operating in this way. As we shall be exploring in more detail below, some prominent theories—most notably Newton's theory of universal gravitation—rely on action at a distance forces, whereas others—most notably Maxwell's electromagnetism and Einstein's general theory of relativity—make a virtue of *not* relying on them.

To add to an already complex picture, a further complicating factor is the manner in which scientific developments can influence metaphysical doctrines concerning forces and causes. Hume was famously sceptical with regard to causation. He is associated with the doctrine that nothing *makes* anything else happen, despite our natural tendency to think otherwise. If Hume is right, in a very real sense there are no forces—as we intuitively conceive them—to be found in nature at all, and the laws of nature do not constrain or necessitate, they merely reflect regularly occurring patterns among objects and events. The counterintuitive Humean view of causes and forces has always appealed to those hard-nosed empiricists who are wary of believing in anything that cannot be directly perceived. As we shall see in due course, it also receives support from the four-dimensional conception of the universe derived from Einstein's relativity theories.

In line with this chapter's title, the bulk of what follows will be historical in character. Inevitably, this might give rise to questions such as these.

> Why bother looking at the conceptions of force that can be found in old and discarded scientific theories? We want to distinguish the true the nature of force from the fictions and myths surrounding it. We want to know how force figures in the actual territory, as opposed to maps of it that we know to be erroneous. Given this, shouldn't we be concerning ourselves solely with *today's* physics?

This might be an option if contemporary physicists were in agreement as to the fundamental nature of physical reality, but as is widely acknowledged, this is far from being the case. General relativity is our best theory of the large-scale structure of the universe and gravity, quantum mechanics is our best theory of the (very) small-scale structure of reality. But in their current forms the two are radically incompatible, and finding a theoretical framework capable of accomodating both—the task Einstein worked on (fruitlessly) in the final decades of his life has proved very difficult.[1] There is certainly no shortage of intriguing speculation as to the form a "quantum theory of gravity" might take, ranging from classical canonical approaches to string/M theories, loop quantum gravity and causal set theory. But since there is no agreement as to which of these very different approaches is closer to the truth, contemporary physics is unable to provide an answer to the simple but basic question: "What is the fundamental nature of physical reality?"

In the light of this, it is well worth taking a brief look at how the conceptions of force have figured in earlier scientific theories and controversies. Since very different modes of physical interaction have been seriously considered by earlier scientists and natural philosophers, a historical survey provides some indications as to the modes of interaction which could easily feature in the physics of the future. No less importantly, a number of historical debates concerning the intelligibility (or possibility) of certain particular forms of physical interaction also have contemporary relevance—as we shall see in due course.

Ancient Forces

Anyone seeking to make sense of the universe will find that ordinary observable objects and processes pose a sizable number of challenges. Why do things fall when dropped? Why doesn't the moon fall out of the sky? How does the sun manage to rise every day? Why does fire rise upwards rather than downwards (or sideways)? If you push a stone across a table, why does it stop? What path does an arrow take when it flies through the sky? How can lodestone and amber affect things without touching them? Why is it that water can be absorbed by a cloth or sponge, but not by a hunk of rock? The ancient Greek natural philosophers were very much interested in making sense of the world, and devised a sizable number of very different explanatory schemes, several of which went on to have considerable influence in the millennia to come.

Following the lead of Parmenides, the early Greek atomists—most prominently Democritus, Epicurus and Lucretius—drew a sharp distinction between being and non-being. The latter is identified with pure nothingness in the form of an infinitely vast and utterly empty space, or *void*. Being comes in the form of a very large number of very small material atoms, indestructible and impenetrable, varying in shape and size. The universe as a whole consists of nothing more than atoms moving through the void. Democritus held that atoms have an in-built tendency to move, and so are in constant motion in all directions and do not require constant pushing.

[1] Or as Rovelli (2008, 4) puts it: "In spite of their empirical success, GR and QM offer a schizophrenic and confused understanding of the physical world. The conceptual foundations of classical GR are contradicted by QM and the conceptual foundation of conventional QM are contradicted by GR. Fundamental physics today is in a peculiar phase of deep conceptual confusion."

Epicurus would later account for gravitational effects by holding that atoms have an innate tendency to move downwards. When atoms collide they sometimes rebound off one another, but sometimes stick together—a process made possible by their shapes: Democritus believed that some atoms had hooks. For the atomists, all compound physical things—such as planets, rocks, liquids and animals—are the product of atomic collisions and subsequent adhesions, or as Aristotle concisely summarizes: "The atoms act and suffer action whenever they chance to be in contact … and they generate by being put together and intertwined".[2]

How could magnetic and electrical attraction be explained in these terms? As noted above, magnetism *seems* to work directly across spatial intervals, with no intervening material mediation. So can physical forces be transmitted through the void? The atomists were unanimous in wholly rejecting action at a distance. The only things which can causally act on physical bodies are *other physical bodies*, and since the void is entirely devoid of any sort of physical body, no causal influence can be transmitted through it. The atomists were obliged to explain magnetic and electrical forces mechanically. We thus find Lucretius suggesting that lodestones emit invisible streams of particles which displace the atoms of air in their surroundings. The small regions of void that are produced by this process result in pressure differences which lead to pieces of iron in the vicinity of a lodestone to move in the latter's direction.

Plato agreed with the atomists on some issues. In the *Timaeus* he suggests they were right to hold that macroscopic material things are composed of interacting smaller constituents—influentially, he also argued that when accounting for the behaviour of these systems mathematics should be deployed. But he felt unable to accept that purely *mechanical* model of the universe that the atomists were proposing held all the answers.

Plato found it implausible to suppose that essentially random atomic motion could give rise to the regular motions of the sun, moon and planets, or highly complex internally powered lifeforms such as plants, animals and human beings. He was thus led to posit that the physical world was ultimately controlled by an all-pervasive *mind* or *spirit*:

> … we must declare that the only existing thing which properly possesses intelligence is soul, and this is an invisible thing, whereas fire, water, earth and air are all visible bodies; and a lover of intelligence and knowledge must necessarily seek first for the causation that belongs to the intelligent nature, and only in the second place for that which belongs to things that are moved by others and of necessity set yet others in motion. We too, then, must proceed on this principle: we must speak of both kinds of cause, but distinguish causes that work with intelligence to produce that which is good and desirable, from those which, destitute of reason, produce their sundry effects at random and without order. (*Timaeus*, 46)

When it came to explaining in a detailed way precisely how the world-soul was related to the material world Plato had little to say beyond pointing to the analogy

[2] *On Generation and Corruption*, 325a.

with the relationship between human souls and human bodies—a relationship that is itself less than transparent.

Aristotle was the ancient Greek natural philosopher whose views had the greatest influence over the course of the subsequent millennium. Like the atomists Aristotle wanted to develop a "theory of everything", and in his *Physics* he proposes principles which can explain motion and change in all their many and varied forms, on the Earth and in the Heavens. Like Plato, he found the mindless purely mechanical cosmos of the ancient atomists implausible.

For Aristotle the world is very much as it seems to be. The Earth is motionless, sitting at it does at the very centre of the universe, and the sun, stars and planets rotate around it. The most basic kinds physical things are the primary elements (earth, fire, water, air), and material substances composed of these, the prime examples of which are living organisms: cats, dogs, horses, fish, trees and the like. In the Aristotelian scheme all physical things are "hylomorphic", combinations of basic material stuff and *substantial forms*. A dog and an oak tree are both composed of material stuff—no doubt different proportions of the four elements—but they are obviously very different. As well as differing in shape, size, colour and internal structures they differ in what they are able *to do*: dogs have the capacity to jump and run, oak trees do not. According to Aristotle the differences between dogs and trees is due the active principle of organization, the form, which animates and bestows qualitative properties and causal powers on the matter composing them. Taken by itself, basic (or "prime") matter is inert and incapable of doing anything. It is only when it is infused with (or possessed by) a controlling form that it can constitute things of the sorts we are familiar with. All the different types of thing to be found in nature have their own distinctive form.

Aristotle recognized that living and non-living things move in different ways: living things have the capacity to *move themselves*, whereas inanimate objects do not. In explaining why inanimate objects move as they do he appealed to a distinction between *natural* and *non-natural* forms of motion.

He held that the different elements each have their own "proper" or "natural" place" and when an element is removed from its natural place it immediately attempts to return to it. Fire rising is an instance of natural motion—the natural place of fire is above the air, just below the celestial sphere carrying the moon; when a stone falls it too is striving to return to its natural place: at the centre of the universe. Non-natural motion occurs, as one might expect, when something intervenes to prevent an object following its natural course—e.g. when someone catches a falling stone.

Aristotle agreed with the atomists as to the character of non-natural motion: it occurs only by immediate *contact* between mover and thing moved. Or as he puts it: "The immediate agent of bodily change of place must be either in contact with or continuous with the moved object … as we always observe this to be the case" (*Physics*, 242). Like the atomists, Aristotle observed that the vast majority of non-natural causal interactions between material bodies involve contact—generally speaking things move only when they are pushed, pulled, kicked or thrown—and drew the conclusion that *all* such interactions require contact, and so rejected of action at a distance. Aristotle diverged from the atomists on the issue of the void: the Aristotelian cosmos is a plenum, containing no regions of totally empty space. The atomists held that the void is necessary for motion to be possible at all. Rejecting this reasoning, Aristotle points out that objects can move through perfectly easily through fluids—e.g. when we draw our fingers through a pool of water—without creating any gaps or voids.

The New Mechanical View

The Aristotelian system explains—and so makes sense of—all the natural phenomenal we encounter in ordinary life in an intuitively plausible way. It also puts us right at the centre of the cosmos, a view which fits nicely with the theological doctrine that we are the favoured creations of God. However, as a program in natural philosophy, by the 14th and 15th century Aristotelianism had also largely stagnated. On many fronts our understanding of the natural world had made little real progress in centuries. Those who believed radical progress was both desirable and possible—people such as Bacon, Galileo, Hobbes, Boyle, Kepler, Descartes and Newton—also recognized that a necessary first step was the overthrow Aristotelian physics. The full story of "The Scientific Revolution" is a highly complex one, extending as it does over several centuries, and involving a great many thinkers—some famous, many forgotten—operating in different traditions (and in different countries). I will confine myself to outlining just a few key developments that are particularly relevant to the role of force in science as it evolved during this period.

In his recent *The Swerve: How the World Became Modern* (2011) Stephen Greenblatt describes how the discovery of a copy of Lucretius' *On the Nature of Things* in the 15th century led to the rediscovery of ancient Greek atomism. Inspired by the mechanistic vision of Lucretius, natural philosophers such as Gassendi, Descartes, Galileo, Hobbes and Boyle all sought—albeit in differing ways—to explain the totality of natural phenomena in terms of matter, motion and natural laws. Regarding the nature of matter they followed they generally followed in the footsteps of Democritus and Lucretius: matter is composed of invisibly small impenetrable atoms, possessing only geometric properties such as shape and size. In so confining their explanatory resources these advocates of the new "mechanistic" or "corpuscularian" worldview were consciously rejecting key elements then-dominant Aristotelian system. In the new scheme of things appealing to animating forms or the doctrine of natural places was no longer an option.

Robert Boyle was a robust defender of the corpuscularian philosophy and the experimental method. In understanding how a lock or a clock functions, we need appeal to nothing more than the constituent parts they possess, the way these are fitted together, and the way they move. What applies to locks and clocks applies to all physical things: they are mechanical in nature. It was a mistake, argued Boyle, to attempt to base scientific theories solely on a priori metaphysical theories.

First comes the gathering of empirical data, and theories developed to explain the data should be put to experimental test. His own laboratory work on compressing gasses in tubes led to the discovery of what became known as Boyle's law for ideal gases (or $PV = k$), which states that the pressure of a gas tends to increase as the volume it is contained within decreases.

Prior to Newton, the most impressive and ambitious theory of the world in the corpuscularian tradition was due to Descartes. Although the latter is now best known for his purely philosophical works, during his lifetime he devoted the bulk of his intellectual efforts to mathematics and physics—and in the eyes of some he has as much claim to be the originator of modern physics as Newton.[3] All the essentials of Cartesian physics had been developed by the time Descartes completed *Le Monde* in 1633. When he learned of Galileo's troubles with the inquisition he decided to withdraw *Le Monde* from publication—perhaps wisely, since in it he reveals a commitment to the sun-centred Copernican view of the cosmos, and discards Aristotelian heliocentrism. Most of *Le Monde*'s doctrines resurfaced in Descartes' posthumously published *Principles of Philosophy* (1644).

These days Descartes is probably most famous for his mind-body dualism. He argued that by virtue of differing in their essential natures, mental and physical phenomena exist, in effect, in two entirely separate universes. One consequence of this dualism—presumably not a coincidence—is that the physical realm is entirely free from any lingering trace of mind, spirit or animating Aristotelian forms. Other proponents of the mechanical world-view were not so rigourous: Gassendi, for example, despite being an atomist also found it necessary to bring in something akin to Aristotelian forms to account for the differences between living and non-living matter. Descartes was determined to extend the mechanical model to matter in all its forms, and viewed—to the horror of some of his contemporaries—animals as mere machines.

For Descartes the essence of matter is simply spatial extension: "the extension in length, breadth, and depth which constitutes the space occupied by a body, is exactly the same as that which constitutes the body" (*Principles* II, 10). If matter simply *is* space, it makes no sense to suppose that one could remove all the matter from the inside of a bottle (say) and leave a region of empty space (or vacuum or void) behind. Like Aristotle before him, Descartes rejected the possibility of a true void. Descartes also followed Aristotle in holding that matter is in principle infinitely divisible. However, he also believed that matter often takes the form of relatively stable and long-lasting small particles, and it from these particles that macroscopic objects are constructed. These particles also come in different shapes and sizes: the smallest and faster-moving particles constitute fire and flames, the larger ones constitute larger bodies such as tables, chairs and planets.

[3] As one commentator puts it: "While nearly all of Descartes' physics is wrong in detail, his grand attempt is the beginning of theory in the modern sense" Truesdell (1984, 6).

If the universe is a fully-filled plenum is motion even possible? It is, provided the particles surrounding a mobile body are themselves free to move. Of course, as Descartes realized, the displaced particles will themselves need to be able to move, and this will only be possible if the particles in front of *them* are free to move as well, and so on ad infinitum. One way for this to occur—a way which opens the door to dynamic structures that are also comparatively stable—is for moving particles to form continuous circular and spherical matter-streams. Descartes developed a sophisticated cosmology on precisely this basis: he held that the universe consists of vast spinning vortices centred on stars. These vortices carry the planets in our solar system around our sun, and are also responsible for gravitational effects. The Earth's rotation generates a centripetal force directed away from the centre of the planet, which—if left unchecked—would hurl us off the Earth's surface. Fortunately for us the Earth is surrounded by a slowly rotating ethereal matter-field, and the downward pressure from this cancels the outwardly directed centripetal force.

Descartes' laws of motion were particularly influential on future physics. The first tells us that "each thing, as far as is in its power, always remains in the same state; and that consequently, when it is once moved, it always continues to move" (Pr II 37), while the second holds that "all movement is, of itself, along straight lines" (Pr II 3). While these laws may look familiar to contemporary eyes—not least because Newton's first law incorporates them both—to Descartes' contemporaries they were innovative. For Aristotelians, natural motion is along circular paths; this is why the planets stick to their orbits. For Descartes natural motion is straight-line motion. Natural philosophers had previously assumed that rest was more natural than motion, and that moving objects would come to a stop unless something keeps on pushing them. For Descartes motion and rest are equally natural properties; once an object is set in motion it will keep on moving in the same direction forever unless something stops it—after the initial push no further force is needed. Descartes third law describes what happens when material bodies come into contact: "a body, upon coming in contact with a stronger one, loses none of its motion; but that, upon coming in contact with a weaker one, it loses as much as it transfers to that weaker body" (Pr II 40). Descartes here anticipates later energy-conservation principles. The total *quantity of motion* in the universe is fixed, and is invariably preserved in collisions. For Descartes an object's "quantity of motion" is a function of its size (in the guise of volume) and speed. If, as he held, spatial extension and mass are identical, an objects mass is necessarily determined solely by its size.

Cartesian physics is remarkable in several respects. It is highly ambitious, aiming as it does to explain all physical phenomena in terms of a small number of basic principles. It is also highly economical in the resources it draws upon. By reducing the physical world to matter (construed as extension) and motion Descartes' cosmos is free from *forces*. He makes clear in the *Principles* that he does "not accept or desire any other principle in physics that in geometry or abstract mathematics, because all the phenomena of nature may be explained by their means, and a sure demonstration can be given of them."

Newton on Gravity

Although the Cartesian version of the corpuscular programme would continue to find adherents well into the 17th century, the history of physics was about to take a different turn, one that was significantly less hostile to natural forces.

Newton's *Philosophiae Naturalis Principia Mathematica* was first published in 1687, and immediately recognized as a monumental advance. Newton's mechanics is still in use today, as is the calculus—the mathematical innovation which Newton used to calculate rates of change. His theory of universal gravitation allowed him to predict the movements of planets and comets with unprecedented accuracy—as well as explaining why the orbits of planets have elliptical rather than circular orbits. He also made important contributions to optics. As Julian Barbour puts it: "So comprehensive was his genius, it appeared to open all doors into nature, to leave nothing really major to discover. Life after Newton seemed a mere walking through the garden into which his genius had directed us" (1989, 629). We will be focusing here on just one element in this garden—but for present purposes it is the most significant.

In its essentials, Newton's theory of gravity is simple to state: every object in the universe exerts an attractive force on every other object in the universe, no matter how distant. The precise magnitude of this force is directly proportional to the mass of the bodies and inversely proportional to the square of the distance between them—so the more massive the bodies the stronger the force of attraction pulling on them, and the farther apart the bodies are, the weaker the force. In formulating this theory Newton rejected Descartes' purely volumetric conception of mass. For Newton similarly sized objects can differ in mass by virtue of possessing different densities (or "quantities of matter").

Newton's gravitational influence operates instantaneously. Consequently, if the sun were suddenly to vanish (let's not inquire how or why), every object in the universe—no matter how distant—would immediately be affected: the gravitational pull that had hitherto been exerted by the sun would no longer be felt. For similar reasons, every time you raise (say) your right arm, *you* are causing an instantaneous change—very small, but nonetheless real and quantifiable—in every portion of matter in the most distant galaxies.

This is remarkable enough in itself. It can seem almost magical: if Newton's theory is correct, everything in the universe is invisibly (but not intangibly) connected to everything else. But the *kind* of connection we are dealing with here is also very distinctive. On the face of it at least, Newton's gravitational force looks to be acting directly across space, with no intervening or mediating factors. It has all the characteristics of what is known as an action at a distance force. Hence the problem: if proponents of the new mechanical world-view and their Aristotelian predecessors agreed on anything, it was that forces or connections of this kind have no role to play in legitimate science. Following in Aristotle's footsteps Aquinas encapsulated this position nicely: "matter cannot act where it is not". For Descartes, as we have seen, all motion is produced by contact. Hobbes agreed: "There can be no cause of motion except in a body contiguous and moved …" (*De Corpore*, 1655, ix para 7). As did Locke in the first three editions of his *Essay*: "How bodies operate on one another … is manifestly by impulse and nothing else. It being impossible to conceive that body should operate on *what it does not touch*" (1689, II, viii, 11).

Newton himself was well aware that his theory of universal gravitation was radical, and bound to prove controversial. His writings clearly reveal that he would much preferred to have found a mechanical explanation of some sort for gravity—one relying on an intervening aether in the manner of Descartes, for example. But despite much effort and many attempts he had been unable to find a viable model along these lines. Consequently, while endorsing the action at a distance model Newton opted to remain neutral on the mechanism (if any) underlying gravitational attraction. In Book 1 (§11) of the *Principia* he tells us:

> I have not as yet been able to deduce from phenomena the reason for these properties of gravity, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy.
>
> …. The impenetrability, mobility, and impetus of bodies and the laws of motion and law of gravity have been found by this method. And it is enough that gravity should really exist and should act according to the laws that we have set forth and should suffice for all the motions of the heavenly bodies and of our sea.

So far as the reception of the new theory among his contemporaries was concerned, Newton's fears were not misplaced: initially at least, many *did* find the notion of an action at a distance acting across any and all distances difficult to accept. Leibniz, who independently discovered the calculus at around the same time as Newton, fully recognised that Newton's theory was an impressive advance over Descartes'. Nonetheless, he firmly rejected action at a distance, remarking in a letter to Clarke: "That means of communication (says he) is invisible, intangible, not mechanical. He might as well have added, inexplicable, unintelligible, precarious, groundless and unexampled … 'Tis a chimerical thing, a scholastic occult quantity" (Alexander 1955, 162).

However, in the decades to come the hostility to action at a distance gradually faded. Although many mechanical models of a gravitational force which avoided appealing to action at a distance were proposed and investigated, they all proved inadequate. Consequently, it was not very long before most physicists accepted that the universe *was* in fact as Newton had reluctantly proposed: bound together by an invisible, all-penetrating force, acting both instantaneously and without intermediaries.

Dynamism

Leibniz may have been hostile to forces acting at a distance, but he was by no means hostile to forces per se. In fact, he was one of the leading 17th century advocates of the new *dynamic* conception of matter. During the early phases of the scientific revolution the corpuscularian mechanical theorists construed atoms in essentially the same way as Democritus and Lucretius: their only properties were size, shape and impenetrability. It is by virtue of being impenetrable that moving atoms bounce off one another after colliding, and for the atomists impenetrability was taken to be a primitive and unexplainable property. Leibniz found this wholly passive conception of matter problematic on a number of fronts.

For Descartes and Newton when a particle is moving inertially—i.e. when not subject to any external forces—it will continue to move in a straight line forever. The object's continuing to move is not grounded in any inherent capacity or power possessed by the object. For Leibniz even inertial motion should be viewed as essentially involving a force or power, a mode of activity whose manifestation is simply the object's *continuing on* moving. (In his later writings it becomes clear that this active power is what we now call *kinetic energy*.)

Leibniz also argued that collisions between classical atoms were profoundly problematic. These atoms were standardly construed as being totally rigid and incompressible. When one incompressible and inelastic atom strikes another, both will undergo an *instantaneous* change in direction. If, as Newton and Leibniz both believed, forces are proportional to accelerations then we immediately encounter a problem: an instantaneous acceleration is an infinite acceleration, requiring infinite forces. We can avoid this difficulty, suggested Leibniz, by construing atoms as point-like particles surrounded by short-range spheres of repulsive forces. When moving particles approach one another these repulsive forces *gradually* slows them down and the particles never actually come into contact. If we willing to acknowledge the existence of compressible repulsive forces inter-atomic collisions are no longer problematic.

In his 1699 *Confessions of Nature* Leibniz pointed out that orthodox atomists had a problem explaining how atoms manage to stick together to constitute compound objects such as table and chairs:

> … Democritus, Leucippus, Epicurus, and Lucretius of old, and their modern followers … asserted that the whole cause of cohesion in bodies may be interweaving of certain shapes such as hooks, crooks, rings projections, and, in short, all the curves and twists of hard bodies inserted into each other. But these interlocking instruments themselves must be hard and tenacious in order to do their work of holding together the parts of bodies. Whence this tenacity? Must we assume hooks on hooks to infinity?

The alternative dynamical solution is to hold that atoms possess both repulsive and *attractive* forces, operating at different strengths at different distances. A theory along these lines was elaborated in considerable detail by Roger Boscovich in his *Theory of Natural Philosophy, Reduced to the Single Law of the Forces existing in Nature* (1758). Boscovich proposed that a strong repulsive action at a distance force operated over very short distances whereas particles separated by very large distances particles were attracted by a force accurately described by Newton's law of gravity. He also held that additional attractive and repulsive forces operated at small scales—albeit at progressively different distances—and hoped to explain phenomena such as cohesion, evaporation and fermentation by appealing to them.

Since Boscovich envisaged these forces as inhering in spatial points—rather than material atoms of any kind—he, in effect, reduces the physical world to a dynamic spatially extended field of force.

Greatly impressed by Newton's achievement in accounting for gravity Kant showed none of Newton's own hesitation in accepting action at a distance—he unhesitatingly endorsed it throughout his career (Friedman 1992, p. 1). In publications spanning several decades Kant sought ways of accomodating Newton's innovations with his own evolving philosophical doctrines, and was ultimately led —in his *Metaphysical Foundations of Natural Science* (1786)—to adopt a position similar in some respects to that espoused by Boscovich. In his early *Thoughts on the True Estimation of Living Forces* (1747) Kant claims:

> There would be no space and no extension, if substances had not force whereby they act outside themselves. For without a force of this kind there is no connection, without this connection no order, and without this order no space.

In making material substances and forces central Kant is evidently working within a Newtonian framework, but Newton held that causally interacting physical objects exist within an all-embracing substantival space. In claiming that space is not foundational or primitive, but a product of the connections between objects generated by forces—which was his intent here—Kant is going well beyond Newton. Kant goes on to make the provocative and intriguing suggestion that force is responsible for the dimensionality of reality: "It is probable that the threefold dimension of space is due to the law according to which the forces in the substances act upon one another."

In his *Physical Monadology* (1756) Kant firmly rejects the passive matter of the corpuscularians. He claims that impenetrability is not a primitive inexplicable feature of matter, but essentially involves an active cause in the form of an action at distance repulsive force. He goes on to argue that a force of attraction must also exist between objects, for if it didn't the material contents of the universe would be dispersed to infinity by the action of the repulsive force. It is the interaction between these attractive and repulsive forces which determines the boundaries of material bodies.

These claims are reiterated and developed more fully in the later *Metaphysical Foundations*. His proposition 7 is a resounding endorsement of action at a distance: "The attraction essential to all matter is an immediate action through empty space of one matter upon another." Kant goes on to defend action at a distance forces from the objection "that matter cannot act *where it is not*". Far from being a contradiction this is a truism: *everything* that has an effect on something else is acting where it is not, and this includes a billiard ball that induces another ball to move by colliding with it. In his *Physical Monadology* Kant had taken the fundamental constituents of matter to be point-like material substances, surrounded by a sphere of repulsive force emanating from the material points. This picture is rejected in the *Foundations*. The defining characteristic of matter is impenetrability—a portion of matter *just is* an impenetrable region of space—and for Kant impenetrability is created by an expansive or repulsive force. Consequently, a region that is pervaded by a repulsive force is thereby pervaded by *matter* as well. In which case, the *Monadology* view, and its distinction between material points and force-filled regions of space is simply incoherent. On Kant new view matter is a continuum *every point of which* exerts an expansive force on its surroundings, or as Kant puts it: "… every part of it contains repulsive force, so as to counteract all the rest in all directions, and thus to repel them and to be repelled by them".[4]

Maxwell, Einstein and the Vindication of Locality

In advocating a wholly force-based account of matter, and taking the universe to be entirely pervaded by forces, Kant was in certain respects anticipating the "field theories" which would be developed in the 19th century, most notably by Faraday and Maxwell in their investigations into electricity and magnetism. However, as we shall see the fields advocated by Faraday and Maxwell differ in one key respect from those proposed by Kant. So far as the nature of forces, and the ways they propagate through the physical world, this difference will prove very significant.

In 1820 Oersted's observed that variations in the current flowing through a wire will cause a compass needle to alter its direction; his subsequent discovery that a wire carrying a current acts as a magnet confirmed that magnetism and electricity are closely connected. Although a number of scientists had suspected as much, Oersted's results triggered a period of increased interest in electromagnetic phenomena, and the most extensive and impressive of these investigations were carried out by Michael Faraday. The latter's *Experimental Researches in Electricity*, published in 1844, brought together many of his results, which included the discovery of induction, the fact that current can be generated in a wire by moving a magnet in the wire's proximity—the vast bulk of the world's electricity is now produced by generators working on precisely this principle. Faraday was fond of carrying out a simple experiment: if you spread iron filings over a sheet of paper, and place a magnet under the sheet, a pattern similar to the one depicted in Fig. 11.1 will result. The manner in which such patterns come into being inspired Faraday's conviction that electricity and magnetism were caused by stresses and strains in a space-pervading invisible aether, transmitted (very probably) at a finite speed.

In a diary entry in 1845 Faraday used for the first time the term "field" in this connection, but he had previously used formulations such as "lines of magnetic force" or "magnetic curves":

[4] Kant (1994, 503); for more on Kant's view of matter in the *Metaphysical Foundations* see Michael Friedman (1992, 2013).

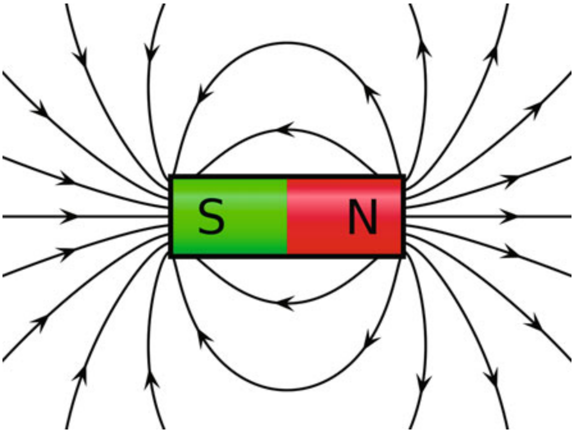


Fig. 11.1 Magnetic "lines of force" extending through the space surrounding a magnet

> I will now endeavour to consider what the influence is which paramagnetic and diamagnetic bodies, viewed as conductors, exert upon the lines of force in a magnetic field. Any portion of space traversed by lines of magnetic power may be taken as such a field, and there is probably no space without them. The condition of the field may vary in intensity of power from place to place, either along the lines or across them … and I have formerly described how this may, for a certain limited space, be produced.

If you look at a magnet surrounded by empty space you will not see anything resembling these lines of force emerging from it, but for Faraday they were nonetheless present, as real and *powerful* physical phenomena in their own right.

The task of devising a mathematical framework capable of accomodating Faraday's field conception and the diverse experimental results concerning electromagnetic phenomena that had by now accumulated fell to James Clerk Maxwell, who succeeded brilliantly. The essentials of Maxwell's comprehensive new theory of electromagnetism were presented in a series of papers which appeared between 1661 and 1665. One particular discovery of Maxwell's stands out. Maxwell's equations captured the manner in which changing magnetic fields give rise to changing electrical fields in their vicinity, and vice versa. Maxwell realized that this mode of interaction would give rise to a self-sustaining and self-propagating wave phenomenon in the electromagnetic field. By drawing on already known results concerning the basic properties of electricity and magnetism Maxwell was able to calculate from first principles the expected velocity of this wave: it turned out to coincide almost exactly with current estimates of the speed of light in a vacuum. Maxwell did not shy from drawing the obvious but nonetheless remarkable conclusion: light *is* a form of electromagnetic radiation. Although Maxwell appreciated that it was likely that only a small part of the electromagnetic spectrum would be constituted by visible light in his *Treatise* he provided no indications as to how go generate higher and lower frequency waves. It was not long before other scientists were attempting to do just this, and Hertz became the first person to transmit and receive radio waves in a series of experiments conducted between 1886 and 1889. The rest is history.[5]

At first glance the electromagnetic forces introduced by Faraday and Maxwell may appear to be similar to Newton's gravitational force: both are invisible, both extend through seemingly empty space. However, they are in fact profoundly different in character: whereas a force as envisioned by Newton and Kant directly connects spatially distant objects, electromagnetic forces always act *locally*: they have to pass *through* the regions of space separating objects they influence. According to Faraday and Maxwell, a magnet creates a pattern among iron filings because it generates a spatially continuous field which unfolds at a finite speed through nearby space—it may operate over a distance, but it does not act at a distance. As this passage from the preface to Maxwell's *Treatise* makes clear, they were well aware of their divergence from the Newtonian conception of gravitational force:

> Faraday in his mind's eye saw lines of force traversing all space where the mathematicians saw centres of force attracting at a distance: Faraday saw a medium where they saw nothing but distance: Faraday sought the seat of the phenomena in real actions going on in the medium, they were satisfied that they had found it in a power of action at a distance impressed on the electric fluids. (1954, Vol 1, p. ix)[6]

Irrespective of its other merits, the new theory of electromagnetism was not vulnerable to the criticism that it relies on forces of an occult or magical kind—the criticisms levelled at Newton's action at a distance gravitational theory when it first appeared.

Although, not surprisingly, Maxwell's account of light was soon widely accepted by physicists, it also gave rise to a serious difficulty. One of the foundation stones of classical physics is that the laws of nature are blind to uniform straight line velocities. Experiments conducted in a laboratory on a moving train will produce exactly the same results as the same experiments conducted in a stationary laboratory; as Galileo and Newton realized, this is the reason why the Earth's motion around the sun is not obvious to those of us confined to the surface of the planet. If Maxwell was right, and the speed of light is a consequence of basic physical laws, then anyone measuring the speed of light-beam should always get the same result—299,792 km/s—no matter what their own state of motion is. But this too seems

[5] As Richard Feynman put it in the second volume of his *Lectures on Physics*: "From a long view of the history of mankind—seen from, say, ten thousand years from now—there can be little doubt that the most significant event of the 19th century will be judged as Maxwell's discovery of the laws of electrodynamics. The American Civil War will pale into insignificance in comparison with this important scientific event of the same decade."

[6] By "the mathematicians" Maxwell is referring here to theorists in Germany and France, such Weber, Gauss and Ampere who construed electrical and magnetic forces in a Newtonian action at a distance fashion. Maxwell returned to this theme in the concluding paragraph of his *Treatise*, where he observes "In fact, whenever energy is transmitted from one body to another in time, there must be a medium or substance in which the energy exists after it leaves one body and before it reaches the other …" (1954 Vol. II, 493).

bizarre, both by the standards of common sense and classical physics. According to the latter, if a beam is measured as travelling at 299,792 km/s by a scientists who is stationary (with respect to the Earth), then a scientist on a train who measures the speed of the same beam but who is travelling at 50,000 km/s in the opposite direction to the beam should find that the latter is travelling at 349,000 km/s.

It seems something has to go: either Newton's classical mechanics doesn't apply to light (and other forms of electromagnetic radiation) or the speed of light cannot be a basic physical constant.

A compelling solution to this conundrum was put forward in 1905 by Einstein, in the guise of his Special Theory of Relativity (STR). According to the latter, the speed of light *is* a basic physical constant, and has the same value for all observers, irrespective of their state of motion. To make sense of this, Einstein proposed that observers moving relative to one another will measure time and space differently, e.g. if you are moving relative to me then time (as measured by clocks and your body) will pass more slowly, events which are simultaneous for me will *not* be simultaneous for you. In more general terms, subjects who are moving at a constant speed relative to one another will possess their own "frames of reference"; each of these frames of reference will divide spatial and temporal intervals differently, and —crucially—all these frames of reference are equally valid. So from my perspective two events E1 and E2 might be simultaneous, but from yours these same to events will *not* be simultaneous, and both perspectives are equally legitimate.

It didn't take long for the full metaphysical implications of STR to emerge. In September 1908 Hermann Minkowski—one of Einstein's maths teachers at the Zurich polytechnic—began his talk to an assembly of German mathematicians and scientists thus: "The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a union of the two will preserve and independent reality" (1954, 75–91). The union of the two that Minkowski went onto propose took the form of a four-dimensional spacetime continuum. Within this continuum there is no privileged universe-wide present, and all spatio-temporal locations are fully and equally real—those lying in the future included. Persisting objects—just as lumps of rock or human bodies—are themselves four-dimensional objects, existing as worldlines (or collections of such) embedded in the four-dimensional spacetime continuum. It is only because all times are real that different inertial reference frames can generate different but equally valid ways of dividing events up between past, present and future. What would soon become known as "Minkowski spacetime" would also became the standard way of interpreting STR in physics.

As Einstein was well aware, the relativization of simultaneity does not sit easily with Newton's account of gravity. For Newton, as we have seen, gravity is an action at a distance force that directly connect every object in the universe. Moreover it is a force which operates *instantaneously*, there being no delay between gravitational causes and effects. If simultaneity is relative in the way Einstein proposed, then events which are simultaneous—and so related by Newton's gravitational forces—in one frame of reference will not be simultaneous in others. Clearly, a new theory of gravity was required, one which did not require instantaneous interactions.

It took Einstein a decade of hard work to devise a new account of gravity, in the guise of his General Theory of Relativity (GTR), which made its first appearance in 1915. Einstein's key move was radical: he solved the problem posed by Newton's gravitational force by abolishing it entirely. According to GTR gravity is not a *force* at all: material objects under the influence of gravity do not attract one another. Instead, a massive body such as the sun or a planet creates a spatiotemporal distortion in its vicinity. For a useful (if only partial) analogy think of the way in which a heavy iron ball will produce a curved region in a previously flat rubber sheet or mattress on which it is placed. In a similar fashion, a massive body will induce curvature in the surrounding region of four-dimensional spacetime, an effect that lessens with distance—just as with Newton's gravitational force. In the absence of significant mass a region of spacetime will be entirely flat—just like a mattress.

According to Einstein, the gravitational effects that were previously attributed to the effects of a force are the products spacetime curvature. In GTR the principle of inertial motion advocated by Descartes and Newton is fully retained: objects that are not subject to any external forces will continue to move in a straight line, at the same speed, forever. However, when we are dealing with curved spaces what counts as a "straight line" is not as straightforward as is the case in flat space. In a flat space a straight line in the familiar Euclidean sense—a line with no bends or curves—will also be the shortest distance between two points it connects. In many curved spaces there are no straight lines in the Euclidean sense at all. If for example we take as an example of a curved space the surface of a sphere, then all the lines in such a space will be curved. Even so, in such a space it remains the case that for any two spatially separated points some connecting lines will be longer than others—e.g. a line stretching straight down from the north pole to a point on the equator will be shorter than a windy "S" shaped line between the same two points. There are also paths of shortest distance in four-dimensional spacetime—though inevitably these are harder to visualize—and according to GTR objects that are falling freely (i.e. which are not subject to any forces) will follow these paths of shortest distance. This is precisely what the planets are doing when they orbit the sun—and similarly for an apple that falls from a tree.

Although Einstein's GTR and the Newton's theory of gravity make very nearly the same predictions in most ordinary circumstances, there are some divergences, and in all such cases Einstein has invariably triumphed over Newton. An early instance was Einstein's prediction that starlight travelling towards the Earth should be deflected by 1.75 arc seconds due to the spacetime curvature created by the sun —a tiny but still measurable amount—a prediction which made the headlines when it was experimentally confirmed by Eddington in 1919. More recently, in 2016 the discovery by LIGO (the Laser Interferometer Gravitational-Wave Observatory) of the existence of gravitational waves—predicted by GTR but hitherto unobserved—also made headlines around the world. Gravitational waves are ripples in the fabric spacetime; those detected by LIGO are thought to originate in the collision between two massive black holes—enigmatic entities whose existence was also predicted by GTR.[7]

Time, Dimensions and Causes

In his *Treatise of Human Nature* (1739) and *Enquiries Concerning Human Understanding* (1748) the philosopher David Hume set out to undermine our common sense beliefs concerning the nature of causation. When we see a moving pool ball strike a stationary one, and the stationary one moves off—perhaps going into a pocket, perhaps not—we are naturally inclined to think that the first ball *made* the second move. Causal interactions such as these are not just a matter of one event being followed by a second, they involve a kind of necessitation: given the first event, the second *had* to happen. As Hume realized, the idea that causation involves necessitation naturally extends to the way we think of natural laws—indeed it largely explains why we talk of "laws" at all. In Newtonian mechanics, for example, it's natural to think that objects fall under the influence of gravity because they *are made to*—by the attractive force of gravity. The laws of nature don't just reflect regularities in how objects behave and interact, they *govern* the movements of objects.

This way of thinking may come very naturally to us, but it is unjustified—or so Hume argued. Think again of what precisely you see when you watch two pool balls collide. Do you really see one ball *making* the other ball move? Or merely one ball moving until it comes into contact with the other, and the other ball moving off? Surely only the latter, Hume urged—and the same applies for all the causal interactions we observe. We are inclined to think the first ball *makes* the second move only because we have perceived lots of similar interactions in the past. In such situations the second ball always moves away when hit by the first, and so in the current case we *expect* the second ball to move—and this expectation is the source of our conviction that the ball in question *has* to move when struck. When we combine this analysis of why we tend to think causation involves necessitation with the fact that we never actually observe any necessitation, we should conclude—or so Hume argued—that causal necessitation does not actually exist in the world, it is simply projected into the world by us. All that exists in the world are certain patterns of events—regular successions, as Humeans call them—and to the

[7] See Dainton (2010) for a more detailed introduction to Einstein's relativity theories.

extent that laws of nature exist in the world they consist of nothing more than these regular successions.[8]

Quite what stance Hume really adopted viz a vis causation remains controversial, but the Humean doctrine that no trace of natural necessitation is to be found in nature is an influential one in contemporary metaphysics. Indeed it enjoys a good deal more popularity now than it did in the 18th and 19th centuries—it was not for nothing that Hume complained of his *Treatise* falling "dead-born from the press, without reaching such distinction as even to excite a murmur among the zealots".

In comprehending why so many of Hume's contemporaries found his causal scepticism difficult to take seriously it is illuminating to consider an imaginary game or pastime. You have in front of you a photograph of Leonardo's *Mona Lisa*. Your task is to construct a metre square representation of this famous work using nothing more than 1 cm wide coloured toy building blocks, one row at a time, from the bottom up and from right to left. This is by no means an impossible task, provided you have enough bricks in the appropriate colours—and happily this is the case, you have more than enough bricks for the task at your disposal. There is however a twist: the rules of the game are quite specific when it comes to how you are to go about choosing which blocks to use. Each successive row of your construction will be composed of 100 blocks, and these have to be selected at random from a container containing tens of thousands of variously coloured blocks. To make matters still worse, once a block is placed in the frame destined to house your picture it is not permitted to remove and replace it with another block; its location is permanent. Needless to say, as you embark on your task you are not optimistic of success: your chances of replicating the *Mona Lisa* by this method are astronomically small.

In Hume's period—as in most others prior to our own—it was universally accepted that time differs from space because time *passes* or *flows* whereas space does not. What does the passage or flow of time involve? It can be characterized in a variety of ways, but there are two key ingredients. First, there is the claim that the present time is metaphysically privileged: perhaps only present events are real, perhaps they are real in a way that past events are not. The second thesis is the seemingly self-evident truism that presence is transitory: what is happening *now* will soon not be happening now because the events in question will soon sink into the past.

For anyone who thinks about time in this common sense sort of way, it will be natural to assume the cosmos comes into existence only gradually, in a succession

[8] In his analyses of Newton's mechanics in *De Motu* (1721) and *Siris* (1744) George Berkeley argues along similar lines to Hume: "Those who assert that active force, action, and the principle of motion are really in the bodies, maintain a doctrine that is based upon no experience, and support it by obscure and general terms, and do not themselves understand what they wish to say" (*De Motu*, §31). In his *Treatise* (§32) Berkeley observes that "When we perceive certain ideas of sense constantly followed by other ideas, and we know that his is not of our own doing, we forthwith attribute power and agency to the ideas themselves"—the relevant "ideas" here are (presumably) the objects of immediate perception.

of momentary universe-wide phases or layers, with each newly created present phase giving way to another as time passes. The process as we are now envisaging it is more fine-grained than the *Mona Lisa* game's, but metaphysically it is analogous. And precisely the same potential problem arises. As we have just seen in the imaginary *Mona Lisa* case, in the absence of tight constraints on the elements chosen for each successive line of blocks, the result will almost certainly be total anarchy: a picture without recognizable forms or patterns. The same applies in the case of the real universe. If *it* were to come into being in a succession of phases or layers, in the absence of tight constraints on the contents of new layers the odds are astronomically high that the result will be utter chaos. Since our world is not chaotic —at least in the extreme sense that is relevant here—we have no reasonable option but to conclude that the process of phase-creation is a tightly constrained one.

It also seems reasonable to conclude that it was considerations along these lines which—in part at least—made it difficult for Hume's contemporaries to take his causal scepticism seriously. In this period the idea that time flows was not seriously questioned. Newton, for example, in the *Principia's* Scholium writes: "Absolute, true and mathematical time, in and of itself and of its own nature, flows uniformly and by another name is called duration."[9] True, in this period many would have followed Descartes in supposing that an all-powerful and benevolent God is directly responsible for re-creating the world instantaneously from moment to moment, which makes the orderliness of the universe a product of divine choice. But the increasing numbers of philosophers and scientists in the 18th and 19th centuries who were reluctant to grant God any overt role in their theories an alternative source of natural order had to be found. A very natural alternative—almost unavoidable in the circumstances—is to take the required constraints to be located *in material world itself*, whether in the guise of universe-wide natural laws to which all physical processes conform, or inherent causal powers that reside in and determine the behaviour of material things.[10]

These days, as we saw in the previous section, thanks to Einstein's relativity theories the majority of physicists assume that our universe takes the form of a four-dimensional spacetime. In such a universe there is no ontological difference between past, present and future: all objects and events are equally real, there is no temporal passage and no privileged present. As a consequence such a universe cannot *come into being* in a succession of momentary phases, in the way that Descartes believed. If the universe comes into being at all—as opposed to existing eternally, an issue which remains unresolved in contemporary cosmology—it can only do so *as a whole*, with past, present and future all being created together.

When we conceive of the universe in this four-dimensional manner the need to explain why chaos is avoided as new slices of reality enter existence simply

[9] Newton was by no means alone. For example in the first *Critique* Kant observes that "space alone is determined as permanent, but time, and thus everything in inner sense, continually flows" (B291).

[10] For some contemporary arguments along these lines see Foster (1982) and Strawson (1982).

vanishes: on the four-dimensional view there are no new slices of reality being created moment-by-moment. If future happenings are already fully real, it makes no sense to suppose that causes bring their effects into being—causes and their effects are both (timelessly) parts of the four-dimensional manifold of events. Holding that law-like regularities are underpinned by a necessary connection of some kind may still be an option, but since positing such a connection lacks any real explanatory value it looks to be redundant. As a consequence a powerful consideration which undermined the case for the Humean regularity view of causation itself vanishes.[11]

If this is not the entire rationale for why the Humean view is taken more seriously than was the case pre-Einstein, it may well be a significant part of it.

Quantum Theory

Quantum theory, currently our best theory of the micro-realm, emerged only gradually in the first three decades of the 20th century. The theory defies easy summary, and remains mired in controversy: there is a still-expanding number of "interpretations" of the theory, each providing very different accounts of what quantum mechanics truly implies about the nature of physical reality. We will focusing here on some of the more obvious implications concerning the nature of physical interactions and causation.

The development of quantum theory was initially triggered by a cluster of puzzling discoveries concerning the behaviour of light and other forms of radiation, and the structure and composition of atoms. The first step took place in 1900 when Planck solved baffling puzzle concerning so-called "black-body" radiation by positing that energy-levels did not form a continuum—as generally assumed hitherto—but rather came in multiples of a very (very) small unit, or *quantum*. In 1905 Einstein successfully resolved a puzzle concerning the photoelectric effect by arguing that rays of light are composed of discrete quanta as well—the particles which would soon be called *photons*. But while the considerations advanced by Einstein for taking light to be composed of particles were very plausible, there remained powerful reasons for supposing that light must also have a *wave*-like nature. Even before the advent of Maxwell's theory, the two-slit experiment devised by Thomas Young in 1801 showed that light-rays produce interference patterns very similar to those produced by water waves—see Fig. 11.2.

This was all very baffling: how can anything be both a wave and a particle? It was not until the 1920s, with the breakthroughs of Heisenberg and Schrödinger, that the new quantum mechanics was put on a solid mathematical footing. The equation proposed by Schrödinger doesn't (directly) tell us how a particle—an

[11] In his more recent (2012, 5–6), while Strawson acknowledges that adopting a four-dimensional conception of spacetime requires a re-conceptualization of causation and natural laws, he argues that natural necessity—of a sort—does still have a role to play in the new temporal context.

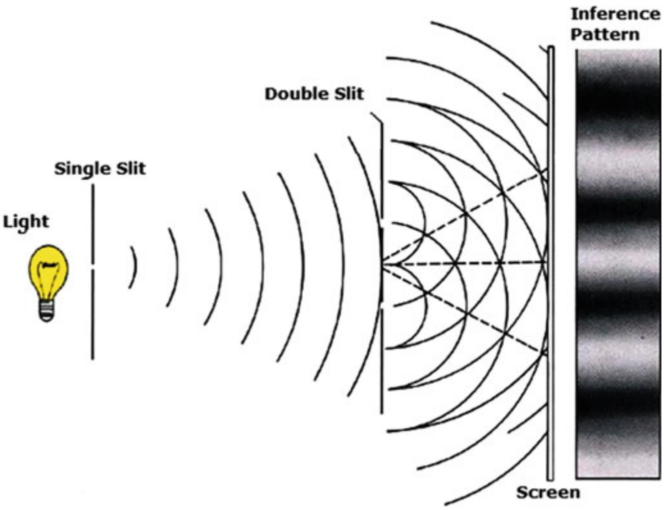


Fig. 11.2 Thomas Young's double slit experiment

electron, say—behaves, but rather how a wave evolves over time. Schrödinger's waves are unlike water or sound waves by virtue of extending through all of space, but just as with classical waves they associate certain numerical values with specific spatial locations. These numbers don't tell us anything definite; they supply only the *probability* that a particle will be detected at a particular location, or possess a certain momentum—if we happen to measure it. Prior to a measurement of position or momentum all the different possibilities and probabilities exist in a "superposition" and the particle does not have a definite position or momentum at all. The measurement process is said to *collapse* the wave function. Or at least this is how things stand on the orthodox (or "Copenhagen") interpretation of quantum mechanics, which is still to be found in textbooks—though there are alternatives, as we shall see shortly.[12]

Earlier classical physics theories, most notably Newtonian mechanics, were entirely deterministic. In principle, if you were supplied with accurate data concerning the locations, masses velocities of all the particles in the universe you could

[12] The wave function in quantum mechanics is fundamentally different in nature to the space-pervading waves found in classical theories such as Maxwell's electromagnetism. The wave function for a physical system exist in an abstract mathematical "Hilbert" space, which possesses 3N dimensions, where "N" is the number of particles in the system—since there are billions of atoms in a drop of water, the dimensionality of these Hilbert spaces will typically be very large indeed. If quantum mechanics provides a complete and correct account of physical reality at its most fundamental level, then if the wave function is the most basic ingredient in quantum theory, shouldn't we conclude that our universe in fact has 3N dimensions, where "N" stands for the number of particles in the universe? So called "wave function realists" argue for precisely this conclusion—for more on this debate see Ney and Albert (OUP 2013).

use Newton's equations of motion to predict precisely how the universe would evolve from that point in time till any point in the future. Since according to quantum theory a particle's wave function provides us with an exhaustive account of its physical properties, if the theory is true we have no option but to accept that at the atomic level reality is inherently probabilistic and indeterministic. Even if you knew everything there is to know about the distribution and motions of particles throughout the universe at one moment in time you would not be able to predict precisely what is going to happen over the next few seconds.

Empirical studies of the ways atomic- and sub-atomic particles such as electrons, photons, protons, neutrons behave all suggest that reality *is* indeterministic in precisely the way quantum mechanics predicts. If in a series of experiments high-energy protons are fired into the nuclei of a succession of wholly indistinguishable hydrogen atoms there is no unique outcome of these collisions, but rather a number of different outcomes, occurring with just the frequencies predicted by quantum mechanics.

If the micro-world is as indeterministic as it appears to be, there are obvious implications for our understanding of causation. So far as fundamental physical particle interactions are concerned, when an event E1 causes E2, it will never be the case that E1 *makes* E2 happen in the strong sense of "given that E1 occurred E2 *had* to happen, it was *necessary* that E2 occurred as well." It seems that causation at this level—if we assume it still exists in any form at all—must be viewed as inherently probabilistic rather than deterministic. For anyone accustomed to thinking of the world in a deterministic way, this will be a revisionary step.[13] However, on reflection is can easily seem to be a very natural one: smoking may do no more than *raise the probability* of one's getting cancer, but it can still count as *causing* cancer—or so most of us are prepared to accept. Partly in response to developments in quantum mechanics philosophers have developed a number of different probabilistic accounts of causation. One option, for example, is to holds that we can still regard particles as possessing causal powers, but these powers take the form of *probabilistic dispositions* to behave in certain ways in certain circumstances.[14]

It is by no means the case that all physicists are happy with the indeterministic world bequeathed to us by quantum mechanics as standardly construed—a disquiet Einstein famously expressed by claiming that God "does not play dice with the universe"[15] Given this, it is not surprising to find that alternative ways of making sense of the basic mathematical framework quantum mechanics have long been sought. Although some of these alternatives do restore determinacy after a fashion, they do so in ways which bring their own costs.

[13] Hume appears to be in this category, given that in section VII of his *Enquiry* he offered this by way of a characterization of causation: "We may define a cause to be *an object*, *followed by another*, *and where all the objects similar to the first are followed by objects similar to the second*."

[14] See Popper (1990); for a survey of different approaches to probabilistic causation see Hitchcock (2010).

[15] Einstein made the remark in a letter to Max Born in 1926.

One of the more influential of the alternatives is the “de Broglie-Bohm approach”, first advanced by de Broglie in 1927, then later re-discovered and elaborated by Bohm in the early 1950s. According to this view, the standard form of quantum mechanics is incomplete: in addition to the wave function there is a quantum potential which acts as a “pilot wave” guiding particles along their trajectories. As a consequence, particles *always* have a quite definite location and velocity—something that is very much not the case under the Copenhagen interpretation. Since changes in one part of a physical system can instantaneously induce changes in the system’s entire pilot wave, which in turn affects how particles will move, the de Broglie-Bohm version of quantum theory is decidedly non-local: a system’s pilot wave might easily extend through very large regions of space, or even the whole universe. It is also important to note that the theory makes precisely the same empirical predictions as orthodox quantum theory—given the latter’s empirical success if it didn’t the de Broglie-Bohm approach would be unviable. Consequently, there remains a sense in which the behaviour of individual particles in a given context remains inherently probabilistic.

If interest in the de Broglie-Bohm approach has been on the rise in recent years, interest the *many-worlds* interpretation—based on Everett’s work in 1957—has soared, partly but not wholly because it is currently favoured is cosmological circles. According to the many worlds theorists there is no collapse of the wave function when a particle is detected by a piece measuring apparatus. Rather, *all* the many potential trajectories which have a finite probability in the particle’s wave function are in fact realized, albeit in different worlds (or sub-worlds) which branch off from this one. Although the many-worlds view certainly solves the problem of explaining how a piece of measuring equipment *can* provoke the collapse of a wave function, the ontologically profligate manner in which it does so renders it implausible in many people’s eyes. Even if we are prepared to overlook that issue, the many-worlds view restores determinacy in a novel (and disturbing) fashion: no possible outcome of a physical interaction *fails* to be realized.[16]

Quantum Strangeness

The two slit phenomenon provides a striking manifestation of the sheer *weirdness* of the realm of the quantum. In this experimental setup a source is able to emit particles—electrons, let’s suppose—either singly, or in great numbers *en masse*. The source is aimed at a detector screen, and whenever an individual electron strikes the screen it registers as a small but visible white dot. In between the source and the screen there is a metal barrier with two narrow vertical slits, which can be opened or closed independently by the experimenter. If both slits are open and

[16] For more on the many worlds interpretation see Vaidman (2014). Lewis (2016) provides accessible introductions to several of the leading alternatives to the Copenhagen interpretation.

electrons from the source arrive at the two slits *en masse*, almost immediately an interference pattern—in the guise of alternating illuminated stripes each consisting of many white dots—will appear on the screen, in the manner depicted in Fig. 11.2.

Remarkably, if the settings for the source are changed, and electrons are emitted only one by one, an interference pattern is still created on the screen, it simply takes longer to appear since the electrons are now arriving singly rather than in large numbers. If, however, the experimenters makes another adjustment to the settings and closes one of the slits leaving the other open, a quite different pattern emerges on the screen. Under these conditions *no* interference pattern is created; instead the electrons create a circular cluster-patter on the region of the screen behind the open hole. The same result occurs if an experimenter places a detector at one or both of the slits, with a view to finding out which slit an electron is passing through.

Put yourself in the position of a particle that has only just been emitted by the source. On average, the trajectory that you will take towards the screen will differ depending on whether one or two slits are open in the intervening barrier. But how at this point—before you have even begun your journey towards the latter—do you know how many slits are open? How do you know whether or not there is a detector at one of the slits? The interference patterns formed by water or sound waves are a straightforward consequence of the combined interactions which take place between myriad simultaneously existing particles. Such a process obviously cannot explain the interference pattern which gradually builds up when electrons are emitted one by one—so what does explain this effect?

Quantum mechanics can provide answers. An electron's trajectory is controlled by the wave function for the entire system, and the system's wave function when one slit is open is quite different from the wave function that exists when both slits are open. In the latter case parts of the wave function pass through both slits and the resulting ripples of probability interfere with one another. It is this interference structure in the wave function which is responsible for the interference pattern generated by electrons striking the screen—it is not difficult to see how this comes about since it is the wave function determines the probability of particles appearing at different locations on the screen.

On the Copenhagen interpretation, the electrons have no definite position from the time they are emitted from the source till the time they strike the screen. In contrast, for proponents of the de Broglie-Bohm approach the electrons always have a definite position throughout their journey, even if we only discover their location when they hit the screen; when both slits are open there is also a fact of the matter concerning which slit each electron passes through—even when we are not making any attempt to detect. On this view it is the guiding pilot wave of an electron that passes through both slits and is responsible for the creation of an interference pattern on the screen.[17]

[17] What of the many worlds interpretation? On one view—see Deutsch (1997)—each of the different potential electron trajectories contained within the wave function correspond to actual outcomes in different worlds, and the interference pattern exists because of the ways the electrons in different worlds interact with one another.

We saw earlier that in advocating a dynamic conception of matter Leibniz mocked the ancient atomists and their followers in the mechanical tradition for holding that the "whole cause of cohesion in bodies may be interweaving of certain shapes such as hooks, crooks, rings projections and, in short, all the curves and twists of hard bodies inserted into each other." If the competing interpretations of the two slit experiment clearly demonstrate anything it is that Leibniz was right: interactions in the micro-realm are governed by mechanisms that are quite unlike anything dreamt of by the ancient atomists. Equally, they also go far beyond anything dreamt of by dynamists such as Newton and Kant.

There is a further implication of quantum theory that is very relevant so far as the nature of physical interactions is concerned: it is now widely agreed that the theory is fundamentally and irreducibly *non-local*. In this context a theory is *local* if it rules out action at a distance influences of any kind. In practical terms, for theories of the local sort if an event E1 exerts an influence on event E2 some distance away, then the effect of E1 will invariably be mediated by a process which passes through the intervening space—whether it be in the manner of a bullet moving from gun to target, or ripples crossing a pond. Since according to Einstein's special theory of relativity nothing can travel faster than the speed of light, it is natural to assume that all transmissions or influences between spatially separated events must occur at either light-speed or sub-light speed. This locality constraint is difficult to square with the much-discussed phenomenon of quantum entanglement. Present purposes will be served by a simplified schematic outline of this subtle effect.

Electrons have a quantum property known as "spin", a form of angular momentum (which, confusingly, does not involve electrons actually rotating). Spin can exist in any spatial orientation, but for present purposes we can restrict our attention to just two of these, which we can label *spin-up* and *spin-down*. Quantum mechanics tells us that it is possible for two electrons to interact in such a way that their spins are thereafter correlated—or "entangled"—in a distinctive way, at least until one or other of them interacts with something else.

Viewing matters from the perspective of the Copenhagen interpretation, when a pair of entangled electrons X and Y comes into being each of them has a 50% chance of being spin-up or spin-down, and their spin-states exist in a superposition until one or other of them comes into contact with a suitable detector. As a consequence, prior to a measurement being taken neither electron possesses a determinate spin. However, if at some point in time electron X encounters a suitable detector and is found to have spin-up, then a measurement conducted on electron Y a moment later will find that it has spin-down; if on the other had X turns out to have spin-down, then Y will be measured as having spin-up. Measuring X's spin results in an *instantaneous* collapse of the wave-function that had hitherto encompassed both particles, and this collapse is such that Y is guaranteed to have an opposite spin-orientation to X. Entangled particle-pairs are connected in this sort of way irrespective of how far apart they happen to be.

More generally, Ismael and Schaffer provide this usefully succinct characterization of the phenomenon: "The components of a system in an entangled state behave in ways that are individually unpredictable, but jointly constrained so that it is possible to forecast with certainty how one component will behave, given information about the measurements carried out on the other(s)" (2016). It was Schrödinger who first wrote of particles related in this way as *entangled*, and he found it problematic: "Measurements on separated systems cannot directly influence one another—that would be magic" (1935, 16). Einstein famously characterized this mode of interaction as "spooky action at a distance" and he too was less than happy with it.[18] He thought it likely that the relevant phenomena could be explained by a purely local theory, but never succeeded in finding one. More importantly, since the 1980s there has been a succession of increasingly sophisticated experiments that all point in one direction: to quantum entanglement's being a real physical phenomenon.[19] For Raymer this outcome "is a highly curious even shocking result. It brings home the truly revolutionary nature of quantum physics" (2017, 139).

So far as Einstein is concerned, it is perfectly understandable why he was far from welcoming with regards to quantum non-locality. It certainly does not sit easily with his special theory of relativity's ban on faster than light causal transmission. More significantly, with his general theory of relativity Einstein had successfully eliminating Newton's action at a distance gravitational force, and explained gravitational effects in terms of purely local fields. By so doing Einstein vindicated—or so it initially seemed—one of the main tenets of both the ancient atomists and the scientific revolution's mechanical theorists: the long-influential conviction that the only way things can only influence one another is by touching one another. Einstein was fully aware of the significance of such an achievement.

Adopting a longer historical perspective sheds a different light on these developments. During the centuries-long reign of Newton's theory of gravity the majority of physicists had no trouble at all in accepting that the workings of the universe were governed by an action at a distance force, and nor did leading philosophers, most notably Kant. Since only a decade or so separates the arrival of Einstein's general theory of relativity—and the ensuing demise of Newtonian gravity—from the advent of quantum mechanics and entanglement, the undisputed reign of locality in modern physics was really rather brief.[20]

[18] The two particle form of entanglement was introduced by Einstein et al. (1935) paper "Can Quantum Mechanical Description of Physical Reality be Considered Complete?", but non-locality had worried Einstein for longer. As Cramer (2016, §6.2) relates, in the 1927 Solvay conference Einstein introduced his "bubble paradox". On the orthodox view, there are circumstances in which a photon's wave function will take the form of an expanding sphere; the sphere will continue to expand until there an interaction with another particle, at which time the entire wave function instantaneously vanishes. Einstein asked how the parts of the wave function at some—potentially considerable—distance away from the detection even "know" they should disappear at precisely this instant?

[19] Particularly relevant here, since they close-off various loopholes in previous tests, are the recent results reported in Hensen et al. (2015) and Giustina et al. (2015).

[20] For helpful and encouraging comments on earlier drafts my thanks to Galen Strawson and Shyam Iyengar.

Appendix: The Standard Model

Our current best theory of matter is known as the *Standard Model of Particle Physics*, which takes the form of a quantum field theory (QFT). This field theory originated in work done on quantization of the electromagnetic field in 1926–1927 by Born, Heisenberg, Jordan and Dirac, and was gradually extended to cover other forces and fields over the next half century or so. The Standard Model received a noteworthy—and much publicized—confirmation when the Higgs boson was discovered at CERN in 2013.

According to the Standard Model all material things are composed of three families of particles: quarks, leptons (e.g. electrons and neutrinos) and force carrying bosons (such as electrons and muons). Hadrons are particles made up of multiple quarks: the *baryons* have three quark constituents—e.g. the protons and neutrons familiar from chemistry fall into this category, whereas the generally short-lived *mesons*—such as the pion—are composed of just two quarks. QFTs are so-called because their fundamental ingredients are entities known as *quantum fields*, and particles tend to be viewed as nothing more than patterns of activity within these fields—with different species of particle being associated with different types of quantum field. From the perspective of QFT the universe consists of a number of different overlapping quantum fields each of which extends through all of space.

The Standard Model provides an account of three of the known four forces in nature. These are the *strong force* which binds the quarks, the *weak force* responsible for the transformation of massive quarks and leptons to lighter particles, and the more familiar *electromagnetic force*, which has a potentially infinite range. As for the force-carriers, here is what CERNs introductory guide to the Standard Model has to say:

> Three of the fundamental forces result from the exchange of force-carrier particles, which belong to a broader group called “bosons”. Particles of matter transfer discrete amounts of energy by exchanging bosons with each other. Each fundamental force has its own corresponding boson—the strong force is carried by the “gluon”, the electromagnetic force is carried by the “photon”, and the “W and Z bosons” are responsible for the weak force. Although not yet found, the “graviton” should be the corresponding force-carrying particle of gravity. The Standard Model includes the electromagnetic, strong and weak forces and all their carrier particles, and explains well how these forces act on all of the matter particles.[21]

In some respects this conception of the physical world is radically revisionary with respect to our common sense ways of thinking. One would never guess just by looking at it (or touching it) that a lump of rock consists of trillions and trillions of vibrations taking place in invisible fields. It is also natural to assume that a region of empty space—a cubic metre midway between two galaxies, say—is truly empty. According to the Standard Model even the emptiest region of space is in fact filled

[21] https://home.cern/about/physics/standard-model.

with the quantum fields which—due to quantum uncertainties—continually generate extremely small, very short-lived ("virtual") particles in large numbers.

However, so far as the nature and role of*forces* are concerned, the picture drawn here may seem to be reassuringly familiar. At the most fundamental level, we are told that the world is being held together by forces. In its attractive mode the electromagnetic force ensures that positively charged protons and negatively charged electrons remain bound within atoms. It is electromagnetic repulsion between the electron "shells" surrounding atoms which prevents our feet from falling through floor. And in the case of the quarks composing protons and neutrons, the strong force binding them is *so* strong that the quarks in question can never be separated from one another. Given that we are all acquainted with the nature of force from our own experience, it seems that our experience—in this respect at least—is providing us with a reliable guide to the nature of reality.

In fact, drawing this reassuring conclusion would be premature. The impressive empirical successes of the Standard Model—the prediction of the Higgs boson is by no means the first of these—have convinced most physicists that the theory accurately reflects some important aspects of the way the world really works, but there remain plenty of unresolved problems.

The theory does not incorporate either dark matter or dark energy, which remain mysterious. Also, the Standard Model has yet to incorporate gravity. As CERN note in their introductory guide: "… the most familiar force in our everyday lives, gravity, is not part of the Standard Model, as fitting gravity comfortably into this framework has proved to be a difficult challenge." This is nicely understated: the problem of reconciling quantum theory with general relativity remains unsolved, despite receiving the attentions of many of the best minds in physics over a period of many years.

Since no one yet knows what a viable quantum gravity theory will look like we are similarly ignorant as to the character of the theory which will succeed the Standard Model.

Also, as we have already seen, quantum theory poses notorious problems of interpretation, which all extend to the Standard Model simply because it *is* a quantum theory. Indeed, the Standard Model generates several new problems of its own. Quite what the best mathematical formulation of it will turn out to be remains controversial—there are a number of competing alternatives. Calculations using the theory tend to produce physically unrealistic infinities; although these have been partially tamed by "perturbation" techniques the suspicion remains that a better theory will not have this consequence. The Standard Model includes a large number of parameters that need to be determined experimentally—the theory provides no clue as to why these parameters have these particular values rather than others. Estimates for the energy of the vacuum derived from the Standard Model turn out to be enormously larger than the value predicted by GTR. Also, and significantly from a metaphysical standpoint, the basic ontology of the Standard Model is very much open to debate. Contemporary theorists remain divided on the question of whether the basic entities in QFTs are fields or particles—there are considerations which point in different directions.[22]

Given the current state of play little is very clear, but one thing does emerge with at least some clarity. Interpretations of the Standard Model in terms of particles consisting ultimately of excitations in fields which interact by exchanging other particles, do provide an account of the basic nature of reality which is intuitively appealing by virtue of being easily visualizable. It may even be that this a picture along these lines proves to be correct; but it is equally possible that it does nothing of the sort.[23]

We saw earlier that Einstein's relativity theories have implications for the nature of time that also impact on our understanding of forces. If, as many have concluded, in the light of Einstein we have no option but to conclude that we live in a four dimensional block universe, the future is as real as the past, and we can no longer view causes (or forces) as bringing their effects into existence. There is another important respect in which the nature of time and the nature of forces and causes are interrelated, one that is entirely independent of relativistic considerations.

Thanks to the work of Euler, Lagrange, Hamilton and others a comprehensive alternative mathematical framework for carrying out Newtonian mechanics was developed in the 18 and 19th centuries. On this alternative picture, the role of forces —both of the impact and action at a distance variety—is supplanted by global "variational principles" such as the principle of least action (in mechanics) or least time (in optics). Since these principles minimize (or maximize) properties of an object's *entire path* through space over an interval of time, they presuppose a four-dimensional view of nature according to which the future is no less concrete and real than the present.

Since in the case of classical mechanics the "Newtonian" and "Lagrangian" approaches are completely equivalent, we cannot draw any implications in that domain concerning the nature of time from the fact that the success of the Lagrangian methodology.[24] However, variational principles not only play a key role in all the main formulations of quantum field theories, they are also at the heart

[22] For more on the difficulties confronting QFT see https://plato.stanford.edu/entries/quantum-field-theory/.

[23] Nima Arkani-Hamed, who has recently pioneered impressive new geometry-based ways of performing calculations in QFT makes the point thus: "… there are more and more people trying to explain quantum field theory in an accessible way … [but] they're explaining a point of view about the subject which is thirty or forty years old and which is almost certainly not going to be the way we think about it in the future. … the one thing that is almost certainly *not* going to be the case is that the story is that *The big deal is that there are those different fields and there are these particles that are excitations of the field*." Burton (2013), 377. See Wolchover (2013) for an accessible introduction to Arkani-Hamed's work on the amplituhedron, the higher dimensional geometrical entity underlying the new QFT methods.

[24] For more on variational principles and the metaphysical conundrums to which they give rise see Smart and Thebault (2013)—also see Chiang's (2002) sci fi story.

of many attempts to reconcile quantum theories with relativity. If this remains the case, and no alternatives to the variational approaches emerge, this could be taken as compelling evidence that nature is itself four-dimensional, and that global variational principles—rather than forces as traditionally conceived—have explanatory priority. This said, anyone who finds this conception of time unacceptable on metaphysical grounds will still have the option of holding that quantum theories should be interpreted only instrumentally, i.e. as useful tools for predictive purposes, rather than reliable guides to the nature of reality.[25]

References

H.G. Alexander (ed.), *The Clarke-Leibniz Correspondence* (Manchester University Press, Manchester, 1955)

J. Barbour, *Absolute or Relative Motion? Vol. 1: The Discovery of Dynamics* (Cambridge University Press, Cambridge, 1989)

J. Barnes (ed.), *The Complete Works of Aristotle* (Princeton University Press, Princeton, 1984)

G. Berkeley, *The Works of George Berkeley*, *Bishop of Cloyne*, ed. by A.A. Luce, T.E. Jessop (Thomas Nelson and Sons, London, 1948–1957)

I. Born (trans.), *The Born-Einstein Letters* (Walker and Company, New York, 1971)

J.V. Buroker, Kant, the dynamical tradition, and the role of matter in explanation, in *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 1972 (1972), pp. 153–164

H. Burton, *The Power of Principles: Physics Revealed: A Conversation with Nima Arkani-Hamed* (2013)

S. Carroll, *The Particle and the End of the Universe: The Hunt for the Higgs and the Discovery of a New World* (Oneworld, London, 2012)

T. Chiang, Story of your life, in *Stories of Your Life and Others* (Vintage, New York, 2002)

J.G. Cramer, *The Quantum Handshake: Entanglement, Nonlocality and Transactions* (Springer, Dordrecht, 2016)

B. Dainton, *Time and Space*, 2nd edn. (Routledge, London, 2010)

R. Descartes, *Philosophical Writings of Descartes*, trans. by J. Cottingham, R. Stoothoff, D. Murcoch, A. Kenny (Cambridge University Press, Cambridge, 1984–1989)

D. Deutsch, *The Fabric of Reality* (Penguin, New York, 1997)

A. Einstein, B. Podolski, N. Rosen, Can quantum mechanical description of physical reality be considered complete? Phys. Rev. **47** (1935)

M. Faraday, *Experimental Researches in Electricity*, vol. 3 (Taylor and Francis, London, 1837–1855)

R. Feynman, *Lectures in Physics*, vol. 2 (Addison-Wesley, Reading Ma, 1964)

J. Foster, Induction, explanation and natural necessity, in *Proceedings of the Aristotelian Society* 83 (1983)

M. Friedman, *Kant and the Exact Sciences* (Harvard University Press, Cambridge Mass, 1992)

M. Friedman, *Kant's Construction of Nature* (CUP, Cambridge, 2013)

S. Greenblatt, *The Swerve: How the World Became Modern* (Norton, New York, 2011)

[25] For a recent defence of this approach to the realm of the quantum see Healey (2017).

M. Giustina et al., Significant-loophole-free test of bell's theorem with entangled photons. Phys. Rev. Lett. **115** (2015)

R. Healey, *The Quantum Revolution in Philosophy* (Oxford University Press, Oxford, 2017)

B. Hensen et al., Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. Nature **526**(7575) (2015)

M. Hesse, *Forces and Fields* (Nelson, Edinburgh, 1961)

C. Hitchcock, *Probabilistic causation, Stanford Encyclopedia of Philosophy* (2010), https://plato.stanford.edu/entries/causation-probabilistic/

T. Hobbes, *De Corpore*, trans. by A.P. Martinich (Abaris Books, New York, 1981)

D. Hume, *A Treatise of Human Nature*, ed. by L.A. Selby-Bigge, Revised, P.H. Nidditch (Clarendon Press, Oxford, 1975)

D. Hume, *An Enquiry Concerning Human Nature*, ed. by T.L. Beauchamp (Oxford University Press, Oxford, 1999)

J. Ismael, J. Schaffer, Quantum holism: nonseparability as common ground. Synthese (2016)

M. Jammer, *Concepts of Force* (Dover, New York, 1999)

I. Kant, *Theoretical Philosophy, 1755–1770* (Cambridge University Press, Cambridge, 1992)

I. Kant, *Metaphysical Foundations of Natural Science*, trans. by M. Friedman (Cambridge University Press, Cambridge, 1994)

I. Kant, *Critique of Pure Reason*, trans. by P. Guyer, A. Wood (Cambridge University Press, Cambridge, 1997)

M. Kuhlmann, *Quantum Field Theory*, Stanford Encyclopedia of Philosophy (2012) https://plato.stanford.edu/entries/quantum-field-theory/

G.L. Leibniz, *Philosophical Papers and Letters*, vol. 2, trans. and ed. by L. Loemker (Springer, Dordrecht, 1989)

P.J. Lewis, *Quantum Ontology: A Guide to the Metaphysics of Quantum Mechanics* (Oxford University Press, Oxford, 2016)

J. Locke, *An Essay Concerning Human Understanding*, ed. by P. Nidditch (Oxford University Press, Oxford, 1975)

T. Maudlin, *Quantum Non-Locality & Relativity*, 3rd edn. (Wiley Blackwell, Chichester, 2011)

J.C. Maxwell, *A Treatise on Electricity and Magnetism* (Dover, New York, 1954)

E. McMullin, The origins of the field concept in physics. Phys. Perspect. **4**, 13–39 (2002)

H. Minkowski, Lorentz et al., Space and time, in *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity* (Dover, New York, 1952), pp. 75–91

G. Musser, *Spooky Action at a Distance* (Farrer, Straus & Giroux, New York, 2015)

I. Newton, *Principia*, trans. by A. Motte, Revised, F. Cajori, (University of California Press, Berkeley, 1962)

A. Ney, D. Albert, *The Wave Function: Essays on the Metaphysics of Quantum Mechanics* (Oxford University Press, Oxford, 2013)

D. Oriti, *Approaches to Quantum Gravity: Toward a New Understanding of Space, Time and Matter* (Cambridge University Press, Cambridge, 2008)

K. Popper, *A World of Propensities* (Thoemmes Press, Bristol, 1990)

M.G. Raymer, *Quantum Physics: What Everyone Needs to Know* (Oxford University Press, Oxford, 2017)

C. Rovelli, *Unfinished Revolution*, ed. by Oriti (2008)

B.T.H Smartand, K.P.Y Thebault, On the metaphysics of least action (2013), https://arxiv.org/abs/1511.03429

G. Strawson, Realism and causation. Philos. Q. **37**(148), 253–277 (1987)

G. Strawson, *The Secret Connexion* (Oxford University Press, Oxford, 2014)

C. Truesdell, *An Idiot's Fugitive Essays on Science* (Springer, New York, 1984)

L. Vaidman, *Many-Worlds Interpretation of Quantum Mechanics*, Stanford Encylopedia of Philosophy (2014)

E. Watkins, *Kant and the Metaphysics of Causality* (CUP, Cambridge, 2005)

R.S. Westfall, *Force in Newton's Physics: The Science of Dynamics in the Seventeenth Century* (Mcdonald, London, 1971)

N. Wolchover, A jewel at the heart of quantum physics (2013) https://www.quantamagazine.org/physicists-discover-geometry-underlying-particle-physics-20130917/

W. Yourgrau, S. Mandelstam, *Variational Principles in Dynamics and Quantum Theory* (Dover, New York, 1979)

D.J. Zeyl (trans.), *Plato*: *Timaeus* (Hackett Publishing, Indianapolis, 2000)

Chapter 12
Map and Territory in Physics: The Role of an Analogy in Black Hole Physics

W. G. Unruh

The generic territory this paper will concern itself with is that of the physical world, and the map is that of theoretical physics: the theories, primarily mathematical, that one generates to describe, to predict new aspects of, that physical world. That such a map is even possible, and furthermore that such a map is such an accurate representation of the physical world is something that amazed and surprised physicists even before the time of Newton. Already Pythagoras astonished both himself and the intellectual world by mathematizing an aspect of the world, that of harmony of musical notes. Two notes produced by different lengths of identical musical strings, such that those lengths bore small whole number ratios with respect to each other, would sound harmonious, while those with arbitrary ratios sounded inharmonious and clashing. Understanding the origin of this mathematization of the physical world formed one of the primary puzzles which exercised the minds of top physicists for 2000 years. That the eventual solution told us as much about the peculiarities of the human mind, as it did about the physical world does not detract from the guiding light that Pythagoras's observation shone in the development of physics (Cohen 1984).

At the same time, analogy has played a guiding role in the rational understanding of the world. In terms of the central metaphor of this book, that of human understanding seen as the interplay in geography between the map and the territory, the question is, "If the map of two regions is the same, how much can we say about the similarity of the territory that the maps describe?"

One of the most astonishing features of modern 20th and 21st century physics has been how similar the mathematical tools are which are used to describe what, on the face of it, are utterly disparate phenomena. Quantum Field theory, developed to describe the quantum mechanics of electromagnetism, and which eventually became the dominant paradigm of elementary particle physics, has also come to dominate the theoretical structure of condensed matter physics.

W. G. Unruh (✉)
CIfAR Cosmology and Extreme Gravity Program, Department of Physics,
University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada
e-mail: unruh@physics.ubc.ca

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_12

In this paper I want to show how one can use the mathematical analogy between two seemingly disparate areas of physics to cast light on both.

Back Holes were one of the most surprising predictions of Einstein's theory of gravity. That theory began with Einstein's insight that gravity, rather than being a force, was actually closely associated with the nature of time. Newton had described the gravity we experience daily as a mysterious force, an "action at a distance", which caused two massive bodies to feel a force of attraction to each other. Einstein (already by 1908) realized that gravity could instead be described as the inequable flow of time from place to place. One often hears that gravity can cause clocks to tick at different rates from place to place. But that is a perversion of the story that his theory tells. Instead it is precisely the ticking of time differently from place to place that is the gravitational field which we usually feel (Unruh 1995). Combining this with Minkowski's description of special relativity as combining distances in time and distances in space into one unified notion of distances in space-time, and with Newton's realization that the motion of matter in the absence of external forces follows straight lines (the shortest distance between two points), Einstein showed how all of Newton's theory of gravity could be subsumed into the law that matter causes time to flow differently from place to place. Of course Einstein's theory, General Relativity, is more complex since if time can flow differently from place to place, then spatial distances can also change from time to time (leading to cosmology, where the distances between each object in the universe can increase or decrease with time without the objects themselves moving, and to the existence of gravitational waves where distance changes can propagate at the speed of light).

Only a few months after Einstein had laid out his theory, Karl Schwarzschild, a German soldier on the Russian front of WWI, found the first exact solution of the equations. He showed that the consequences of the theory were even more dramatic than anyone had expected. It took almost 50 years for physicists to realize that his solution implied that one could have regions of space which could be entirely out of contact with the rest of the universe. Even at the speed of light, anything inside what is now called the horizon in Schwarzschild's solution, cannot communicate or interact with anything outside (unless that outside object also falls into the horizon). In honour of this behaviour, Wheeler popularized the name "Black Hole" for this phenomenon. But almost immediately after these objects had been named, another shock was delivered. Hawking (1977) argued that, if one takes seriously the behaviour of the aforementioned quantum fields near the black hole, it ceased to be black. It radiates, and surprisingly, it radiates as though it were a hot body, with a temperature inversely proportional to the mass. Thus a solar mass black hole has a temperature of about 10^{-6} K, but an earth-mass black hole radiates with a temperature about a million times higher, while the black hole in the centre of our galaxy has a temperature of the order of the coldest temperature ever achieved in terrestrial labs.

As stated, that temperature is a function of the mass of the black hole. The mass of the black hole is, via $E = mc^2$, expressible in terms of the energy of the black hole—the total energy which has fallen in to make the black hole. This suggests that the black hole is a thermodynamic object, with an entropy. Using the expression Hawking found (in units in which $\mathrm{G} = \mathrm{c} = \hbar$)

$$T = \frac{1}{8\pi M} \tag{12.1}$$

one finds that the entropy is just one quarter of the area of the black hole, with the area expressed in terms of the Planck area (the area expressed in purely in terms of $\hbar$, G, c). As with any thermodynamic object, this entropy limits the efficiency with which one can use the black hole to convert heat into work (Unruh and Wald 1982). The understanding of Hawking's result has been a driving force in theoretical physics in the past 45 years.

Bekenstein (1973) had suggested that black holes should have an entropy by following Wheeler's unsupported suggestion that black holes, as absorbers of entropy, should also act as thermodynamic objects and had an entropy. This idea ran into the road block that black holes are black. They do not have any temperature except 0. While black holes had formal features which suggested the laws of thermodynamics, at best everyone took these as formal unphysical analogies. Hawking's unexpected result shocked and inspired the theoretical community. Black holes are thermodynamic objects. This result was too surprising to be false, but in the past 40 years, the understanding of the source of thermodynamic aspect of black holes has remained largely a mystery. One of the greatest mysteries is the entropy. Entropy was introduced in the mid 19th century to explain the operation of heat engines. Later in the 19th century, Maxwell, Gibbs, Boltzmann and others explained entropy in terms of statistical mechanics. For them, the entropy is related to the number of different states the system could have at a given energy or temperature consistent with the macroscopic parameters one could access in using the system in a heat engine. But what is the entropy of a black hole? How does it relate to statistical properties of a black hole, and what are those microscopic degrees of freedom needed to give it a statistical interpretation?

However, Hawking's temperature rested on a strange aspect of quantum field theory in the vicinity of the black hole. The evolution of quantum fields is deterministic. The thermal emission must arise from some aspect of the initial state of the field, which was assumed to be the vacuum state. What aspects of that vacuum state result in the thermal emission after the black hole has formed? Hawking essentially operated backwards. Given the final state, of the field, what aspect of the initial state could have produced it? To find it, one can evolve the final state backward in time. And because of the linearity of the field which he used to calculate, one can do this mode by mode. Given any mode of the field (some distribution of the field obeying the classical equations of motion) one must see where it came from in the initial state. It cannot come from inside the black hole (nothing can get out of the black hole by definition of what a black hole is). But it comes from the direction of the black hole. It must therefore come from a vicinity closer and closer to the horizon of the black hole. In fact, the equations of motion say it comes from a region exponentially closer with a scale of the radius of the black hole, and a time scale of the light-travel time across a distance of the order the size of the black hole. It continues to get closer and closer to the horizon until one gets to a time when the black hole forms, when that mode can escape out toward infinity. By that time its wavelength is tiny, and frequency extremely high. For example, for a solar mass black hole, a thermal mode, of frequency near the maximum of the thermal spectrum, emitted one second after the formation of the black hole via collapse, must have originated from a quantum vacuum fluctuation in the initial vacuum state with a frequency of order e^{10^5}. This would have an energy of the order of e^{10^5} times the mass of the whole universe. Clearly, for modes of this energy, the assumptions that the field is a simple linear, non-interacting field is extremely suspect. Does this mean the prediction of thermal radiation is wrong?

This problem with Hawking's derivation was clear very soon after his discovery. It has also misled many researchers throughout the years into believing that Hawking's result depends on high energy Planck scale physics. There certainly seems no way of avoiding this conclusion if one takes his derivation seriously.

In 1972 I was asked by Denis Sciama to give a colloquium at Oxford on black holes. Desperately trying to think of some way of making some of the properties of black holes approachable by the audience who had never heard of such objects, I thought of an analogy, that of a waterfall. If one imagines a waterfall so high that the velocity of the water somewhere exceeds the velocity of sound at some surface, then that surface acts very much like a black hole horizon as far as sound is concerned. Sound cannot escape out of that surface, since the sound there is swept back over the waterfall at the same rate as it is trying to escape. Furthermore, any sound trying to escape from just outside that surface takes a long time to get out. The closer it is to that surface that the sound is emitted, the longer it takes to escape. Both of these features are similar to what happens to light near a black hole. No light can escape from behind the horizon, and the time it takes for the light emitted just outside the horizon gets longer and longer the closer the emission is to the horizon. The sound waves emitted nearer and nearer the horizon are bass-shifted, just as light emitted nearer and nearer the horizon is red-shifted.

This analogy was just that, an analogy whose only purpose was to try to clarify some features of a black hole. It indicates a similarity between sound and light, but as it stands it does not indicate that the two territories share a map, a detailed mathematical similarity. In 1980, I was assigned a course on Fluid Mechanics to teach. One evening, while preparing my lecture for the next morning, my mind wandered back to that analogy and I decided to try to see how well the analogy actually worked. Was it more than a pretty picture? To do so I wrote the equations of motion of an irrotational fluid, separating them into some time-independent background flow and a small linear perturbation around that flow. Those perturbations were to represent sound waves. Introducing the velocity potential (possible because the flow was irrotational), and eliminating the fluctuations in the density between the resultant two differential equations, I got an equation for that velocity potential which looked just like the equation for a scalar field in a background spacetime. In this case that effective spacetime is determined by the background flow and density of the fluid, not by the relation between spacetime and gravity as in Einstein's theory.

Furthermore, one could imagine quantizing those linear perturbations, the sound waves. Such a quantization of sound waves is standard practice in condensed matter physics, where the quantized sound excitations are called phonons. One could then follow Hawking's derivation of the thermal emission by a black hole, step by step, for this quantum field (the velocity potential) in this effective spacetime metric. If the fluid flow was such that at some place the velocity of the fluid exceeded the velocity of sound, that effective metric looked in many ways like that of a black hole, with a Killing horizon (i.e., a horizon defined by the condition that the vector denoting the time displacement symmetry becomes null in the effective metric). A straightforward calculation shows that this quantum field should also produce a thermal flux of phonons, just as the black hole produces a thermal flux of photons. In the latter case the temperature is proportional to the inverse mass of the black hole. In this case the temperature is equal to

$$T = \frac{1}{4\pi c}\frac{d(c^2 - v^2)}{dx} \tag{12.2}$$

where x is the distance along the flow lines of the fluid which go into the horizon where $v^2 = c^2$ (Unruh 1981).

One thus has the same map—the propagation of the field in a spacetime—and the same conclusion—the quantized small fluctuations of that field result in thermal emission from the horizon, with the temperature of that emission determined by properties of that background spacetime. The same map of the two diverse territories implies that unexpected features of the territories also seem to be the same. This conclusion that sonic horizons would also produce a thermal quantum spectrum is also a surprising conclusion, but in both the black hole and the dumb hole cases, the problem of ultra high frequencies in the initial states is the same. If maps are identical then the territories, at least to the extent that the maps are accurate, must also be identical.

But this conclusion in the case of dumb holes (the name given to such sonic analogs to black holes) is clearly wrong. In the case of the sound waves, one can understand the emission of the thermal radiation in the same way. Tracing back the modes of the sound which are thermally excited in the future, one finds again that the horizon is a one-way membrane, at least in the simple model of sound derived from the Navier-Stokes equations. Those modes cannot come from inside the horizon, and must therefor be squeezed more and more against the horizon as one goes into the past. The bass-shift of the outgoing waves near the horizon implies an exponential squeezing of the modes against the sonic horizon, just as the light in the black hole case is squeezed against the horizon because of the red shift of the radiation emitted by a source falling into the black hole. But in the case of sound waves, we understand that the hydrodynamic equations are an approximation. At short wavelengths the fluid cannot be described by a continuous density with some velocity, but rather must be described as a conglomeration of distinct, spatially separated atoms. Sound waves ultimately are a description of the average motion of those atoms around some background equilibrium flow. And sound waves cannot have a wavelength shorter than the average distance between the atoms.

The equivalence of the maps in the sonic and the black hole case breaks down. Or does it? After all one has the gut feeling, which goes all the way back to Planck, that at some scale, quantum gravity effects should come into play in the case of the black hole. This can be seen to be in analogy to atomic effects coming into play in the case of the dumb hole.

One of the worries about the black hole is that perhaps those quantum gravity effects could destroy the thermodynamic edifice erected around black holes via Hawking's discovery. If Hawking radiation really depends on those exponentially large frequencies and exponentially tiny distances which his derivation requires, then the necessary alteration of the theory at those scales by the effects of quantum gravity might destroy the effect he discovered.

It is precisely here that the sonic model might come to the rescue. We understand precisely how the hydrodynamic equations break down, and we understand, at least in theory, what a truer description of a fluid is. It is the collective motion of a bunch of atoms. The calculations of how the fluid behaves in terms of the individual atoms might be horrendously complicated but, unlike the case for quantum gravity, we have a strong faith that the essentials of the theory of fluids are known. So we can ask, "Does the thermal radiation emission by a dumb hole survive the generalization of hydrodynamics to a fully atomic description of the fluid?" If it does not, then one has no faith that the Hawking effect would survive a fully quantum treatment of gravity. If the prediction of dumb-hole thermal radiation does survive, then it may give us clues as to how the black hole thermal radiation might also survive the effects of small scale quantum gravity.

When I wrote the paper which resulted from my evening's distraction from lesson preparation, I realized the potential usefulness of the dumb-hole model in deepening our understanding of black holes. But I had no idea how to actually carry out a calculation treating the atoms of the fluid as fully quantum objects. I tried to imagine how I would even start to carry out a fully non-linear quantum treatment of 10^{25} interacting atoms. Fortunately I gave a seminar at the University of Texas where Ted Jacobson was in the audience. About 10 years later, he realized that one of the key effects of the atomic nature of the fluid was to change the dispersion relation of sound waves, i.e., instead of the velocity of sound, whether phase or group velocities, being a constant, independent of frequency or wavelength, the atomic nature of matter caused the velocity of sound to change at short wavelengths. How it changes depended on the particular nature of the fluid. For liquid helium, for example, both the group and phase velocities would, at short enough wavelengths, decrease from their values at long wavelengths. For a Bose Einstein condensate fluid on the other hand, the velocity of sound would increase as the wavelength became shorter and shorter. It was this realization which allowed people to begin to answer the question as to what the effect of the atomic nature of the fluid on the analog to Hawking radiation could be.

In the above description of how the horizon affects the modes which eventually come away from the horizon in a thermally excited state, the key was that the modes got squeezed up more and more against the horizon as one propagated those modes backward in time, until one got those absurdly high frequencies and wavelengths. The change in the dispersion relation, the change in the velocity of sound with frequency, means that, while initially those modes are again squeezed against the horizon, eventually their wavelength becomes sufficiently short that their velocity is no longer the same as the velocity of the fluid. If their velocity decreases with frequency, those waves must have been swept in from outside the horizon. If the velocity increases with frequency, those waves must have travelled from inside the horizon out to the horizon. In either case, that squeezing of wavelength ends once the wavelength reaches the value where the velocity of the waves changes from the velocity of sound at long wavelengths.

What Jacobson's observation meant was that the modes of propagation of the sound waves always remained in a regime in which they acted like linear sound-waves, with wavelengths much longer than the inter-atomic spacing. One did not have to worry that the highly non-linear regimes of the inter-atomic interactions would destroy ones ability to do calculations. In general the equations can still not be solved analytically, but they can be solved numerically. Soon after Jacobson's observation, both I (Unruh 1995), and then Corley and Jacobson (1996) did just that and found that the change in the dispersion relation at high frequency had essentially no effect on the thermal emission at low enough frequencies the quantum sound emission behaved just as in the hydrodynamic approximation. Although at high frequencies, radiation begins to deviate from thermal, at lower frequencies thermal spectrum is a very good approximation. The thermal spectrum is insensitive to the behaviour of the equations of the field at short spatial or temporal scales. The thermal behaviour of the emission from horizons is a robust phenomenon. This suggests strongly that the concern, that Hawking's derivation requires a specific behaviour of the fields at arbitrarily high frequencies or arbitrarily short spatial scales, is misplaced. Hawking radiation is a low frequency, large (relatively) distance phenomenon. It is not a magic road to Planck scale physics.

One can understand this in a hand-waving way by the following argument. Consider a mode of the field which begins life far from the location of the future black hole, and which has a very high frequency. Our assumption is that a mode begins in its ground, or vacuum, state. Because of its high frequency, it sees the surrounding metric change on scales which are of much lower frequency and longer spatial scales than its own. By the quantum adiabatic theorem, a quantum system which is perturbed on time scales much longer than its own does not change its state. If it begins in its ground state, it remains in its ground state. As the mode propagates near the horizon of the black hole, this adiabatic behaviour of the surrounding spacetime continues until the frequency has been red-shifted by its propagation along the horizon to a value which is the same order as the rate of change of the surrounding metric (the time scale and spatial scale of the curvature of the black hole). It is only at this point that that the time-dependence of the surrounding spacetime begins to change the state of that mode of the field, creating particles (excitations away from the ground state of that mode). If this argument is correct (and no rigorous derivation exists which demonstrates that this argument is correct), then the Hawking radiation is truly a low energy, long wavelength process.

Entropy

As part of his thesis project under John Wheeler, Jacob Bekenstein argued that black holes should have an entropy. Wheeler had argued that because black holes could absorb the entropy of the matter falling into the black hole, it should also have entropy itself. Otherwise one could get rid of an arbitrary amount of entropy from the external universe, and perhaps violating the second law of thermodynamics. Since a classical black hole has a zero temperature, and since a zero temperature heat bath can (barring the third law) absorb and arbitrary amount of entropy, Wheeler's argument was somewhat shaky. Bekenstein however ran with the idea. Hawking had just shown that the laws of classical General Relativity, together with the requirement that matter always have positive energy, implied that the surface area of a black hole must always increase. Since entropy (by the second law) must also always increase, it was very suggestive to Bekenstein that perhaps there was some relation between the area of a black hole and its entropy. He generated a number of arguments that this identification of entropy and area was more than an analogy. However, this analogy foundered on the problem that if the black hole had an entropy, and since it certainly had energy, it must also have a temperature. Classical black holes have at best a zero temperature. Geroch pointed out that if one regarded the area as the entropy one could violate the second law of thermodynamics if the black hole temperature was zero.

It was Hawking's discovery that quantum field theory implied that black holes did have a temperature, a temperature moreover which was a function of the mass of the black hole that gave a way out of this impasse. Using the standard thermodynamic relation, $dE = TdS$, one found that the entropy must be equal to 1/4 of the surface area of the black hole, as measured in Planck units. Various arguments showed that this entropy was more than just a fluke. In particular, if one operated a heat engine with the gravitational field of the black hole being used to convert heat energy to work, then such a heat engine obeyed the standard Carnot efficiency if the surface area of the black hole was the entropy required in the Carnot argument. The entropy of the black hole is a real thermodynamic entropy.

The big question then was whether or not the arguments of Maxwell, Gibbs and Boltzmann, that entropy is related to the uncertainty of microstate of the system under the constraint that the few degrees of freedom used by the heat engine be fixed, were correct. What are these internal degrees of freedom of a black hole? Or, alternatively, is the entropy of a black hole not of any statistical origin, but is a "pure" entropy, unrelated to a counting of the microscopic degrees of freedom?

It is these questions which the sonic model can perhaps also shed light on. For the sonic analog, there is no relation between the energy in the waterfall, and the temperature. There is then also no entropy associate with a dumb hole. Yet, in both the black and dumb hole cases, one finds that the fields living on this spacetime (e.g. photons in black holes, and phonons in the dumb hole case) are emitted with a thermal spectrum. That thermal spectrum is not the result of the dynamics of any hidden degrees of freedom of the spacetime, but is a direct consequence only of the smooth metric structure which determines the equations of motion of the quantum fields.

In ordinary statistical mechanics, there is an intimate relation between the microscopic degrees of freedom of the thermal object and the thermal radiation emitted by that object. It is precisely those microscopic degrees of freedom which create the radiation which escapes from the body. It is because those micro degrees of freedom move and change that the radiation is created. In the case of both the black holes and the dumb holes this is not the case. The background metric does not change. It is not due to its alterations, due to its thermal excitation, that the radiation is created. Rather it is because of the quantum field's motion over the smooth surface of the spacetime that the radiation is created. To me this suggests that the entropy (which, as I said, is a genuine thermodynamic entropy in the case of black holes) is not the result of microscopic degrees of freedom, but is fundamentally thermodynamic entropy, unrelated to any microscopic degrees of freedom.

Experiment

My original paper on the sonic analog was entitled "Experimental Black Hole evaporation?" What excited me was the possibility that one could, in a terrestrial laboratory, carry out experiments which were directly related to the thermal emission by black holes. No matter what the approximations used to solve the theory, they are approximations and one is never sure how accurate they are. Furthermore there can be additional physical effects which are not included. One example is that the viscosity of a fluid might affect the thermal radiation. Or turbulence in the fluid, or a host of other effects. In the presence of quantum and classical fluctuations, the exact location of the horizon is uncertain. Do those fluctuations in the position of the horizon affect the thermal radiation? If the high frequency behaviour of the field (e.g. its squeezing against the horizon) changes the horizon then one might expect that the location of the horizon could be important. The waves could be squeezed up against the position of the horizon at one time, only to have the horizon shift so that those squeezed waves are now either inside or outside the horizon. If the claim above is true, that the thermal emission is not a high frequency phenomenon, but represents the reaction of the field at low frequencies and long wavelengths to the changes in the metric field, then one would not expect the exact location of the horizon to be important. This is a question that, potentially, experiments could resolve.

There have now been a number of experiments to look for the thermal nature of the radiation (Daniele Faccio et al. 2013). One set of experiments, initiated by Germaine Rousseaux, and carried to completion by a group at the University of BC (Weinfurtner et al. 2010), used water as the medium for creating a dumb hole, and used the surface gravity waves as the field which carries the thermal emission. Of course the quantum emission would be impossible to see. Its temperature (of the order of 10^{-12K}) is far colder than the temperature of liquid water, but a stimulated emission experiment could be carried out. As Einstein, with his A and B coefficient analysis, showed, knowledge of stimulated emission is sufficient to also understand the spontaneous emission in a system. In these experiments the alteration of the dispersion relation was created by the transition from shallow water waves to deep water waves. The experiment showed that the spectrum of the quantum emission, assuming that Einstein's analysis is correct, would be thermal, with a temperature of the order of 10^{-12} K.

Another recent experiment was by Jeff Steinhauer (2016) using BECs. He looked for fluctuations in the density of the BEC as the measurable quantity of the created quantum phonons. In his case the experiment was too noisy to be able to see a thermal spectrum, but there was a suggestion that there was entanglement between the waves travelling in opposite directions, away from the horizon. Such entanglement would be expected for the creation of Hawking radiation by a horizon, and would be a signature that the process creating those fluctuations was quantum, and not simply the amplification of some classical noise source.

An additional path has been the attempt to use light in a medium to form a black hole type horizon by altering the index of refraction in the medium (see for example Belgiorno et al. 2010). Since the media are solids one cannot have the medium flowing with different velocities. Instead one must have a region in which the velocity of the light is changed, with that region travelling at almost the velocity of light. In most of the experiments of this nature this is done by using an intense region of light whose non-linear interaction with the medium changes its refractive index. So far this promising approach has not yet exhibited quantum emission.

Conclusion

All maps are approximations to the territory they describe, including the mathematical maps which physics use to describe their territory, the world. That the maps which describe different territories can be similar at a certain level of approximation allows us to gain understanding of a poorly understood territory by applying the lessons from the better understood territory. This is the role that analogy has played throughout history. What we see in the example which this article has looked at it that that understanding can come from the differences as much as, or perhaps even more so, than from the similarities.

Acknowledgements I would like to thank the Natural Science and Engineering Research Council of Canada for their research support during this work. I also thank the Canadian Institute for Advanced Research for support and for keeping me in a stimulating environment for carrying out research. A vast variety of fellow researchers interested in the topic that has come to be called Analog Gravity have pushed me, stimulated me, kept me honest, and helped to do far more than I could have done on my own.

References

J.D. Bekenstein, Black holes and entropy. Phys. Rev. D **7**, 2333–2346 (1973)

See for example Belgiorno et al. Phys. Rev. Lett. **105**, 203901 (2010)

S. Corley, T. Jacobson, Hawking spectrum and high frequency dispersion. Phys. Rev. D **54**, 1568–1586 (1996)

For a history of this search see for example H.F. Cohen *Quantifying Music* (Springer, 1984)

For an exposition of this see W. Unruh *Time Gravity and Quantum Mechanics* p 23 in Time's Arrow Today ed. by S. Savitt (Cambridge University Press, 1995)

For an introduction to both the theory and experiments in Analog gravity, see the volumes **Analogue Gravity Phenomenology: Analogues Spacetime and Horizons, from Theory to Experiment** ed. by Daniele Faccio et al (Springer, 2013) or **Analogue spacetimes: The first 30 years** ed. by V.M.S. Cardoso, L.C.B. Crispino, S. Liberati, E.S. Oliveira, M. Visser (Editora Livraria da Fisica, Sao Paulo, 2013)

S.W. Hawking, Black hole explosions? Nature **248**, 30 (1974) and S.W. Hawking, Particle creation by black holes. Comm. Math. Phys. **43**, 99 (1975). For a popular exposition see also S.W. Hawking, The quantum mechanics of black holes. Sci. Am. **236**, January p34 (1977)

J. Steinhauer, Observation of quantum Hawking radiation and its entanglement in an analogue black hole. Nat. Phys. **12**, 959 (2016)

W.G. Unruh, Experimental black hole evaporation? Phys. Rev. Lett. **46**, 1351 (1981)

W.G. Unruh, Sonic analogue of black holes and the effects of high frequencies on black hole evaporation. Phys. Rev. D **51**, 2827 (1995)

W.G. Unruh, R.M. Wald, Acceleration radiation and the generalized second law of thermodynamics. Phys. Rev. D **25**, 942 (1982)

S. Weinfurtner, E. Tedford, W.G. Unruh, G. Lawrence, Measurement of stimulated Hawking emission in an analogue system. Phys. Rev. Lett. **106**, 021302 (2010)

Chapter 13
Topological Foundations of Physics

Joseph Kouneiher

Introduction

The idea that the universe is governed by precise causal or dynamical laws, is a very old one, and it was for a big part due to Galileo, Descartes, Newton and others (Kouneiher 2017). Here, we should understand the word *laws* as a set of true principles that form a strong but simple and unified system that can be used to predict and explain. *it's a way to understand a great many complicated phenomena in a unified way, in terms of a few principles* (Anandan 2002). As we know Newton had formulated a highly successful set of laws for material particles, known today as Newton's laws of motion and gravitation. So it was natural for Newton to unify the behavior of light and his laws for material particles by making the hypothesis that light consisted of material particles, called corpuscles. But many observations of the light behavior push the physicists to abandon Newton's ontology of corpuscles and replace it by waves, while keeping his basic assumption that light obeyed *deterministic laws*. Huygens's principle was the first dynamical or causal law to govern the propagation of a wave, as opposed to Newton's laws that governed the propagation of material particles.

The introduction of electric and magnetic fields by Faraday and Maxwell gave a tremendous boost to the wave theory, and the causal deterministic laws obeyed by those fields were formulated mathematically by Maxwell. Moreover, light waves were recognized as special cases of this electromagnetic field, and Maxwell's laws justified Huygens's principle. The price to pay by keeping the *paradigm of natural laws* was that the universe had to be regarded as a strange mixture of material particles and fields.

Physicists lived with this dual ontology even when an inconsistency was found between the two sets of laws that governed material particles and fields. This inconsistency, first clearly recognized by Einstein, was that the symmetries of the

J. Kouneiher (✉)
Nice and Cote d'Azur University, ESPE-Université de Nice Sophia Antipolis,
89, Av. GeorgesV, F-06064 Nice cedex 01, France
e-mail: joseph.kouneiher@unice.fr

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_13

laws of mechanics that governed material particles were not the same as the symmetries of the laws of the electromagnetic field. Einstein required that both symmetries should be the same, and asserted the primacy of fields over particles by requiring that the laws of mechanics should be modified so that they have the same Lorentz group of symmetries as the laws of the electromagnetic field: **This was the first time in the history of physics that symmetries took priority over laws** in the sense that the laws were modified to conform to the symmetries.

Moreover, the existence of universal symmetries for all the laws of physics enabled the construction of a physical geometry having the same symmetries, namely the Minkowski space-time. The idea of turning groups into basic building blocks for the geometric formulation of Physics is simply the natural result of pushing ahead the old usage of imposing the compatibility of observer in the same way Differential Geometry itself considers admissibility of local chart. The requirement of a definite structure in the set of observers, or atlas, delimites seriously the nature of physical laws in that they must be formulated in terms of say $GL(n;R)$-tensors, although this requirement is not restrictive enough so as to actually *predict* dynamical laws. However, the condition of having defined an associative composition law in a set of *active* transformations of a physical system really predicts in many cases its dynamics, and can accordingly be considered as a basic postulate.

The use of Riemannian geometry in relativity, during the first part of the XX century, ensured its status as an essential pillar of mathematics; the same applies to Hilbert space in quantum mechanics and the notion of symmetry, in the broadest sense, leading to the systematic use of group representation theory. At first, the use of Riemannian geometry by physicists was done so implicitly, but it was unavoidable that global problems would eventually be asked. This, of course, implies topology.

In addition to the role that topology plays in understanding the singularity problem in general relativity and in standard models, there is no doubt that the principle factor in topology rise in physics is due to the growing importance and supremacy[1] of gauge theories in physics. These theories have played a central role in the understanding and formulation of the theory of elementary particles. Theses theories can be described with the help from bundles of true topological objects. Such a journey can be characterised by the process of abstraction in physics, granting it greater performance in terms of describing nature, as Dirac described and predicted (Dirac 1938–39).

In this article we will build on this and go one step further and shall discuss the relationship maintained by physics and topology.[2] We want to explore thus the

[1] This of course is another matter and not our topic here. However, it is important to recognise that the significance of non-abelian gauge theories was revived with the proof given by the renormalisation of non-abeliennes gauge theories by Hooft (1971, 1994) and Hooft and Veltman (1972) (1971–72). This lead to the resurrection of previous papers on the subject and of the standard theory with the gauge group SU(2)xU(1) electromagnetic and weak interactions and also of quantum chronodynamics (QCD) with the gauge group SU(3), the preferred model in confining quarks that are considered responsible for strong interactions.

[2] Listing was the first to use the term Topology (actually he used *Topologie* for he wrote in German) in a letter to a friend in 1836 see Pont (1974), and Listing (1847, 1861). Remark that the term

cohomological aspect of physics, more precisely to show the cohomological foundations of the physics. One aspect will be the question concerning the quantization and the topological foundations of some theories.

Such a journey can be characterised by the process of abstraction in physics, granting it greater performance in terms of describing nature, as Dirac described and predicted (Dirac 1938–39)

> It seems likely that this process of increasing abstraction will continue in the future and that advance in physics is to be associated with a continual modification and generalisation of the axioms at the base of the mathematics rather than with a logical development of any one mathematical scheme on a fixed foundation.

Cohomology and Invariants

Homology and cohomology have emerged as the main instruments of algebraic topology. Their influence goes far beyond the geometric topology. They give the natural expressions in algebraic geometry, in differential geometry, and in algebraic theory of numbers. The cohomology depends on the local structure of the variety. It gives an account of forms, defines them. It connects the continuous to the discontinuous. But the most remarkable is universality.

It is in an article of 1895 that Poincaré (1895, 1901, 1902) defines for the first time, on differential manifolds, chains (or sub-varieties) which he qualifies as homologous (see Herreman (1997)). Its definition was somewhat imprecise, but the notion he used covered exactly the current acceptance: two closed chains are homologous if their difference is a boundary. However, Poincaré's text did not reveal the idea of Cohomology. The reason for this is that on a manifold, we can obtain completely cohomology from homology by Poincaré's duality. Roughly, Poincaré's duality connects the local statements of cohomology to the global statements of homology.[3]

Poincaré's work did not remain unnoticed, but was not considered until the 20s. During the next 20 years, different (co) homologous theories more or less general and more or less competing (singular, singular, Čech …). It was not until 1925 and the work of Emmy Noether to understand that these Topological numbers are only one facet of a richer structure: the homology group. Thanks to this contribution, a whole algebraic formalism develops, with among others the notion of differential

Topology was not commonly present in literature prior to the 1920s (instead we find the Latin terms *geometria situs* and *analysis situs*). The first important use of a topological argument could be argued to go back as far as Euler's solution (Euler 1736) to Konigsberg's bridge problem in 1736. During the elaboration of his answer, Euler had the idea to associate a graph to the problem giving birth to what we now call graph theory and so, with this example, he presented one of the first combinatorial topological problems. Today graph theory is a subject in its own right.

[3] The homology appeared as a redoubling of abstraction; The homological forms have doubled the algebra of the geometric forms which they enveloped. We can distinguish quite clearly two movements: a birth of geometry or algebra followed by homological stabilization. From a logical point of view, a geometric object and an homological object have the same nature.

complex in a greater generality. Naturally, this open breach allows the introduction, by different authors, in the following decade, of different types of homologies (and cohomologies).

The transition from homology to cohomology was initially an attempt to generalize Poincaré's duality. In a very surprising manner, the multiplicative structures on the differential manifolds can be translated very well into more abstract situations Cohomology (which homology does not allow at all).

There is a pleiad of cohomology leading to the same results. However, as its name implies, it is the dual of homology. The latter is based on the global properties of a variety. Some homological entities are known to all: the orientation of varieties, the connected component of a point in a topological space (an object in one piece). Another example: $H_1(\mathbb{T}^2)$, the set of homology classes of degree 1 of a torus of dimension 2. Imagine the surface of a tire on which we draw a line; This line makes turns and returns but, to finish it returns to its starting point. Two paths define the same *class of homology* when they can be continuously deformed one to the other, by contiguity. The set of all path classes will be $H_1(\mathbb{T}^2)$.

As soon as a coordinate system, made of a meridian (a) and a parallel (b) on the torus, is chosen, each path would be coded by two numbers. But other references (a'), (b') are equally valid. So, we obtain a discrete lattice isomorphic to $\mathbb{Z} \times \mathbb{Z}$, set of pairs of relative integers, from an infinite variability of the continuous results.

So the mathematical idea is that cohomology is equivalent to a limited form of homotopy, where a person will push and pull on circles and spheres like lassos until they hit an obstruction—a hole, namely homology, that detects holes in manifolds. It turns out that cohomology and homology have their roots in the rules for electrical circuits formulated by Kirchhoff in 1847 (Kirchhoff 1847).

The relativity group that exchanges the coordinate systems is $SL_2(\mathbb{Z})$, the set of 2×2 matrices with integer coefficients of determinant 1 (For example, $a' = 2a + b$, $b' = 3a + 2b$). This group does not act on the loops traced on $\mathbb{T}^2$, it acts only in homology. The rule of addition over $\mathbb{Z} \times \mathbb{Z}$ translates the concatenation of paths.

In general, there are more algebraic structures with cohomology. For example, the cohomology group of degree 1 of the torus $\mathbb{T}^2$, $H^1(\mathbb{T}^2)$ is constructed from the representations: at each loop, assigning a real or complex number, with "composition constraint": when a path is obtained by putting two paths end to end, its number must be the sum of the numbers of the two components, and *deformation constraint*: two close paths have the same number. The set of "numerical assignments" and the "constraints" form $H^1(T^2, \mathbb{R})$ or $H^1(T^2, \mathbb{C})$, depending on the quality of the numbers employed. It is now an affine plan. The intersection of the paths on the torus in pairs provides this plane with an area unit. A secondary geometry appeared, the group $SL_2(\mathbb{Z})$ is a distinguished part of its symmetries.

Homology, can be considered as a general technique in mathematics used to measure the difficulty that certain sequences of morphisms have to be exact. The idea is precisely to note that if an morphism α on a module M has $\alpha^2 = 0$, then $Im\ \alpha \in Ker\ \alpha$. Everything is in this remark! For then we can characterize the elements of M which are in the kernel of α (they are called cycles) but which are not in the image of α (they are called boundaries). So we form the quotient of modules

$$H(M, \alpha) = \frac{Ker\ \alpha}{Im\ \alpha}$$

Called the homology of M for α. This construction allows us to characterize the cycles that are not boundaries. It also allows to associate a sequence of abelian groups or modules with a mathematical object like a topological space or a group.

Often, we have a richer situation in which M is graduated in the form $M = \oplus M_n$ and α is a morphism of degree -1 (resp. $+1$) which is decomposed into the morphisms: $\alpha_n : M_n \longrightarrow M_{n-1}$ (resp. $\alpha_n : M_n \longrightarrow M_{n+1}$). Then we define the homology of M (respectively the cohomology) of M by a formula similar to the quotient above, with this time $Im\ \alpha_{n+1} \in Ker\ \alpha_n$ (resp. $Im\ \alpha_{n-1} \longrightarrow Ker\ \alpha_n$). We obtain Thus the homology modules $H_n(M; \alpha) = Ker\ \alpha_n / Im\ \alpha_{n+1}$ (respectively of cohomology $H^n(M; \alpha) = Ker\ \alpha_n / Im\ \alpha_{n-1}$).

The example of Poincaré in his work corresponds to taking for M_n the formal linear combinations with integer coefficients of oriented polyhedra of dimension n, and for α the operation which consists in taking the boundaries (in the geometric sense) of these polyhedra. These boundaries being themselves formal combinations of oriented polyhedra. We know that the boundaries of a boundaries is empty, which gives $\alpha^2 = 0$. Having fixed bases in the $\mathbb{Z}$-modules M_n (where possible), Such a module is identified with a power $\mathbb{Z}^{d_n}$, and the morphisms of $\mathbb{Z}$-modules α_n to matrices of size $d_n \times d_n$ whose entries are integers.

Now, in algebraic topology, we are interested to find the relationships between the cycles and boundaries in various dimensions of a topological space, chain complex and cochain complex are introduced as an algebraic means of this. Homological algebra includes thus the study of chain complexes in the abstract, without any reference to an underlying space. In this case, chain complexes are studied axiomatically as algebraic structures.

A chain complex is a sequence[4] of abelian groups or modules $\ldots M_2, M_1, M_0, M_{-1}, M_{-2}, \ldots$ connected by homomorphisms (called boundary operators or differentials) $\partial_n : M_n \longrightarrow M_{n-1}$ (sometimes we use $d_n : M_n \longrightarrow M_{n-1}$), such that the composition of any two consecutive maps is the zero map: $\partial_i \circ \partial_{i+1} = 0$ (or $d_n \circ d_{n+1} = 0$).

The elements of the kernel $ker\ \partial_i$ are called cycles. The elements of the image $Im\ \partial_{i+1}$ are called boundaries. Every boundary is a cycle. The homology groups of the complex M_* are then, by definition: $H_i(M_*; \partial_*) = ker\ \partial_i / Im\ \partial_{i+1}$.

A variant on the concept of chain complex is that of cochain complex. A cochain complex is a sequence of abelian groups or modules $\ldots, M^{-2}, M^{-1}, M^0, M^1, M^{+1}, \ldots$ connected by homomorphisms $\partial^n : M_n \longrightarrow M^{n+1}$ such that the composition of any two consecutive maps is the zero map: $\partial^{n+1} \circ \partial^n = 0$.

The index n in either M_n or M^n is referred to as the degree (or dimension). The only difference in the definitions of chain and cochain complexes is that, in chain complexes, the boundary operators decrease dimension, whereas in cochain complexes they increase dimension.

[4] We note sometimes this sequence as (M_*, ∂_*) (or M_*, d_*).

Differential forms provide a modern view of calculus. They also give you a start with algebraic topology in the sense that one can extract topological information about a manifold from its space of differential forms, it's called cohomology. For instance, de Rham cohomology is a good way to detect the "shape" of a domain. Indeed, we say that a vector field F in $\mathbb{R}^3$ is conservative if $F = \nabla f$ for some scalar-valued function f. This has natural applications in physics (e.g. electric fields). It's easy to see this happens if and only if line integrals of F are path independent, if and only if line integrals around closed loops vanish, etc. On a simply connected domain, F is conservative if and only if $\nabla \times F = 0$ (use the freshman version of Stokes' theorem). On a non-simply connected domain, this may fail (e.g. $\mathbb{R}^3$ minus a line). The extent to which it fails is of course the de Rham cohomology of the domain. Notice also that the differential forms are necessary for the development of cohomology theory in the context of manifolds without getting into the aspects which depend on metric notions.

The late 1930s and early 1940s witnessed the rise of homological algebra. This contributed largely to the emergence of notions of category and functor, ubiquitous notions in algebra and logic afterwards. Indeed, the tensor products of modules, the exact sequences and the functors Hom and Ext allowed remarkable progress both in calculating the homology group and to conceptualize what eventually become the homological algebra in Henri Cartan and Eilenberg works in the 1950s (Cartan 1956). Algebra topology, as its name indicates so correctly, proposes to study the topology of space by using algebraic concepts, such as homology groups, but also homotopy groups.

The invention of the cohomology of sheaves by Leray has had the same success in all the algebraic geometry (Leray 1950). Various generalizations have been devised: cohomology of groups (with surprising connections to geometry), equivariant cohomology, etal cohomology …which shows, if need be, that cohomological notions have spread widely in almost all mathematics, and even in theoretical physics.

Another contribution that later led to other paths is that of the axiomatization of simplicial homology by Eilenberg and Norman Steenrod in 1945 (Eilenberg 2011). This work allowed us, on the one hand, to show that some of the other homologies defined in this context are isomorphic to simplicial homology, and on the other hand it has generated more generalizations, *the generalized homologies*, of which the K-theory is only an example.

In parallel with these developments in the algebraic topology domain, the works in algebras has conceptualized different essential notions, such as the extension of abelian groups: an extension of the abelian group F by the abelian group H is an abelian group G containing F such that H is identified with the quotient G/F. In other words, we have an exact short sequence of abelian groups

$$0 \longrightarrow F \longrightarrow G \longrightarrow H \longrightarrow 0$$

An essential aspect of modern Mathematics and physics is the studies of the invariants. Indeed, classify the invariants became a central issue in physics and mathematics.

In the introduction to his book The Principles of Quantum Mechanics, Clarendon Press, Oxford, 1930, the young Dirac (1902–1984) wrote:

> The important things in the world appear as invariants . . . The things we are immediately aware of are the relations of these invariants to a certain frame of reference . . . The growth of the use of transformation theory, as applied first to relativity and later to the quantum theory, is the essence of the new method in theoretical physics. Dirac 1938–39.

A fundamental tool which will play an essential role to find and calculate those invariants is the Cohomology and homology. Indeed, Cohomology[5] plays a fundamental role in modern mathematics and physics (Kouneiher 2010). As we saw, Cohomology is an example of a local—global structural connection that permeates mathematics.

Cohomology is used in physics[6] (Bennequin 2010) to compute topological structure of gauge fields for instance, like the electromagnetic field in the Ahranov-Bohm effect. Here, the electron encircles a magnetic flux, which you can measure in the self interference pattern of the electron. That is amazing because the electron never actually passes through a magnetic field. Maxwell's equations tell us that the only interaction between the magnetic field and the electron is local—when the electron passes through the field. So what is going on?

The magnetic field is the curvature of a vector potential, which is a gauge connection and "lives" in the cohomology of the underlying manifold. Because of its cohomology, it is a real thing with measurable consequences in physics experiments. To understand this we will use Stoke's theorem. The field strength B is the curvature of the vector potential A, $B = dA$. The magnetic vector potential A is closed $[B = dA = 0]$ but not exact (i.e. A cannot be written as the curvature of another form everywhere i.e. A not equal dQ, though $A = dQ$ in patches and each patch contains a different Q, which must be patched together like a quilt). (Differential) cohomology is the group structure built up from the vector space of forms that are closed modulo the vector space of forms that are exact. The fact that idea manifests itself in physics was a huge surprise in the 50s, 60s and 70s. We still seem to be continually surprised by cohomology popping up in physics everywhere we look. That is a deeply intuitive way to see cohomology. The electron and the field interact locally in a trivial way yet sense non trivial global structure of the system, namely the encircled magnetic flux, which acts as an obstruction, when the path of the electron in spacetime forms a closed circle around it. Adding up local variations yields global data, even far away

[5] Usually the non vanishing of a cohomology class in algebra, geometry, and topology, express some "failure". Indeed, Often in math you wish something were true, but in general it is not. But, the quantification of how badly it fails, help us towards finding out a more precise statement that holds generally. The size (or dimension) of the corresponding cohomology group is a measurement of how many ways things can go wrong. If it is nice or if you can understand it completely, then you may be able to analyze all the possible failure modes exhaustively, and use that to prove something interesting. This idea can be applied in an amazingly broad set of contexts. This explain in some way the use of Cohomology to describe quantization.

[6] In deciding to extend the concepts of homology and cohomology outside the ideal world of mathematics, We are led to accept the use of some analogies. In physics as in mathematics, a universal concept of homological object does not exist.

from where the global data is most manifest, namely at the obstructions and holes. The magnetic flux here plays the role of the hole or obstruction.

A single electron may encircle a single flux in quantum physics because the electron can be in two places at once—it can take the path to the left of the flux and to the right, at the same time, and meet itself at the other side. The result is a full circle. Quantum mechanics also says that the electron picks up a phase from the vector potential, which physically manifests itself in self interference patterns of the electron. But why would an interference pattern emerge when the electron only travels through space in regions where the magnetic field is zero?

The vector potential adds up along the path(s) in the phase of the electron interacting with the electromagnetic gauge field. That is, around a path encircling the magnetic flux, you add up A, which by Stoke's theorem equals adding B in the enclosed disk, which is non zero because of the enclosed magnetic flux.

This helps to explain why the Maxwell equations in electrodynamics are closely related to cohomology, namely, de Rham cohomology based on Cartan's calculus for differential forms and the corresponding Hodge duality on the Minkowski space. Since the Standard Model in particle physics is obtained from the Maxwell equations by replacing the commutative gauge group U(1) with the noncommutative gauge group $U(1) \times SU(2) \times SU(3)$, it should come as no great surprise that de Rham cohomology also plays a key role in the Standard Model in particle physics via the theory of characteristic classes (e.g., Chern classes which were invented by Shing-Shen Chern in 1945 in order to generalize the Gauss–Bonnet theorem for two-dimensional manifolds to higher dimensions).

It is very clear now that the gauge-theoretical formulation of modern physics is closely related to important long-term developments in mathematics pioneered by Gauss, Riemann, Poincaré and Hilbert, as well as Grassmann, Lie, Klein, Cayley, Elie Cartan and Weyl. The prototype of a gauge theory in physics is Maxwells theory of electromagnetism. The Standard Model in particle physics is based on the principle of local symmetry. In contrast to Maxwells theory of electromagnetism, the gauge group of the Standard Model in particle physics is a noncommutative Lie group. This generates additional interaction forces which are mathematically described by Lie brackets.

We also emphasize the methods of invariant theory. In terms of physics, different observers measure different values in their experiments. However, physics does not depend on the choice of observers. Therefore, one needs both an invariant approach and the passage to coordinate systems which correspond to the observers, as emphasized by Einstein in the theory of general relativity and by Dirac in quantum mechanics. The appropriate mathematical tool is provided by invariant theory.

The Idea of Motion Mechanics

The concept of an object's motion, i.e. the change of its position in space over time, is intimately linked with two other concepts in physics: space and time. Therefore the three concepts of time, space and motion must be considered together.

Recently in history, discussions on the nature of space and time, have been dominated by two points of view: relational concepts and absolute. From these view points, space can be seen as:

(i) a positional quality of the world of material objects, or
(ii) a container in which live all material objects.

The same thing can be done with time: it can be seen as:

(i) a well ordered quality of the world of material events;
(ii) a container of all material events.

In truth, most natural philosophers have adopted either the absolute point of view (like Newton) or the relational view (like Leibniz and Huygens) if not both.

We can note that there is a temporal order common to both points of view, before the advance of the theory of special relativity, and that it was either inherent to the nature of events themselves or due to their immersion in the temporal continuum, presumed as unique and non problematic, for successive events produced in the same place. Concerning events considered to be happening in different places, we need a convention to allow for the introduction of global time.

The possibility to define a unique temporal order lies with the ability to introduce a universal relation or simultaneous absolute, between these events, to different positions in different places. This relation is one of equivalence (simultaneous relation) for it divides the events into equivalence classes of simultaneous events. We can then use this simultaneous relation to define simultaneous events as happening at the same global time. In this way we transfer the local temporal order of events to different places. In consequence, in Galilean-Newtonian physics, the concepts of local time and global time merge into the concept of an absolute and universal time. We know today, that Einstein included in the notion of absolute or universal time the source of the incompatibility of Newtonian mechanic's principle of relativity with Maxwell's electrodynamics. From this fact, he developed a new type of kinematics, in which temporal intervals "local and global", as well as spatial intervals, are no longer universal or absolute: since there is no longer the notion of an absolute or universal time, clocks can not be perfectly synchronized, and the concept of simultaneous events must be approached with precaution. Even though arbitrariness (in conventional terms) enters in all definitions of global time and of (relative) simultaneous events we must have, for each inertial referential, the choice of the best convention allowing the expressions of nature's laws to be simplified. This convention, applied to each inertial referential of reference, implies that each referential has its own relative simultaneousness, using a system of equivalent clocks at rest in each referential and the speed of light in a vacuum as a signal for synchronization. The global time interval between two events now depends on the referential, in the same way as the spatial interval between two (non simultaneous) events in the Galilean-Newtonian kinematic. Now that the simultaneous events are relative, the spatial interval between two events is also relative. Despite the fact that in special relativity we have more symmetry between properties of space and time, there still remains a big difference between space and time. Effectively, (proper) local time that flows between two events, depends on the path (type time) between them; but here the shorter the path, the longer the corresponding time, the shorter the distance.

The concepts of Galilean-Newtonian time (and space) as well as in special relativity are founded on the fact that the structure of the kinematic referential is independent of the dynamic process. But in all cases, the kinematic structure is established once and for all by a symmetry group of space and time transformations, a group in which the dynamic laws of all closed systems stay invariant: in a Galilean-Newtonian case, this group is the Galilee's inhomogeneous group; in the case of special relativity, it is of the Lorentz's inhomogeneous group (also called Poincare's group).

In special relativity, we look at space-time transformations using a 4-dimentional formalism, in which the coordinates of space and time are represented by a 4-dimentional space-time. In this formalism, a point in space is represented by a universe line/worldline: a one dimensional curve in space-time representing the history of this point through time. Done in such a way that a three dimensional space-time is represented by the convergence of all these universe lines (without intersection) covering all of space-time or fibration of space-time. A moment in time is represented by the unique intersection of a hyperplane surface with all the congruence curves. We reasoned that a family of these hyperplanes (without intersection) on all of space-time represents a (global) time variable or leaves of space-time. Proper time along all universe lines is local time associated with a series of events along that universe line.

In special relativity's 4-dimentional space-time, each inertial referential is represented by a different leaf in space-time in hyperplanes at the same instants, as well as the three corresponding spaces defined by the convergence (fibreation) of parallel lines of type time that are "pseudo-orthogonal" to hyperplanes. Lorentz's special transformations (boosts), which express the relation between two inertial referentials (each represented by a sheet and a fibration), are represented by a "pseudo-rotation" that transforms a sheet and fibration into another.

A closer examination of the kinematic structure of space-time shows that it is made up of two distinct yet interconnected structures: one chrono-geometric and the other inertial. The chrono-geometric structure models mathematically spatial geometry and the measure of global time, while the inertial structure models mathematically the behaviour of physical objects undergoing no exterior forces, i.e. behaviour characterized by Newton's first law (law of inertia). The structure is described by the field of refined connection. These two structures obey certain conditions of compatibility. In other words, these conditions assure that the set of free falling particles can be used to construct clocks and rules for measurement.

The inertial structure is common to both Galilean-Newtonian kinematics and special relativity; the inertial structure associated with a refined (linear) space, allows us to define straight parallel lines, parallel hyperplanes and the equality of parallel vectors of space-chrono-geometrical structure consists of a leafing by simultaneous hyperplanes (representing all events of the same absolute time) plus a degenerated 4-dimentions metric system (of 3rd rank) representing Euclidian geometry, which applies itself to all inertial referential. Special relativity's (or minkowskienne) chrono-geometry consists of a non-degraded pseudo-metric with 4-dimentions of signature 2.

In general relativity, the two parts of kinematic structure, chrono-geometric and inertial, lose their formally fixed character: both become dynamic structures. There are two reasons for this: general relativity is a gravitational theory and gravity makes the inertial structure more dynamic (this equally applies to the Newtonian theory of gravity from a four-dimensional view point). Also, in general relativity, the chono-geometric structure is uniquely linked to the inertial structure, so if the first becomes dynamic, the second must do the same.

As we have seen, the concept of inertial structure is founded on the behaviour of bodies in free fall, but the presence of gravity cancels out this concept of motion. Let's begin first by noting that all non-gravitational forces (electric, magnetic) can be cancelled or hidden, but that gravity is universal and so can't be cancelled or hidden. Secondly, it has the same effect on the motion of all bodies. This would not be a problem if it were possible to specify a class of inertial referential independently from the concept of free movement. But inertial referential can't be defined independently from inertial motion, which is a free movement. The only way to solve this problem is to admit that we can't distinguish, in absolute terms, the effects of inertia and gravity on the movement of bodies. There exists a unique inertial-gravitational field such as "free falls" which are motions of a body subjected only to this field.

To take into account gravity in Galilean-Newtonian inertial structures, like we have seen previously, we must generalize the inertial-gravitational structure. But due to the inclusion of gravity, this structure is no longer fixed, but is subject to dynamic field equations specifying how the presence of matter affects the inertial-gravitational field. In the Newtonian theory of gravitation, compatibility conditions between chrono-geometrical kinematical structures remain fixed and the interial-gravitational structure stays dynamic and still viable, but it now only serve to fix the inertial-gravitational structure when the chrono-geometrical structure is given. It gives certain flexibility in the choice of the former to impose a 4-dimentional equivaliance of Newton's law of gravitational attraction on the inertial-gravitational structure. Contrastingly, in special relativity, relations of compatibility between chrono-geometrical and inertial structures are completely restrictive: the chrono-geometrical structure determines, in a unique fashion, inertial structure. When we incorporate gravity and make it more dynamic, we face a choice: either abandon the unique relation between chrono-geometrical structures and inertial structures or preserve it, which also makes chronogeometrics more dynamic.

In general relativity, this last approach is chosen. The metric tensor field leads to both chrono-geometrical structures and inertial-gravitational structures and follows a set of field equations linking the curvature tensor to a tensor describing the sources of gravitational interactions, called energy-momentum tensors. These field equations for the refined tensor curvature resemble 4-dimentional Newtonian equations. But in the case of general relativity, as previously specified, the refined structure, which includes the tensor curve, is completely determined by the metric tensor field. Also in general relativity, all depends on the metric system.

In general relativity, no space-time structure has yet to be chosen, and there exists no preferred symmetrical kinematic group to represent the symmetries that retrain punctual class transformations possible for the underlying 4-dimentional space-time. (In general relativity we can't call this space mathematical for there are no physical properties before the metrical tensor is selected).

In physics this means that a 4-dimensional topological space does not suffice, we need tensor fields or geometric objects in the space on which different differential operations can be applied. To carry out these operations independently from the coordinate system choice, we need a differential structure on the underlying topological space. We are lead to consider a variety of 4-dimensional differentials as sub-jacent mathematical structures that have the diffeomorphic group (differentiable homeomorphisms) as their symmetry group. We consider the general covariance of all field equations on these differential varieties; the invariance beneath the diffeomorphic group. This request constitutes an important aspect of what is called general covariance of the entire theory.

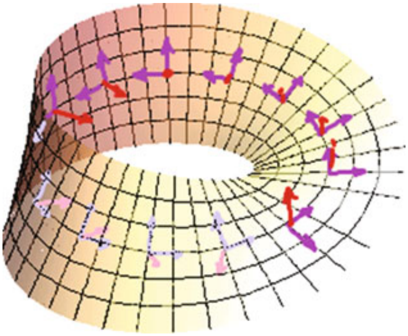
The Idea of Fiber Bundle

A Fibre bundle is a way to construct 'products' of topological spaces. They are useful in that you can build up more complicated spaces from simpler spaces. The notion of fiber bundle was introduced in the 1930s. It is one of the most important notion in Topology. The words fiber (Faser in German) and fiber space (gefaserter Raum) appeared for the first time in a paper by Seifert in 1932 (Seifert 1932), but his definitions are limited to a very special case. The main difference from the present day conception of a fiber space, however, was that for Seifert what is now called the base space (topological space) of a fiber (topological) space E was not part of the structure, but derived from it as a quotient space of E. The first definition of fiber space is given by Hassler Whitney in 1935 (Whitney 1935) under the name sphere space, but in 1940 Whitney changed the name to sphere bundle.[7] The theory of fibered spaces, of which vector bundles, principal bundles, topological fibrations and fibered

[7] W. S. Massey (1920) listed five definitions of fibre space (Massey 1999): (a) fibre bundles in the American sense (Steenrod 1951); (b) fibre spaces in the sense of Ehresmann and Feldbau (Ehresmann 1934; Feldbau 1939); (c) fibre spaces as defined by the French school (Cartan 1956); (d) fibre spaces in the sense of Hurewicz and Steenrod (Steenrod 1951), and (e) fibre spaces in the sense of Serre (1951). Each of these competing definitions developed out of a mix of examples and problems of interest to the research community in topology, often marked by a national character. We will consider the origins of each of these strands and the relations among them.

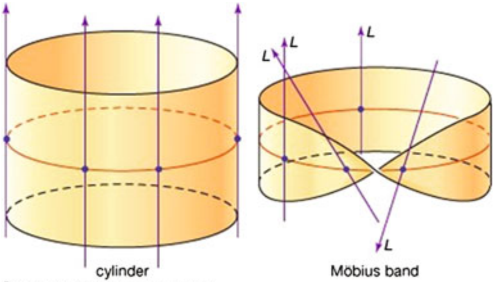
manifolds are a special case, is attributed to Seifert (1932), Hopf, Feldbau (1939), Whitney (1935; 1940), Steenrod (1951), Ehresmann (1934), Ehresmann and Feldbau (1941), Ehresmann (1983), Serre (1951), and others. Fiber bundles became their own object of study in the period 1935–1940. The first general definition appeared in the works of Hassler Whitney (1935). Whitney came to the general definition of a fiber bundle from his study of a more particular notion of a sphere bundle, that is a fiber bundle whose fiber is a sphere of arbitrary dimension.

In physics, they are used to represent Gauge Theories and ‘constrained vector fields’. A Möbius strip is a good example of fiber bundle:



As you walk along the Möbius Strip, at each point p you have a velocity, which is simply a vector tangent to the strip, e.g. a purple vector in the following picture: If you have a constant velocity (e.g. the length of the purple vector doesn't change as you move through space), you will find that once you go around the circle once, your velocity will have ‘flipped’ (relative to your initial velocity vector). This represents the non-orientability of the Möbius Strip.

In the above description, we've assumed that the Möbius Strip is an ‘irreducible object’ itself (e.g. it is a manifold). However, we can construct a Möbius Strip by gluing a little line (called a fibre) to each point on a circle, as long as we ‘flip’ the fibers once we have gone half way along the circle. If we don't flip the fibers, we get a cylinder:



© 2006 Encyclopædia Britannica, Inc.

The Möbius Strip is the simplest, non-trivial example of a Fibre Bundle. It is a way to attach some type of object at every point of a topological space. In the above case, our topological space was $X = S^1$, the circle, while the fibers were simply line intervals $Y = [0, 1]$. Note that the above represent two different 'products' of the spaces $X \times Y$—One, the cylinder, is simply the Cartesian product (pairs of elements of X, Y), while the other has a 'twist' that is an extra constraint on how we construct a product.

Example Vector Bundles: Given a manifold M we would like to attach a k-dimensional vector space to each point in such a way that locally on the manifold, the vector spaces look like the product space $M \times \mathbb{R}^n$. A priori, these vector spaces do not really talk to each other in the sense that there is not a well-defined way to add and subtract vectors that lie in different vector spaces, i.e. the vector spaces attached to different points on the manifold. So, we introduce a connection on the bundle to be able to subtract points near each other on the manifold. This allows us to take derivatives, and defines the covariant derivative $D_\mu = \partial\mu + A_\mu$, where the matrices A_μ are the Christoffel symbols (or gauge fields, etc.) in a certain coordinate system on the manifold and the bundle's frame. In physics literature, the fields A_μ are often introduced as a fudge factor meant to impose the condition of gauge covariance on D_μ. Covariance means the transformation $\psi \longrightarrow U\psi$ and the corresponding $D_\mu\psi \longrightarrow UD_\mu\psi$, which in the bundle language corresponds to changing coordinates on the bundle.

One way to think of fiber bundles is that they are the data to globally twist functions (on spacetime, say) where global twist is much in the sense of global anomaly and the like, namely an effect visible on topologically nontrivial spaces when moving around non-contractible cycles. The concept of monodromy which may be more familiar to physicists is closely related: monodromy is something exhibited by a connection on a bundle and specifically by a flat bundle. For a discrete structure group (gauge group) every bundle is flat, and in this case non-trivial bundles and non-trivial monodromy come down to essentially the same thing (see also at local system).

More explicitly, suppose X denotes spacetime and F denotes some space that one wants to map into. For instance F might be the complex numbers and a free scalar field would be a function $X \to F$. For the following it is useful to talk about functions a bit more indirectly: observe that the projection $F \times X \to X$ from the product of F with X down to X is such that a section of this map is precisely a function $X \to F$. We think of $X \times F \to X$ as encoding the fact that there is one copy of F associated with each point of X, and think of a function with values in F as something that, of course, takes values in F over each point of X. One says that $X \times F \to X$ is the trivial F-fiber bundle over X. If F is a vector space and all transition functions are linear maps on the fibers, then one calls this a vector bundle.

The point being that more generally we may add a global twist to the F-valued functions by making the space F vary to some degree as we move along X. For a fiber bundle one requires that it doesnt change much: in fact the word fiber in fiber bundle refers to the fact that all fibers (over all points of X) are equivalent. But the point is that any F may be equivalent to itself in more than one way (it may have automorphisms), and this allows non-trivial global structure even though all fibers look alike.

In this sense, a general F-fiber bundle on some X is defined to be a space P equipped with a map $P \to X$ to the base space X (e.g. to spacetime), such that locally it looks like the trivial F-fiber bundle, up to equivalence. To say this more technically: $P \to X$ is called an F-fiber bundle if there exists a cover (open cover) of X by patches (e.g. coordinate charts!) $U_i \to X$ for some index set I, such that for each patch U_i (with $i \in I$) there exists a fiberwise equivalence between the restriction $P|_{U_i}$ of $P \to U_i$, and the trivial F-fiber bundle $F \times U_i \to U_i$ over the patch U_i.

To say this again in terms of sections: this means that a section of P is locally on each (coordinate) patch U_i simply an F-valued function, but when we change patches (change coordinates) then there may be a non-trivial gauge transformation that relates the values of the function on one patch to that on another patch, where they overlap.

Even if this may seem a bit roundabout on first sight, this is actually something at the very heart of modern physics, in that it embodies the two central principles of modern physics, namely

- the principle of locality;
- the gauge principle.

The first roughly says that every global phenomenon in physics must come from local data. In the above discussion this means that any globally F-valued thing on spacetime X must come from just F-valued functions on local (coordinate) charts $U_i \hookrightarrow X$ of spacetime. But—and this is key now, second, the gauge principle says that we may never strictly identify any two phenomena in physics (neither locally nor globally) but we must always ask instead for gauge transformations connecting two maybe seemingly different phenomena. Hence combining the gauge principle with the locality principle means that if an F-valued something on spacetime is locally given by plain F-valued functions, then it should be globally given by gluing these F-valued functions together not by identification but by gauge equivalence. The result may be a structure that has global twists, and the nature of these global twists is precisely what an F-fiber bundle embodies.

Fiber Bundles and the New Conception of the Classical Idea of Motion

The simplest bundle structures found in physics are those of Galilean and Newtonian mechanics: here the base or independent variable that determines the change of all other dynamic characteristics of an object in motion is time. The base of mechanics is one-dimensional whilst the fibers are three dimensional: they are the "fluent" spatial coordinates of a point in motion. The fibers are a set of dependent variables where the variations characterize mechanical processes whatever complexity and which are interpreted as defined spatial translations of certain objects relative to others.

This purely descriptive division of dependent and independent variables of mechanics serves as a base to put forward important and profound theoretical structures where forces are the only cause of dynamic state change in an object in motion. Classical dynamic laws (Newton's equations) applied to the second derivatives of each mobile point's coordinates allows us to discover the forces that act on the point—the new elements of physical reality that are important in the comprehension of theoretical mechanical motion. Using the law of a force's action (given through experience or a certain theory) and the results concerning the initial state of the mechanical motion of bodies, we can calculate without ambiguity (or guess) the smallest details and the results of all the point's interconnected motions of whatever complexity (in principal at least, using powerful calculators to solve concrete dynamic equations).

From Topological point of view, the simplest structures are the bundle spaces of mechanics for uniform motion (or constant speed) or uniformly accelerated. In this case the fibrations are reducible to a trivial (Cartesian) space of parameters: the distance covered by an object moving uniformly is always equal to the "trivial" product of the speed times the time, and if the body has a uniformly accelerated motion, its speed is equal to the acceleration times the time. However, in this last case, the bundle space of the distance covered by an object with uniformly accelerated motion is structurally more complex: it ceases to be a trivial global bundle, becoming only trivial locally. The complexity of motion mechanic's quality, due to the transition between motion at constant speed and motion at constant acceleration, reveals an abrupt quantative jump in the complexity of bundle space describing them as: the element (or differential) of distance covered by an object with uniform acceleration is only equal locally, in the neighbouring infenitesimal space around each point, to the product of a variable speed (increasing or decreasing) and the element (differential) of time. To obtain the whole journey covered by a body with uniform acceleration, we must be capable to add all theses infinitesimal products, i.e. calculate the integrals of these differentials within certain limits.

Hence, classical analysis in its totality seems, from a modern topological point of view, to be a systematic method to "calculate" all (extensive) variable quantities in trivial local bundle spaces that had been formerly studied by means of diverse theoretical problems of analytical mechanics. Additionally, the structure of modern algebraic topology allows us to understand the singular role played by systematic theoretical construction of dynamics for inertial referentials of reference that single-handedly allowed us to correctly formulate these basic laws (Newton's first, second and third laws of motion). An interesting characteristic of inertial referentials is that the second derivatives of their relative motions to the time coordinates is nil; referentials are free in relation to the action of external accelerations and their constant speeds of mutual relative translation differ from each other.

Modern algebraic topology considers that the second derivative in function with the temporal coordinates equals zero as a sort of simplicity or topological triviality, i.e. dynamic acyclicity of mechanic's initial (dynamic) complex. Mathematically, it is characterized precisely as the result of the iterative application of abstractedly defined operators to the dynamical system's coordinated associated to the objects, with dynamic boarder conditions (dynamic differentials) equal to zero. From this aspect, all forces modifying the mechanic state of objects in motion acquire a totally new mathematical interpretation—being topological (or cohomological) measures of deviation for the dynamic systems studied with respects to inertial motion (corresponding to the highest state of topologic-dynamic triviality). In terms of algebraic topology even the quantitative magnitudes of forces are not always essential in predicting the results of motion, particularly for its global nature, in comparison to the generalized geometric (topologic) characteristics that determine the place and degree of deviation of a mechanical system from a defined inertial state as the state of motion with the greatest cohomological simplicity: dynamic acyclicity.

A large number of studies on mechanic's topological structure, made these last few years, show that these structures play a decisive role, including in the solution to concrete dynamic problems. All seems to be linked to a simple variety, varieties with the simplest (simplex) geometric dynamic structures such as the repeated application of operations with dynamic (dynamic differential) boarders that give zero.

Newton's first law of mechanics expressed as such and leads to the necessary presence, for all dynamic systems, of kinematic simplexes that are physically interpreted as the most natural state of an inertial motion, topologically the most elementary and undisturbed by forces. The second law interprets forces as cohomological measures of behaviour deviation of systems in motion with respects to "inertial states" as the maximum of topological-dynamical simplicity. The third law needs global anti-symmetry of force action: deviation with respects to inertial motion states don't appear alone, but always accompanied by similar deviations with respects to inertial motion states of objects with an opposite sign. A similar topological interpretation can be given to Lagrange's or Hamilton's equations of dynamics even though it requires additional topological concepts and complex mathematics.

We shall now offer a brief methodical analysis of the basic concepts and laws from another fundamental theory of physics—the classical theory of electromagnetic fields. Proof that this theory is also based on simplistic structures and probably on one of the most interesting results of current theoretical physics. The principle focus of classical electromagnetism is on the behaviour of force fields in space and time as being and elementary part of physical reality revealed by mechanics.

The main laws describing the behaviour of Maxwell's equations are closely interconnected: as in mechanics, thanks to bundle space structure, the foundation (the set of parameters vary independently) is extended to field theory: the base is represented here by set of 4-dimensional points of a space-plus-time continuum (not simply the time of mechanics). Like with fibre (all independent fundamental physical variables of this base), it is represented by force vectors acting on a single charge (or current) in each point at a given moment (meaning, electrical and magnetic field amplitudes). Field equations for certain combinations of force field derivatives, with respects to coordinates and time, allow for the discovery, in this bundle space, of new, more profound and fundamentally more significant elements on charges and currents generating force fields in physical reality.

Bundle space of field theory has, as with mechanics, a universal property that is applicable to all solutions, a property of great importance for methodology: when the laws of charge and current motion in space and time are known, at least empirically, the local structure of field equations allow for the calculation (in principle with the level of precision required) of special distribution of all physical field combinations as well as their temporal dependence. Electromagnetic fields on bundle space have properties that have spatio-temporal symmetry (like the invariance of field quantities with respects to translations and rotations in a 4-dimensional space-plus-time continuum), which leads, according to the Hamel-Noether theorem, to a large number of conservation laws for energy, impulsion, angular momentum, etc.

The recent discoveries of bundle space's simplistic structures in field theory are of great interest for the defined isomorphic topology of dynamic structures in mechanics and electromagnetic. The methodological study of physic's foundations has to this day failed to understand the mysterious fact that electromagnetic equations can be formulated equivalently (and in this case solved quicker) in terms of auxiliary quantities special to 4 dimensions: electromagnetic potentials. Its means of introduction is nearly analogue, from an algebraic topological point of view, to the means of introduction for reference inertial referentials in mechanics: it rises from a defined class of 4 dimensional potentials that is equally simplistic. The requirement of the potential's gauge-invariance to be satisfied, leads us always to select the latter, as much like the inertial referentials but in a more arbitrary way, even if the additive class is defined. All concrete 4-dimensional potentials of the same class differ from one another by 4-dimensional vector, whose components satisfy a wave equation without a source (or that the right hand side equals zero). The topological signification of this condition is analogue to the topological signification of the inertial motion state in mechanics: it identifies the simplistic structures in electrodynamic bundle space. The wave equation's left had side (Alembert's operator) can be represented as the result of a repeated application of field potentials of a certain abstract operator with 4 dimensions having dynamic boarder conditions (of 4 dimensions). The fact that it is zero for arbitrary additional quantities (of 4 dimensions) changes the gauge of field potentials and the simplistic character signature.

Like with bundle space of mechanics, the fiber bundle space of electrodynamics reveals the universal dynamic states of simplistic acyclic topological structures of extremely simple dynamics that characterize the field propagation in the case of simple dynamics in absence of currents and charges. In electro dynamics, these states appear like "standards" of particularly simple dynamic configurations: Maxwell's equations are interpreted in a simplistic field theory like topological (cohomological) measures of deviation with respects to the simplistic "standard" in the behaviour of the analyzed electrodynamic systems. In some ways, this new interpretation is simpler and more "visual" than the formulation (differential or integral) of Maxwell's equations, for they accentuate the purely quantitive (topological) characteristics of dynamic electromagnetic system behaviour. Also, in this type of system, the magnetic force lines have neither beginning nor end; they always appear as concentric circles perpendicular to electrical fields and currents that vary with time, wrapping itself around the former using the common right hand rule. It is the same for the electrical field force lines, they begin (or end) in the electrical charge from which they originate, or they appear as concentric circles perpendicular to all magnetic fields that vary with time.

The common form of Maxwell's equations can be obtained from these purely qualitative (topological) formulations with the help of De Rham's profound theorems, which establish an isomorphism (under defined conditions) between the homological (and cohomological) algebraic groups and the differential groups given by derivatives (or integrals). Coulomb's law, for example, appears much like a consequence of the simple qualitative fact on which the force lines of the static electric field begin or end only at the level of the charge. The electromagnetic analogue of this state of relative mechanical rest can be seen in the state where the field has 4 dimensional potentials that are identical everywhere and invariant which satisfies trivially the right hand field equation equalling zero.

Cohomology and Quantization

One of the most important and vital problems of modern theoretical physics is a study of the basic dynamic structures of the fiber spaces of quantum theory that would be as thorough as in the case of mechanics and electrodynamics.

Quantization essentially was and still motivated by experiment and observations, and curiously enough it is quantum mechanics and quantum field theory which account correctly for experimental observations where classical mechanics and classical field theory gives no answer or incorrect answers. Historically the *ultraviolet catastrophe* represent an important example. Indeed, the paradoxal aspect predicted by classical statistical mechanics, was explained and corrected by quantum mechanics.

However, independently from the experimental input, do we have a formal mathematical reasons and motivations to privileged quantum mechanics rather classical mechanics. Could we have been led to quantum mechanics by just pondering the mathematical formalism of classical mechanics? The following spells out an argument to this effect. It will work for readers with a background in modern mathematics, notably in Lie theory, and with an understanding of the formalization of classical/prequantum mechanics in terms of symplectic geometry. See next Sect. "Coho mology and Quantization" for more details.

Let X be the classical phase space with symplectic form ω. The actual line bundle you have to consider is the prequantum line bundle, which is a line bundle over the phase space equipped with a $U(1)$-connection ∇ such that the curvature of the connection is the symplectic form ω. If you have now a choice of polarization on the phase space, i.e. a split of the coordinates into positions q and momenta p, the actual wave functions of quantum mechanics are the sections of the prequantum line bundle which only depend on position, i.e. are constant on surfaces of constant q.[8]

If you're wondering where time evolution is here: We have only said what the space of states is. The time evolution is of course given by the unitary operator generated by the Hamiltonian (which is encoded in ω) acting on this space of states. Generically, the action of the quantum version of a phase space observable f is given by the covariant derivative (defined by the connection ∇) of the section along the

[8] Also, they should be square integrable in an appropriate sense. This requires discussing how we equip the space of sections with a Hilbert space structure and what measure we integrate against.

vector field X_f associated to f by $df(.) = \omega(X_f, .)$. More precisely, f acts on a section ψ as

$$\hat{f}(\psi) = i\nabla_{X_f}\psi + f.\psi \tag{1}$$

where $\hat{f}$ is now the (pre-)quantum operator associated to f.

Now, lastly, for the (non-)triviality of the bundle: Nothing forbids the prequantum line bundle from being non-trivial, but the classical phase space needs to have nontrivial cohomology for that—complex line bundles are completely characterized by their first Chern class, which is an element in $H^2(X, \mathbb{Z})$. Thus, in most situations you will encounter in typical quantum mechanics applications (where the phase space is just a symplectic vector space, and hence in particular contractible, so most cohomologies vanish), the prequantum line bundle will indeed be trivial.

Indeed, as the discussion shows, quantization as such, if done non-perturbatively, is all about lifting differential form data to line bundle data, this is called the prequantum line bundle which exists over any globally quantizable phase spaces and controls all of its quantum theory. It reflects itself in many central extensions that govern quantum physics, such as the Heisenberg group central extension of the Hamiltonian translation and generally and crucially the quantomorphism group central extension of the Hamiltonian diffeomorphisms of phase space. All these central extensions are non-trivial fiber bundles, and the quantum in quantization to a large extent a reference to the discrete (quantized) characteristic classes of these bundles. One can indeed understand quantization as such as the lift of infinitesimal classical differential form data to global bundle data.

Quantization as Central Extension

In this section we will exhibit an argument which spell out such possibility and develop in more details the topological foundation of the quantization (Schreiber (see (https://physics.stackexchange.com/users/5603/urs-schreiber)). Recall that a system of classical mechanics/prequantum mechanics is a phase space, formalized as a symplectic manifold (X, ω) and we know that a symplectic manifold is in particular a *Poisson manifold*, which means that the algebra of functions on phase space X, hence the algebra of classical observables, is canonically equipped with a compatible Lie bracket: *the Poisson bracket*. This Lie bracket is what controls dynamics in classical mechanics. For instance if $H \in C^\infty(X)$ is the function on phase space which is interpreted as assigning to each configuration of the system its energy—the Hamiltonian function—then the Poisson bracket with H yields the infinitesimal time evolution of the system: the Hamilton's differential equations.

Here, we will be concerned by the infinitesimal aspect and nature of the Poisson bracket. Generally, every *Lie algebra* $\mathfrak{g}$, can be regarded as the infinitesimal approximation of a globally defined object, the corresponding *Lie group* (or generally smooth group) G. One also says that G is a Lie integration of $\mathfrak{g}$ and that $\mathfrak{g}$ is the Lie differentiation of G. Therefore, it is very natural to ask: Since the observables in classical mechanics form a Lie algebra under Poisson bracket, what then is the corresponding Lie group?

Surprisingly, the answer to this question is not a widely acknowledged fact that we could find in the basic educational textbooks: *the Lie group which integrates the Poisson bracket is the quantomorphism group*, an object that seamlessly leads to the quantum mechanics of the system. Notice that Lie integration is not quite unique. There may be different global Lie group objects with the same Lie algebra.

The simplest example and which is of central importance for the issue of quantization, it is the Lie integration of the abelian line Lie algebra $\mathbb{R}$. This has essentially two different Lie groups associated with it: the simply connected translation group, which is just $\mathbb{R}$ itself again, equipped with its canonical additive abelian group structure, and the discrete quotient of this by the group of integers, which is the circle group

$$U(1) = \mathbb{R}/\mathbb{Z}\,.$$

This circle group structure is induced by the discrete and hence quantized nature of the integers. This can be traced back to the heart of what is *quantized* about quantum mechanics. Namely, one finds that the Poisson bracket Lie algebra $\mathfrak{poiss}(X, \omega)$ of the classical observables on phase space is (for X a connected manifold) a Lie algebra extension of the Lie algebra $\mathfrak{ham}(X)$ of Hamiltonian vector fields on X by the line Lie algebra:

$$\mathbb{R} \longrightarrow \mathfrak{poiss}(X, \omega) \longrightarrow \mathfrak{ham}(X)\,.$$

This means that under Lie integration the Poisson bracket turns into an central extension of the group of Hamiltonian symplectomorphisms of (X, ω). And this is true either it is the fairly trivial non-compact extension by $\mathbb{R}$, or it is the interesting central extension by the circle group $U(1)$. For this non-trivial Lie integration to exist, (X, ω) it needs to satisfy a quantization condition coded by the existence of a prequantum line bundle. If so, then this $U(1)$-central extension of the group $Ham(X, \omega)$ of Hamiltonian symplectomorphisms exists and is called the quantomorphism group $QuantMorph(X, \omega)$:

$$U(1) \longrightarrow QuantMorph(X, \omega) \longrightarrow Ham(X, \omega)\,.$$

Unfortunately, this group is not very well known though it contains Heisenberg group as a small subgroup. More precisely, whenever (X, ω) itself has a compatible group structure, notably if (X, ω) is just a symplectic vector space (regarded as a group under addition of vectors), its subgroup of the quantomorphism group which covers the (left) action of phase space (X, ω) on itself would be Heisenberg group $Heis(X, \omega)$, which in turn is a $U(1)$-central extension of the group X itself:

$$U(1) \longrightarrow Heis(X, \omega) \longrightarrow X\,.$$

Surprisingly, and it is a astonishing fact that what represent a hallmark of quantum mechanics has appeared simply by applying Lie integration to the Lie algebraic structures in classical mechanics: if we think of Lie integrating $\mathbb{R}$ the interesting circle group $U(1)$ instead of the translation group $\mathbb{R}$, then the name of its canonical basis element $1 \in \mathbb{R}$ is canonically i, the imaginary unit. Therefore one often writes the above central extension instead as follows:

$$i\mathbb{R} \longrightarrow \mathfrak{poiss}(X,\omega) \longrightarrow \mathfrak{ham}(X,\omega)$$

Consider the simple special case where $(X,\omega) = (\mathbb{R}^2, dp \wedge dq)$ is the 2-dimensional symplectic vector space which is for instance the phase space of the particle propagating on the line. Then a canonical set of generators for the corresponding Poisson bracket Lie algebra consists of the linear functions p and q of classical mechanics, together with the constant function. Under the above Lie theoretic identification, this constant function is the canonical basis element of $i\mathbb{R}$, hence purely Lie theoretically it is to be called i.

With this notation then the Poisson bracket, written in the form that makes its Lie integration manifest, indeed reads

$$[q,p] = i\,.$$

Since the choice of basis element of $i\mathbb{R}$ is arbitrary, we may rescale here the by any non-vanishing real number without changing this statement. If we write $\hbar$ for this element, then the Poisson bracket instead reads

$$[q,p] = i\hbar\,.$$

This is of course the hallmark equation of quantum physics, if we interpret $\hbar$ here indeed as Planck's constant. We see it arises here merely by considering the non-trivial (the interesting non-simply connected) Lie integration of the Poisson bracket. This outcome of the quantization, naturally understood and *derived* from applying Lie theory to classical mechanics, will be the basis of what we will call later the geometric quantization program (Kirillov 1976; Kostant 1976; Souriau 1970).

Let us describe the construction of the quantomorphism group which is the non-trivial Lie integration of the Poisson bracket. Given the symplectic form ω, it is natural to ask if it is the curvature 2-form of a $U(1)$-principal connection ∇ on complex line bundle L over X (this is directly analogous to Dirac charge quantization when instead of a symplectic form on phase space we consider the the field strength 2-form of electromagnetism on spacetime). If so, such a connection (L,∇) is called a prequantum line bundle of the phase space (X,ω). The quantomorphism group is simply the automorphism group of the prequantum line bundle, covering diffeomorphisms of the phase space (the Hamiltonian symplectomorphisms mentioned above).

As such, the quantomorphism group naturally acts on the space of sections of L. Such a section is like a wave function, except that it depends on the entire phase space, instead of just on the *canonical coordinates*. For purely abstract mathematical reasons linked with the nature of the associated Hilbert space, it is indeed natural to choose a *polarization* of phase space into canonical coordinates and canonical momenta and consider only those sections of the prequantum line bundle which depend only on the former. These are the actual wave functions of quantum mechanics, hence the quantum states. And the subgroup of the quantomorphism group which preserves these polarized sections is the group of exponentiated quantum observables. For instance in the simple case mentioned before where (X,ω) is the 2-dimensional symplectic vector space, this is the Heisenberg group with its famous action by multiplication and differentiation operators on the space of complex-valued functions on the real line.

One final remark concern the cohomolgical nature of the Mass: *it parametrizes the extensions of the Galileo group*. Indeed, in classical mechanics, the Galilei group acts on the symplectic manifold of states of a free particle. But in quantum mechanics, we only have a projective representation of this group on the Hilbert space of states of the free particle. The cocycle is the particles mass. In other words, you can not see the mass of a free classical particle by just watching its trajectory, since it goes along a straight line at constant velocity no matter what its mass is. But you can see the mass of a free quantum particle, because its wave function smears out faster if it is lighter! So there is some difference between classical and quantum mechanics. Ultimately this arises from the fact that the latter involves an extra constant, Planck's constant. In slight disguise, one can see this cocycle also control already the classical free non-relativistic particle, in the sense that its action functional is of the form of a 1d WZW model with that cocycle being the "WZW term" that however comes down to be the ordinary free action.

Conclusion

The presence of a cohomological nature in physics in general and in quantum field theory in particular is confirmed by the modern treatment of quantum symmetries, gauge invariances, renormalization, anomalies, the BRST formalism and the numbers associated with the figures (diagrams) via the Feynman integrals. For the elliptic type, Witten sees it as a generalization of the characteristic classes like that of Euler. He deduces the premises of a (rather infinite) geometric definition of the elliptic cohomology which enters a hierarchy:

bordisme et cobordisme ⟶ cohomologie elliptique ⟶ K-theorie ⟶ cohomologie ordinaire

Witten interprets the Jones polynomial with the cohomology (with coefficient in a sheaves) of a space of non-abelian cohomology of Riemann surfaces with boundaries (with coefficient in unitary groups). In general, the topological quantum fields theories in dimension 3 are cohomologies of a new type,[9] so that the spaces of states Φ_M of a quantum field theory seem to be close to cohomology as well. Thus: the spaces Φ_M would be cohomology groups; the bundle $\mathcal{A}$ of the dynamic states would be rather something like an *algebra of operations*.

The study of non-abelian cohomolgy (Giraud 1971) sets is a very active subject. The spaces of Holomorphic bundles on complex algebraic curves are at the heart of fields theories, as well as the curves moduli spaces. Similarly, the moduli spaces of holomorphic bundles on complex algebraic surfaces parametrize the equivalence classes of self-dual connections in real dimension 4, the minimal solutions of the Yang-Mills equations which generalize the Maxwell (Euclidean) equations for the case of strong and weak interactions.

When the rank of the bundles is 1, they are spaces of (almost) ordinary cohomology, but when the rank is >1, they are spaces of non-abelian cohomology. Because of their singular nature, they are unstable. For this we consider the most ordinary cohomology of this cohomology to stabilize the forms (topological gravity in dimension 2, Donaldson polynomial invariants in dimension 4, …).

Fiber Bundles in Physics: In Perspective

As we know the gauge fields in Yang-Mills theory (Atiyah 1979), hence in electromagnetism, in QED and in QCD, and in the standard model of the known universe, are not really just the local differential 1-forms A^a_μ known from so many textbooks, but are globally really connections on principal bundles (or their associated bundles) and this is all-important once one passes to non-perturbative Yang-Mills theory, hence to the full story, instead of its infinitesimal or local approximation.
Notably what is called a Yang-Mills instanton in general and the QCD instanton in particular is nothing but the underlying nontrivial class of the principal bundle underlying the Yang-Mills gauge field. Specifically, what physicists call the instanton number for SU(2)-gauge field theory in 4-dimensions is precisely what mathematically is called the second Chern-class, a characteristic class of these gauge bundles.

YM Instanton = class of principal bundle underlying the non-perturbative gauge field

To appreciate the utmost relevance of this, observe that the non-perturbative vacuum of the observable world is a sea of instantons with about one *YM* instanton per femtometer to the 4th. (see for instance the first sections of Schaefer and Shuryak (1998) for a review of this fact). So the very substance of the physical world, the very vacuum that we inhabit, is all controled by non-trivial fiber bundles and is inexplicable without these.

[9] The cycles carried by a surface Σ are formal combinations of manifolds of dimension 3 bordered by Σ; The partition function Z defines a form of intersection on cycles and homology occurs when we quotient by the kernel of this form.

Also monopole solutions in physics are mathematically nontrivial principal bundles. For instance the Dirac monopole (that appears in Dirac charge quantization) or the Yang monopole.

Similarly fiber bundles control all other topologically non-trivial aspects of physics. For instance most quantum anomalies are the statement that what looks like an action function to feed into the path integral, is globally really the section of a non-trivial bundle notably a Pfaffian line bundle resulting from the fermionic path integrals. Moreover all classical anomalies are statements of nontrivializability of certain fiber bundles.

Actually the role of fiber bundles reaches a good bit deeper still. Quantization is just a certain extension step in the general story, but already classical field theory cannot be understood globally without a notion of bundle. Notably the very formalization of what a classical field really is says: a section of a field bundle. For instance the global nature of spinors, hence spin structures and their subtle effect on fermion physics are all encoded by the corresponding spinor bundles.

Two aspects of bundles in physics come together in the theory of gauge fields and combine to produce higher fiber bundles: namely we saw above that a gauge field is itself already a bundle (with a connection), and hence the bundle of which a gauge field is a section has to be a second-order bundle. This is called gerbe or 2-bundle: the only way to realize the Yang-Mills field both locally and globally accurately is to consider it as a section of a bundle whose typical fiber is $\mathbf{B}G$, the universal moduli stack of GG-principal bundles.

All of this becomes even more pronounced as one digs deeper into local quantum field theory, with locality formalized as in the cobordism theorem that classifies local topological field theories. Then already the local Lagrangians and local action functionals themselves are higher connections on higher bundles over the higher moduli stack of fields. For instance the fully local formulation of Chern-Simons theory exhibits the Chern-Simons action functional with all its global gauge invariance correctly realized as a universal Chern-Simons circle 3-bundle. This is such that by transgression to lower codimension it reproduces all the global gauge structure of this field theory, such as in codimension 2 the WZW gerbe (itself a fiber 2-bundle: the background B-field of the WZW model!), in codimension 1 the prequantum line bundle on the moduli space of connections whose sections in turn yield the Hitchin bundle of conformal blocks on the moduli space of Riemann surfaces. In short all global structure in field theory is controled by fiber bundles, and all the more the field theory is quantum and gauge. The only reason why this can be ignored to some extent is because field theory is a complex subject and maybe the majority of discussions about it concerns really only a small little perturbative local aspect of it. But this is not the reality. The QCD vacuum that we inhabit is filled with a sea of non-trivial bundles and the whole quantum structure of the laws of nature are bundle-theoretic at its very heart.

Acknowledgements For the invitation to contribute to this volume and for discussion I thank Shyam Wuppuluri.

References

M.F. Atiyah, *Geometry of Yang-Mills fields* (Lezioni Fermiane Academia Nationale dei Lincei & Scuola Normale Superiore, Pisa, 1979)

D. Bennequin, *Théories quantiques et fibrations*, dans Vers une nouvelle philosophie de la nature, ed. by J. Kouneiher (Hermann, 2010)

H. Cartan, S. Eilenberg, *Homological Algebra* (Princeton University Press, 1956)

P.A.M. Dirac, The relation between mathematics and physics, in *Proceedings of the Royal Society (Edinburgh)* vol. 59, 1938–39, Part II pp. 122–129

J. Anandan, Causality, symmetries and quantum mechanics. Found. Phy. Lett. **15**, 5 (2002)

Ch. Ehresmann, *Œuvres complètes et commentées*. I-1,2. *Topologie algébrique et géométrie différentielle* ed. by Andrée Charles Ehresmann. Cahiers Topologie 14, Géom. Différentielle 24(1983), suppl. 1, xxix+601 pp

Ch. Ehresmann, Sur la topologie de certains espaces homogénes. Ann. Math. **35**, 396–443 (1934)

Ch. Ehresmann, J. Feldbau, *Sur les propriétés d'homotopie des espaces fibrés*. C.R. Acad. Sci. Paris **212**, 945–948 (1941)

S. Eilenberg, N. Steenrod, *Foundations Of Algebraic Topology* (Princeton Legacy Libirairy, 2011)

L. Euler, Solutio problematis ad geometriam situs pertinentis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, **8**, 128–140 (1736)

J. Feldbau, Sur la classification des espaces fibrs. C. R. Acad. Sci. Paris. **208**, 1621–1623 (1939)

J. Giraud, *Cohomologie non ablienne*, Grundlehren 179 (Springer, 1971)

A. Herreman, Le, statut de la géométrie dans quelques textes sur lhomologie, de Poincaré aux années 1930. Revue dHistoire des Mathmatiques. **3**, 241–293 (1997)

G. Kirchhoff, Ueber die Aufl osung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanisher Ströme gefürt wird. Ann. der Physik und Chemie **72**, 497–508 (1847)

A.A. Kirillov, *Elements of the Theory of Representations* (Springer, Berlin, Heidelberg, New York, 1976)

B. Kostant, *Quantization and Unitary Representations*. Lecture Notes in Mathematics, vol. 170 (Springer, Berlin, 1976)

J. Kouneiher, *Symétrie et fondements de la Physique, l'aspect Cohomologique de la Physique*, in *Vers une Nouvelle Philosophie de la Nature*, ed. by J. Kouneiher (Hermann edition, 2010)

J. Kouneiher, *Asymmetric, Pointless and Relational Space: Leibnizs Legacy Today: Leibniz and the Dialogue between Sciences, Philosophy and Engineering, 1646-2016. New Historical and Epistemological Insights*, ed. by R. Pisano, M. Fichant, P. Bussotti, A.R.E. Oliveira (The Colleges Publications, London, 2017)

J. Leray, L'homologie d'un espace fibré, dont la fibre est connexe. J. Math. Pures et Appl. **29**, 169–213 (1950)

J.B. Listing, *Vorstudien zur Topologie*, Gttinger Studien, 811–875 (1847)

J.B. Listing, *Der Census rumlicher Complexe oder Veallgemeinerung des Euler'schen Satzes von den Polyedern*, Gttingen, 97–180 (1861)

W. Massey, *A history of cohomology theory*, in *History of Topology*, ed. by I.M. James (Elsevier, Amsterdam, 1999)

J.C. Maxwell, *It A Treatise on Electricity and Magnetism*, vol. I and vol II (1873), 3 edn. (Oxford University Press, 1904)

H. Poincaré, Analysis situs. J. Sec. Polyt. **1**, 1–121 (1895)

H. Poincaré, Sur la connexion des surfaces algébriques. C. R. Acad. Sc. **133**, 969–973 (1901)

H. Poincaré, Sur les cycles des surfaces algébriques; quatrième complement a l'Analysis situs. J. Math. pures et appl. **8**, 169–214 (1902)

J.-C. Pont, *La topologie algèbrique des origines à Poincaré* (Presses Universitaires de France, 1974)

T. Schaefer, E. Shuryak, Instantons in QCD. Rev. Mod. Phys. **70**, 323–426 (1998). arXiv:hep-ph/9610451

H. Seifert, Topologie dreidimensionaler geschlossener Rume. Acta Math. **60**, 147–238 (1932)

J.-P. Serre, Homologie singulire des espaces fibrs. Applications. Ann. of Math. **54**, 425–505 (1951)

U. Schreiber, *What quantization*, see (https://physics.stackexchange.com/users/5603/urs-schreiber)

J.M. Souriau, *Structures des systemes dynamiques* (Dunod, Paris, 1970)

N. Steenrod, *The Topology of Fiber Bundles* (Princeton, 1951)

G. 't Hooft, Renormalization of massless Yang–Mills fields. Nucl. Phys. B **33**, 173–199 (1971)

G. 't Hooft, *Under the spell of the gauge principle*, in Advanced Series in Mathematical Physics, vol. 19 (World Scientific, 1994)

G. 't Hooft, M. Veltman, Regularization and renormalization of gauge fields. Nucl. Phys. B **44**, 189–213 (1972)

H. Whitney, Sphere spaces. Proc. Natl. Acad. Sci. USA **21**, 464–468 (1935). https://doi.org/10.1073/pnas.21.7.464

H. Whitney, On the theory of sphere bundles. Proc. Natl. Acad. Sci. USA **26**, 1481–53 (1940). https://doi.org/10.1073/pnas.26.2.148

Chapter 14
Quantum Physics and Time from Inconsistent Marginals

Chiara Marletto and Vlatko Vedral

PACS numbers: 03.67.Mn · 03.65.Ud

When formulating a scientific explanation—be it a fundamental physical theory such as quantum theory or general relativity, or a higher-level scientific explanation such as the theory of evolution—we often take space and time for granted. They appear in the explanation, as elements of the frame of reference with respect to which statements are made; but they are, usually, not explained further. In this essay we speculate about the idea that resorting to the coordinate time is a necessity to explain certain outcomes of measurements that can be observed in physical systems. To introduce this idea, we can first consider how quantum theory and its properties arise as a necessity, to resolve certain inconsistencies in the classical theory of probability, in the so-called problem of marginals.

Marginals are a special case, relevant for probabilistic theories, of a more general concept that is relevant to any scientific explanation. A scientific explanation often supplements one's immediate perception of physical reality. It provides a *global* picture that integrates the snippets one can access via perception into a global picture that is consistent overall. For example, one can perceive directly the dark sky at night; the explanation for that is a cosmological model describing an expanding universe, together with the theory of light propagation. Likewise, one can witness the moon apparently moving across the night sky; this is explained in terms of the planetary motion around the sun, and of the motion of the moon around Earth. A particularly important example are scientific explanations which involve space and time. For example, given the existence fossils, which we can see directly at this time

C. Marletto
Clarendon Laboratory, University of Oxford, Parks Road, OX1 3PU Oxford, UK
e-mail: chiara.marletto@physics.ox.ac.uk

V. Vedral (✉)
Centre for Quantum Technologies, NUS, Singapore, Singapore
e-mail: vlatko.vedral@gmail.com

V. Vedral
Department of Physics, NUS, Singapore, Singapore

C. Marletto · V. Vedral
ISI Foundation, Turin, Italy

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_14

in the evolution of the universe, one can put forward a theory about dinosaurs, in the context of evolutionary biology. That theory is a way of sowing together the snippets provided by fossils, by creating a consistent story in time—a sequence of events. Likewise, in the process of reconstructing a jig-saw puzzle, one can typically only see some partial details of a picture—say that of a cat; those details are sub-pictures separated in space, which are then connected by adding pieces on the ground of one's understanding of the complete picture. So, for example, one makes sure that the two eyes of the cat are spatially located at the same level with respect to each other; that they are above the whiskers; and so on. And this is because one has an overall theory of how a typical cat's head is arranged.

Explanations can take the form of a *probabilistic* theory, where events are assigned a certain probability to occur. For instance, one is provided with the probability of an event occurring (e.g. a cat mewing) and of another event occurring (a dog barking) and then one has to provide a joint probability distribution (e.g. for a dog barking and a cat mewing) that is consistent with the given two 'marginal' probability distributions for the events happening by themselves. In classical probability theory, this is known as 'the marginal problem'. In more formal terms, this problem is concerned with the following question: given probability distributions of individual random variables, find a possible joint probability distribution which is consistent with the given marginals.

For instance, for a binary random variable (one that can assume two possible values), given two marginal probability distributions $p(A) = (1/2, 1/2)$ and $p(B) = (1/2, 1/2)$, we can construct a joint probability distribution like $p(A, B) = (1/2, 0, 0, 1/2)$ which produces the desired marginals $p(A)$ and $p(B)$. There are usually many joint probability distributions consistent with one set of marginals (for instance the distribution $p(A, B) = (1/4, 1/4, 1/4, 1/4)$ could also have produced the desired marginals). Usually, in the context of a scientific theory, there would be a specific explanation that comes with either of those models, giving details of the particular phenomenon that they describe; but the problem of marginals is concerned with the overall question of finding all the global probability distributions consistent with given marginals. This problem is constrained by the conditions that probabilities must be positive real numbers smaller than one; that the sum of the probabilities for all the events must be one; and that the probability for unrelated events to happen one after the other is the sum of the probabilities of the two events.

These are strong constraints. Indeed, there the marginal probability distributions that cannot arise from a joint probability distribution. Imagine for instance three random variables, with the following marginals: $p(A, B) = (1/4, 0, 0, 0)$ while $p(A, C) = (1, 0, 0, 0)$. This cannot be satisfied at the level of $p(A, B, C)$ simply because the marginal of A is maximally mixed (i.e. one has maximum uncertainty) as far as the distribution $p(A, B)$ is concerned, but it is pure, i.e. it is completely determined, as far as far as $p(A, C)$.

Until the advent of quantum theory, all such cases were just considered as inconsistencies in the context of classical probability theory. However, quantum theory has shown that some of those cases can be considered as describing events that correspond to the outcomes of measurements in a quantum system. This is a rather interesting scenario: although marginal distributions cannot arise out of a global probability distributions, they could still be considered as arising from a global quantum state.

From now on, the events we are interested in are a particular outcome for an observable of a system to be observed. So, outcome and event will be used as synonyms. For example, you can think of a binary observable (one that has only two possible outcomes, like the direction of an arrow—which can be pointing up or down); the events we will be referring to will be the event where the direction 'up' is observed, and the event where the direction 'down' is observed.

A quantum state (formally represented as a density operator—a positive hermitian operator, with trace equal to one) can be considered as containing the specification of the probabilities for all outcomes of all possible measurements on the physical system it describes. Now here comes the interesting bit. Classical states represented by classical probabilities can be thought of as vectors. Every element of the vector contains a probability for one particular outcome of a measurement (so, all elements add up to unity). Quantum states are, instead, matrices. This is because to describe completely a quantum system it is not sufficient to just write down the probabilities for the outcomes of the measurements of one observable. What we also need are the amplitudes connecting every two events, each corresponding to an outcome being observed. So if we have two events, in quantum theory we have to present not just p_1 and p_2 (which are the probabilities for the respective events) but we also need another number a_{12} which is the amplitude connecting the two events (just one, because in quantum theory the matrix element a_{21} is the complex conjugate of a_{12}, by symmetry). The remarkable point here is that there are marginal probability distributions that can only be explained assuming that the global picture is given by quantum theory, in the form of a quantum state as well as a collection of quantum measurements of observables. The marginal probability distributions are then interpreted as marginal states emerging from that quantum state. This happens when a measurement (represented by another matrix in quantum physics) is applied to a system in a given quantum state. The product of the matrix representing the state and the one representing the measurement then leads to another matrix whose diagonal elements represent, informally speaking, the classical probability distribution arising out of the measurement. This property of being able to account for marginals by having an overall quantum state is, in jargon, called 'quantum contextuality' (contextuality, because the outcome is context, i.e. measurement, dependent).

Here is a beautiful example, (Klyachko et al. 2007). We have five binary random variables $A_1, \dots A_5$ and every two adjacent ones can either be correlated (both have values 0 or both have values 1) or anti-correlated (one has value 0 and the other one has value 1). We are assuming that the distributions A_1 and A_5 are also adjacent. Because of the constraints, the total number of anti-correlated random variables could therefore be 0, 2, or 4 (the variables one and two, two and three, three and four and four and one are all anti-correlated). Given many different possible random variables, the average number of the anti-correlated ones cannot therefore exceed 4.

Then, imagine that someone allows you to sample from these different distributions, but you can only look at the two nearest neighbours ones at a time. So, you randomly pick the neighbouring random variables and look at the outcome. (This random choice is necessary because if your choice of the neighbouring random variables is known in advance, the distributions could have been arranged in advance to “conspire” to produce any desired average.) You now find that no matter which edge you choose, you find zero-zero with a probability of $1 - \frac{2}{\sqrt{5}}$, zero-one with $\frac{1}{\sqrt{5}}$, and one-zero with $\frac{1}{\sqrt{5}}$. So, the expectation value of the sum of mismatches is $2\sqrt{5} = 4.47 > 4$.

There is clearly no classical joint probability distribution over five random variables that could possibly produce these marginals. But there is a quantum state that can be seen as producing those marginal distributions. Each collection of five random variable is interpreted as a 3D quantum system with an orthonormal basis of quantum states $\{|A\rangle, |B\rangle, |C\rangle\}$. Each is initialised to the state $|C\rangle$. Each neighbouring pair of distributions is assigned an observable, represented by a 1D projector, to $\frac{1}{\sqrt{\sqrt{5}}}|C\rangle + \sqrt{1 - \frac{1}{\sqrt{5}}\left[\cos\left(\frac{4\pi n}{5}\right)|A\rangle + \sin\left(\frac{4\pi n}{5}\right)|B\rangle\right]}, n = 1, \ldots, 5$, which represents the outcome of a measurement. Adjacent projectors commute—meaning that they represent properties that can be measured simultaneously with the maximal precision. If we project, think of the outcome as anti-correlated. Otherwise, it is correlated. The probability to project is formally given as the trace of the projector acting on the quantum state; and quantum theory allows a particular quantum state to reproduce those marginals, which could not possibly arise as marginals of a classical joint probability distribution.

Bell’s inequalities (Bell 2013) are also just another example of this fact that while there can be no global probability distribution given some marginals, there can still exist a global quantum state, or density operator. However, unlike in the previous example, Bell’s inequalities involve additional assumptions about locality, specifically the requirement of “no-signalling” between the two subsystems that are measured. In the Bell case we have four random variables and are given the marginals of every two shared by the two subsystems. These details, however, are not relevant from our perspective. The main point here is that “quantumness” is a necessary element to explain some observed statistics that cannot be accounted for using probability distributions. One must resort to a new physical theory in order to explain marginals that cannot possibly arise out of a global classical probability distribution.

One could then argue that the existence (experimentally speaking) of correlations that cannot be explained as arising from any allowed joint probability distribution is what forces us to resort to quantum theory. This means that we must give up the fact that states of physical systems can be described using just probabilities; and the idea that observables can all be measured at the same time to the same arbitrarily high accuracy. To account for joint measurements, we need to use quantum states—density operators. To describe observables, we need to resort to operators that in general do not commute with one another.

An intriguing possibility now arises. In analogy to the above-described classical-to-quantum transition, suppose now that one has some marginal descriptions given by quantum states; and that there is no global quantum state that can reproduce the marginals. Is there some explanation that one could come up with, which allows still to explain those marginals as arising from a different, global description? This is the main topic of the rest of this essay. We would like to argue, as we said, that the global description must resort to the coordinate time.

First of all, we would like to be clear that, as in the classical case, we can always choose marginals that are globally incompatible. Let us say that we have three subsystems A, B, C and we require that A and B are in a pure quantum state, but also that so are A and C. This is clearly impossible, even quantum-mechanically. However, what is interesting is that we can find incompatible marginals, but they could still arise from physical measurements that are not described by an overall quantum state! In other words, there is a consistent description that would give those as marginals, which is not a quantum state, but something else. And as we shall see, this something else is a sort of generalised quantum state that involves an additional dimension to space, i.e., time.

We now proceed to give an example of this. Suppose that we want to describe a physical process where a single qubit, initially in a maximally mixed state, is then measured at two different times. Each measurement is performed in all three complementary basis X, Y, Z (represented by the usual Pauli operators). The evolution is trivial between the two measurements, i.e. the identity operator. Suppose now that we would like to write the statistics of the measurement outcomes in the form of an operator, generalising the quantum density operator. Because the whole state, as we shall see, is hermitian and unit trace, but not positive, we refer to it as a 'pseudo-density matrix' (Fitzsimons et al. 2015) (see also Leggett et al. 1985; Brukner et al. 2004; Horsman et al. 2016 for different views of temporal quantum correlations).

The state would be represented as as matrix in the following way:

$$R_{12} = \frac{1}{4}\{I + X_1X_2 + Y_1Y_2 + Z_1Z_2\} \ . \tag{14.1}$$

This operator looks very much like the density operator describing a singlet state of two qubits, however, the correlations all have a positive sign (whereas for the singlet they are all negative, $\langle XX\rangle = \langle YY\rangle = \langle ZZ\rangle = -1$). In fact, it is simple to show that R_{12} is not a density matrix, because it is not positive (i.e. it has at least one negative eigenvalue). We can however, trace the label 2 out and obtain one marginal, i.e. the "reduced" state of 1. Interestingly, this itself is a valid density matrix (corresponding to the maximally mixed state $I/2$). Likewise for the subsystem 2. So the marginals of this generalised operator are actually both perfectly allowed physical states, but the overall state is not.

It is important to stress that this is not just an artefact of making two measurements in time on a single system. The same goes for measurements performed at three or more times (Genovese in preparation).

The simple reason why temporal correlations cannot always be written as a density matrix is that the outcomes to measurements performed consecutively in the same basis are always perfectly correlated. That means that we would have the correlation signature of the kind: $\langle XX\rangle = \langle YY\rangle = \langle ZZ\rangle = 1$. However, as we said, there is no allowed density matrix with this signature of correlations: this violates one of the principles of quantum mechanics, because it would require the observables XX, YY and ZZ all to be simultaneously correlated. Therefore in a pseudo-density matrices, although different instances in time can be treated as different qubits (subsystems, more generally), the price to pay is that the resulting overall state can have negative eigenvalues (which, therefore, could not be interpreted as probabilities, at least if we think of probabilities either as representing frequencies or degrees of belief).

Now, it should be stressed that the above marginals alone could be obtained from a overall valid density matrix. For instance, we could have

$$\rho_{12} = \frac{1}{2}\{|00\rangle\langle 00| + |11\rangle\langle 11|\} \ . \tag{14.2}$$

However, this state does not reproduce the two point correlations that exist in R_{12}. As we said, there is no density matrix that can reproduce those correlations (since, as we explained, density matrices do not allow for perfect correlations in all complementary basis measured; only perfect anti-correlations are allowed).

There are many open questions and research avenues to explore regarding pseudo-density matrices. First, what is the set of all allowed pseudo-density matrices? We know the answer for a qubit measured at two different times (Zhao et al. 2017), but we do not have results for higher dimensions or more then two measurements in time. Secondly, can the dynamics of any quantum systems be understood within this picture? Thirdly, we would like to use pseudo-density matrices so as to provide a unified approach to the study of temporal and spatial quantum correlations. However, rather than tackling these questions (which are left to future research), here we would like to focus on a more fundamental issue.

Our discussion has shown two facts, that can now be unified in a general explanation. First, the existence of some marginal distributions that cannot arise from a joint probability forces us to introduce quantum density operators—and the idea that certain observables cannot all be measured simultaneously to arbitrarily high accuracy. Second, the existence of marginal density matrices that cannot arise out of an overall joint density matrix forces us to introduce another dimension to space—time—in which the description of correlations between measurements of different observables requires the use of pseudo-density matrices. The unifying explanation is this. We are in an intriguing position to be able to derive both quantum physics and the existence of time from the fact that marginals cannot be understood from joint states. However, there are some subtleties to be taken into account.

First and foremost, of course, the pseudo-density matrix description does not tell us anything about the ordering of events in time. It is fully time-symmetric, so that we cannot tell which measurement is performed first and which last (though we can tell which ones are in the middle). Secondly, we have in fact been working entirely within the Newtonian "totally ordered" concept of time (where for every two events either one is before the other or they are simultaneous). Relativity, on the other hand, tells us that events in time are only partially ordered. In fact, according to relativity (and in reality), given two events it is possible that neither is in the past of the other (or simultaneous) and this forces us to accept that space must exist and that it must have at least one dimension (because if time only existed and was one dimensional, then all instances would have to be totally ordered). Thus, if we already acknowledge the existence of time as fundamental, arising from the reasoning we presented here, enforcing the relativistic partial ordering requires one to derive the existence of (at least one dimension of) space as a necessity Alfred et al. (1914).

This leads us to considering the following possibility. If we take relativity into account and we are given the pseudo-density matrix of the universe (because no probability distribution and no density matrix would suffice), could we then separate the spatial from the temporal aspects of this pseudo-density matrix? Could space and time arise from the universal quantum correlations? This of course would leave the problem of how to define a measurement (to which the notion of an observable implicitly refers) without referring to a notion of before-and-after. Still, we can explore this line of argument as it is an intriguing conjecture.

The logic would go as follows: we make observations and construct two random variable marginals that cannot be derived from an overall probability distribution. This necessitates the introduction of quantum physics; we then observe that the measurements provide cases where we cannot describe outcomes consistently even if we resort to quantum theory—meaning that the overall state from which the local state arise is not quantum (it is a pseudo-density matrix). This forces us to interpret the dimension in which measurements are made as time. Then we realise that events are only partially ordered in time, which forces us to introduce one dimension of space. And so on. This would clearly have to be extended to then account for 3 dimensions of space. However, we could envisage some kind of relationship, first suggested by von Weizsaecker Carl et al. (1971), that the space has to have three dimensions because a quantum bit requires three real numbers to be fully specified. Both space and time would therefore arise out of the statistics of measurements. Are observables, measurements and the correlations of their outputs a deeper explanation for how the territory around us is charted in terms of space and time?

Acknowledgements CM's research was supported by the Templeton World Charity Foundation and by the Eutopia Foundation. VV thanks the Oxford Martin School, the John Templeton Foundation, the EPSRC (UK). This research is also supported by the National Research Foundation, Prime Ministers Office, Singapore, under its Competitive Research Programme (CRP Award No. NRF-CRP14-2014-02) and administered by Centre for Quantum Technologies, National University of Singapore.

References

A paper in preparation, CM and VV in collaboration with the group of M. Genovese

J.S. Bell, Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy (Cambridge Univeristy Press, 2013)

C. Brukner, S. Taylor, S. Cheung, and V. Vedral, arXiv preprint quant-ph/0402127 (2004)

J. Fitzsimons, J. A. Jones and V. Vedral, Scientific Reports 5, Article number: 18281 (2015)

D. Horsman, C. Heunen, M. F. Pusey, J. Barrett, R.W. Spekkens, arXiv preprint arXiv:1607.03637 (2016)

A. Klyachko, M. Ali Can, S. Binicioglu, and A. Shumovsky, Foundations of Physics, Vol. 37, No. 8, (2007)

A.J. Leggett, A. Garg, Physical Review Letters **54**, 857 (1985)

A. Robb, A theory of time and space (Cambridge University Press, 1914)

C.F. von Weizsaecker, Einheit der Natur (The Unity of Nature) (in German) (1971)

Z. Zhao, R. Pisarczyk, J. Thompson, M. Gu, V. Vedral, J.F. Fitzsimons, arXiv preprint arXiv:1711.05955 (2017)

Chapter 15
Quantum Non-individuality: Background Concepts and Possibilities

Décio Krause and Jonas R. Becker Arenhart

Introduction

It is not an exaggeration to say that quantum mechanics is at odds with most of our received metaphysical notions. In particular, an alleged revision is brought about by the theory on the metaphysical notion of 'individuality'. Certainly, this should figure as being of great interest for metaphysicians and philosophers of science alike. What makes issues even more interesting is that some of the founding fathers of the theory, with their typical philosophical inclinations, suggested that the entities dealt with by the theory had something different regarding individuality: according to them, quantum entities somehow fail individuality. That situation is clearly distinctive from what happened in classical mechanics, for instance (see French and Krause 2006, chap. 3 for a historical overview).

Having such a request for revision on individuality, however, is not the same as having a new approach to individuality right at hand; the founding fathers expressed the failure of individuality in rather vague terms, claiming that quantum entities had 'lost their identity'. In the context of their discussions, it is clear that their target is the very notion of individuality; however, knowing that something is wrong does not always give us any positive sign on how to fix it. Furthermore, the claim that quantum entities 'lost their identities' is at best a heuristic, that may be articulated in a plurality of distinct ways.

Consider Weyl (1950, p. 241) on the possibility of discerning two electrons:

> …the possibility that one of the identical twins Mike and Ike is in the quantum state E_1 and the other in the quantum state E_2 does not include two differentiable cases which are permuted on permuting Mike and Ike; it is impossible for either of these individuals to retain

D. Krause (✉) · J. R. Becker Arenhart (✉)
Department of Philosophy, Federal University of the Santa Catarina, Santa Catarina, Brazil
e-mail: deciokrause@gmail.com

J. R. Becker Arenhart
e-mail: jonas.becker2@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_15

> his identity so that one of them will always be able to say 'I'm Mike' and the other 'I'm Ike'. Even in principle one cannot demand an alibi of an electron!

Thus, electrons are not like people or ordinary objects, of which we could demand an alibi. By 'alibi' we may understand something that would allow us to individuate or, perhaps, to distinguish a particle from other similar items. The idea seems to be that electrons are all just so much alike that nothing discerns them. Schrödinger also remarked that "we cannot mark an electron; we cannot paint it red" (Schrödinger 1964). While a painting over an ant or a twin may well serve as an alibi for its individuality, nothing of the sort works for an electron or another quantum particle. But still, this is not enough for us to determine what is wrong with electrons on what concerns individuality.

Consider Schrödinger again, in another context:

> I beg to emphasize this and I beg you to believe it: it is not a question of our being able to ascertain the identity in some instances and not being able to do so in others. It is beyond doubt that the question of 'sameness', of identity, really and truly has no meaning. (Schrödinger (1996, pp.121-122)).

Schrödinger goes even farther than Weyl, it seems, by claiming that the problem is not the failure of discernibility, but rather that the very idea of identity fails to make sense in some cases. That is, there are situations in which one cannot even say that some objects are the same or different. In the broader context of this sentence, Schrödinger is addressing the issue of identity over time, of whether we may say some entity at a time t_1 is the same as another entity seen at a later time t_2. If we take this quote seriously, then, the claim that quantum particles 'lost their identity' is now to be understood literally. But let us not go so fast.

These quotes are just samples for us to motivate the claim that the idea of a "loss of identity" was indeed widespread among the founding fathers of the theory. This view, that quantum entities somehow lost their identity, was called the *Received View on quantum non-individuality* by French and Krause (2006) (for simplicity, we shall refer to it simply as the RV). What was received was the idea that quantum particles had somehow lost their individuality, that identity does not make sense, or that we cannot always discern those entities. However, as we mentioned, if that slogan is to make sense, we must provide the view with a more detailed and metaphysically articulated development. As a general view, the RV recommends only that quantum particles are different from classical particles on what concern issues of identity and individuality, but does not by itself impose any specific view of identity and individuality that is to be revised. Notice that while Weyl speaks of the lack of an alibi, inducing one to think of a failure of discernibility, Schrödinger speaks of identity making no sense, which could be seen as demanding more profound revisions. The development of the RV, then, may be provided for in a variety of distinct ways, by the clear understanding of the notions of identity and individuality, and their relations. It is to these possibilities that this chapter will be devoted. We shall explore and illustrate how the idea that entities lost their identities may be further clarified and become a workable view of quantum ontology, as described by a possible understanding of quantum mechanics. There is more than one way to do that, but here we shall not enter into the dispute of which of them is preferable, if any.[1]

In order to discuss some of the possibilities open for the metaphysical articulations of the RV, we shall proceed as follows. In the next section we briefly sketch the main reasons related with the claim that quantum mechanics seems to afford a theory of non-individuals: quantum statistics and the permutation symmetry. This will clarify the physics behind the metaphysical developments of the RV. In section "Identity, Individuality, Individuation" we sketch the main concepts to be employed in the metaphysical discussions of the chapter: individuality, individuation, and identity. In section "Schrödinger's Problem" we present one of the most well-known ways to articulate the RV, which is based on a literal understanding of Schrödinger's claim that identity makes no sense for quantum entities. It is the view put forward, for instance, in French and Krause (2006), which can be backed by formal systems of non-reflexive logics. In section "Non-individuals with Identity" we present alternatives to the non-reflexive approach, which may also be candidates to ground a metaphysics that is faithful to the tenets of the RV (although not to Schrödinger's claim that identity makes no sense). We conclude in section "Conclusion".

Quantum Mechanics, Statistics, Permutation, Identity

As we have already mentioned, the RV, as crudely advanced by some of the founding fathers of quantum mechanics, is a very general view that relates identity, individuality, and indiscernibility. Indiscernibility seems to be thought of as one of the main ingredients of the problem; but identity and individuality are also related. In order to untie the knot involving the three concepts involved, we shall begin by presenting the main quantum mechanical facts that led to such considerations.[2] We begin by presenting the classical statistics, whose contrast with quantum case originates the main claims of the RV.

The idea that classical particles are individuals in a very strong sense is famously encapsulated in Maxwell-Boltzmann's statistics. Let us illustrate it with the case in which two particles, labeled 1 and 2, must be distributed in two states, A and B. We have the following four distinct possibilities for such a distribution (where $A(1)$ is an abbreviation for the claim that particle 1 is in state A, and similarly for other cases):

1. $A(1)$ and $A(2)$;
2. $B(1)$ and $B(2)$;

[1] Notice that there is also the option of rejecting the RV and interpreting those entities as individuals; we shall not discuss this option here, but see French and Krause (2006, chap. 4) and French (2015).

[2] We are not here claiming that this understanding of the statistics is not problematic or that it is the only alternative; rather, this is how the RV is typically presented, as a contrast between the classical and the quantum case.

3. $A(1)$ and $B(2)$;
4. $A(2)$ and $B(1)$.

All these possibilities are assigned the same weight, that is, $\frac{1}{4}$.

The fact that permutations do give rise to distinct states, as seen in cases 3 and 4 above, is typically accounted for in terms of the fact that classical particles are individuals. Of course, if such particles were discernible somehow, then their permutation could reasonably be seen as giving rise to distinct states. However, there is a sense in which classical particles may be taken as indiscernibles: classical systems may share all their intrinsic or state independent attributes. Even when this is the case, the classical statistics distinguishes between situations 3 and 4. Then, how to account for such a distinction?

It is here that individuality enters the stage: a permutation gives rise to distinct states precisely because those particles are individuals. There is something accounting for their numerical difference, and making the case that the two situations are different: the particles' individuality. There are many ways to account for such individuality without having to appeal to discernibility by intrinsic properties (which, as we have seen, fails in the classical case). The most typical option appeals to the fact that classical particles have a unique trajectory in space-time once an assumption of impenetrability is adopted. With that, each particle has a unique space-time trajectory, which may be regarded as conferring individuality to it (see French and Krause 2006, chap. 2).

In quantum statistics, on the other hand, permutations of indistinguishable particles are not counted. This gives rise to permutation symmetry and the alleged loss of identity we have been discussing. Usually, the formalism of orthodox QM uses symmetrization postulates: symmetric and anti-symmetric vectors/functions express the lack of identity of particles. For an illustration, let us consider two systems labeled 1 and 2 distributed in two possible states a and b, we can have the following possibilities:

1. $|\psi_1^a\rangle|\psi_2^a\rangle$;
2. $|\psi_1^b\rangle|\psi_2^b\rangle$;
3. $\frac{1}{\sqrt{2}}(|\psi_1^a\rangle|\psi_2^b\rangle \pm |\psi_2^a\rangle|\psi_1^b\rangle)$.

In fact, we have two different kinds os statistics here: Bose-Einstein (BE) for bosons, and Fermi-Dirac (FD) for fermions. The difference comes in the third possibility, bosons have the "+" sign, and fermions have the "−" sign. Also, for fermions only this third possibility obtains, they cannot be distributed according to the first two cases for they cannot be in the same state, they do obey the Pauli Exclusion Principle. Notice that, as in the classical case, to write the vectors we had to label both particles and states. But this does not run counter to the alleged loss of identity? To grant that the labeling on particles has no effect, we use symmetric and anti-symmetric vectors, for bosons and fermions respectively, adding also the Indistinguishability Postulate below (more on reference and labeling in the next section, when we deal with individuation). This is of course a mathematical trick, for what matters for physics is that the expectation value of the measure of any observable $\hat{O}$ for the system in the state $|\psi\rangle$ does not change after a permutation of the labels of the particles. Being P a permutation operator and $|\psi_{12}\rangle$ the vector state for particles 1 and 2, we express this by means of the Indistinguishability Postulate:

$$\langle\psi_{12}|\hat{O}|\psi_{12}\rangle = \langle P\psi_{12}|\hat{O}|P\psi_{12}\rangle$$

The topic of individuality for quantum particles is related to how we understand this postulate. The usual reading, attached to the Received View, regards it as a restriction on the states: there are only symmetric and anti-symmetric states. In this case, only bosons and fermions are possible, and since the operations representing observables always give as a result a vector of the same symmetry type as the one to which it was applied. So, the particles are regarded as non-individuals, nothing can be made to distinguish them, there is nothing there to account for a permutation that could make for a distinct state before and after the permutation. It is this reading of the Indistinguishability Postulate that traditionally underpins the statements advancing that quantum particles loss their identity.

However, it can be argued that this is not the only one reading of the postulate. There is an alternative way of reading it, as imposing a constraint on the observables: only observables commuting with the permutation operators are allowed. In this case, the asymmetric states are not banned, they exist but are inaccessible for particles whose states are represented by vectors of the other symmetry types, since the particles are always in symmetric or anti-symmetric states, and no operator shifts them to some of the asymmetric states. So, in this case, it would be possible, at least in principle, to distinguish the particles (for the distinguishing features would not be observable), and they can be seen as individuals, in some sense (see French and Krause 2006, chap. 4 for these possibilities).

But now comes the question: how can we understand this individuality? In general, it has been argued that the individuality of quantum particles will have to be grounded in some kind of Lockean substratum or non-qualitative *haecceity*. Since particles can be absolutely indistinguishable (and this can be rigorously argued for), it has been argued that the Principle of the Identity of Indiscernibles (PII), famously stated by Leibniz, according to which indiscernible items are identical, fails in quantum mechanics (see French and Krause 2006 for details). Along with the failure of PII goes also the possibility of grounding the individuality of the particles in some set of properties belonging to them. That is, the so-called bundle theories of individuation, according to which what characterizes an individual is a subset of its properties are ruled out in quantum mechanics and so, one must look for help in substrata or in some form of haecceitism.

So, it seems that our options are: accept quantum non-individuality and go on to explain this lack of identity that characterizes it, or take the individuals route, and adopt some kind of principle of individuation which have always been dubbed as mysterious, to say the least, in the history of philosophy. Here, we shall discuss only the non-individuals option. As a first step, we shall disentangle three notions which we have not been very careful to distinguish in the above discussion: identity, individuality, and individuation.

Identity, Individuality, Individuation

We have seen that the 'new' quantum statistics provide the main motivation for the informal claim that quantum particles lost their identities. But that concerns only part of the understanding of what is going on in the physics, and physics, by itself, does not provide for a unique metaphysical characterization of the metaphysics (metaphysics is said to be underdetermined by physics; see also French and Krause 2006, chap. 4). In order to provide for a possible understanding of what is going on in metaphysical terms, it is time to present carefully the relevant metaphysical concepts to deal with the above problems: identity, individuation, and individuality. Another concept that would be relevant for us is identity over time, but we shall not address the issue directly here.

We shall briefly discuss each of the three concepts, and explain the meaning of each expression. Our aim is to attempt to do so in the most uncontroversial terms as possible, because we would like to allow that distinct possibilities of combining those concepts remain open; distinct views on their relationship, then, will correspond to distinct views on identity and individuality. In particular, our main goal is to remain completely neutral as to the relation between identity (a logical notion) and individuality (a metaphysical notion). 'Individuation' shall serve as an umbrella term for diverse epistemic notions of separating an entity, singling it out and discerning it from other entities for the sake of linguistic reference or perceptual attention. The idea is that the epistemology, thus understood, needs have no impact on the metaphysical notion of individuality, although one may enforce one such relation and try to understand the metaphysical notion through some rendering of the epistemic notion (see section "Non-individuals with Identity").[3]

Identity

Identity is taken by us to be a relation between objects. As the tradition goes, identity statements are statements of the form '$a = b$', asserting that objects a and b are one. Such statements are true only when we are dealing with one and the same item as relata. It is just a matter of distinct forms of referring to it as either a or b. This is only a heuristic clarification, of course, not a formal explanation.

Formally, identity is a relation whose understanding depends on the language and the semantics employed. Basically, when it comes to logic, most philosophers adhere to a first-order characterization of the relation of identity. This is due to typical Quinean admonitions against the use of higher-order logic and set theory. The first-order axioms for this relation are well known:

[3] A small note on terminology: individuation is typically taken as synonym for individuality. Here, we distinguish both notions: individuality, as we mentioned, is a metaphysical feature of an entity, while individuation concerns an epistemic act of agents. We hope that the similarity of words won't cause any confusion.

Reflexivity $\forall x(x = x)$
Substitution $x = y \rightarrow (\alpha \rightarrow \alpha[y/x])$, with the known restrictions.

As it is known, these axioms allow for unintended interpretations of identity, where the meaning of the symbol = is not the identity relation over the domain of interpretation, understood as the diagonal of the domain of interpretation (see da Costa and Bueno 2009, pp. 186–187 for a sketch of the argument).

Here we develop a little further da Costa and Bueno's suggestion to show that numerical identity cannot be characterized by a purely syntactical approach. Their example is a simplification of Hodges' (1983, p. 64) (but see also Mendelson 1997, p. 100). In a general setting, it says that for every structure that interpret our first order language with identity having what Hodges call "standard identity" (nothing more than the diagonal of the domain), we can find an elementary equivalent structure which also models (in particular) identity, but where the relation that interprets this concept is not standard identity. So, identity cannot be characterized on purely syntactical grounds. Let us see the details.

Suppose we have a first-order theory with identity and let $\mathfrak{A} = \langle D, R_i \rangle$ be a model for the theory, where the binary predicate of identity is associated the identity of D, namely, its *diagonal*, $\Delta_D = \{\langle x, x \rangle : x \in D\}$ (so it is a *normal* model Mendelson 1997, p. 100; Hodges 1983 calls it *structure with standard identity*). Let $a_1, \ldots, a_n$ be elements not belonging to D. Now we construct a new structure $\mathfrak{A}' = \langle D', R_i' \rangle$ defined this way: to each element $a \in D$, we associate n new ordered pairs $\langle a, a_i \rangle$ $(i = 1, \ldots, n)$. The set D' is then formed by these pairs. Furthermore, to each k-ary relation R in $\mathfrak{A}$, we associate a k-ary relation R' in $\mathfrak{A}'$ and impose that the k-tuple formed by $\langle a^{(1)}, a_i \rangle, \ldots, \langle a^{(k)}, a_i \rangle$ satisfies R' if and only if the k-tuple $a^{(1)}, \ldots, a^{(k)}$ satisfies R. So we are extending all the semantic features of our theory to the new structure. Now, on D', we define the following relation, which can be proven to be a congruence:

$$\langle a, a_i \rangle \equiv \langle b, a_i \rangle \text{ if and only if } a = b.$$

Then the new structure is elementarily equivalent to the original one, and in particular it models the predicate of identity. However, the structure $\mathfrak{A}'$ is not a normal model for the theory, although it can be "contracted" to a normal one (as shown in Mendelson 1997, p. 100 and Hodges 1983, pp. 65–6). In other words, from the point of view of the first-order language, we cannot distinguish between the two structures.

Notice that there are three notions of identity going on here: the identity symbol in the object language, the identity as a diagonal of the domain of interpretation, and the identity relation of the metalanguage, which is used to talk about the other two. We shall come back to these distinctions very soon, given that this has important consequences on how to understand identity. On what concerns the first-order characterization of the meaning of the symbol of identity, one can only grant the intended interpretation if one stipulates, in the metalanguage, that the identity sign is always going to be interpreted in the diagonal of the domain (that is, in technical terms: we stipulate that we are dealing only with normal or standard interpretations of this symbol).

What is more relevant for us at this moment is that these basic properties of the sign of identity (reflexivity and substitution) allow for a minimal characterization of the logical notion. The relation of identity, as minimally characterized, is compatible both with views that attribute to identity an important metaphysical role, as well as with views that consider it to be metaphysically neutral. The basic properties of identity underlying both claims is the same. As we shall see later, from a metaphysical perspective, identity by itself requires nothing of a metaphysical character in itself, although it is also compatible with distinctively metaphysical interpretations that associate identity with heavy metaphysical machinery. The characterization that first-order languages provide allow for some of the minimal properties of identity and also leave it open what else should be added, if something, both from a formal as well as from a metaphysical perspective.

Those properties of identity are also neutral on whether identity should be defined in terms of other notions, or if it is a primitive notion. Taken as a primitive (or even fundamental) notion, or as a defined notion, identity must satisfy at least those two properties. Whatever else is required of identity—that it is reducible to qualitative identity, for instance—is something that is added to those properties. The point is: something failing those properties is not identity.

We may explore the relation between this minimal characterization of identity and the related metaphysical issues by bringing in some of the issues that arise in discussions related to the fact that identity is not characterizable in first-order languages.[4] Notice that we have mentioned that such an attempt to characterize identity involves, in fact, three distinct notions of identity, operating at distinct levels. This brings important questions to our very understanding of identity and its relations with individuality and reducibility of identity. Two important and related issues are as follows:

First: identity seems to be presupposed in our very attempts to characterize identity (be it at first-order or at higher-order languages). The claim is that identity must be previously understood in the metalanguage if we are to understand properly those characterizations in the object language, and even if we are to understand why some attempts to characterize identity in the object language fail. In this sense, some have judged that identity is not only undefinable, but is also a fundamental feature of every conceptual scheme; it is a pre-condition for us to make sense of everything else (see Bueno 2014 for a defense of this view).

Second: being fundamental, identity would be applicable every time we speak—and its use would indeed be *required* if we are to make sense of what we say. Some have gone one step further and suggested that this would confer a kind of primitive, very thin notion of individuality for objects. The idea is as follows: there is a fundamental notion of identity, applying for everything, and the mere fact that we may always meaningfully say that

[4] There are troubles for higher-order languages too; see French and Krause (2006), chap. 6 for a general discussion.

some items are equal or distinct seems to confer a metaphysical power to identity (see, for instance, Dorato and Morganti 2013). Identity and individuality are intimately connected in this view.

Now, those are substantial points on the role of identity and its understanding, which must be disentangled. Our aim is to remain neutral about them for now, and recognize that they are additions to the very minimal notion of identity we are trying to present here. Of course, given that those additions are adopted by some, they must be clearly discussed and articulated; what is relevant for us is that those additions are not encapsulated in identity itself, as minimally characterized. There are alternative understandings of the meaning of identity which do not require such additions (their merits must also be assessed, of course), and depending on how one takes those issues, distinct sets of possibilities for the notions of individuality and individuation will also arise.

The claim that identity is fundamental (and not eliminable), for instance, may be countered by an eliminativist (reductive) approach (see the discussion in Shumener 2017). Some such approaches use Leibniz's law, reducing numerical identity to qualitative identity (also called sometimes indiscernibility or indistinguishability):

$$x = y \leftrightarrow \forall F(F(x) \rightarrow F(y)).$$

Here, items are identical if and only if they share every property of the appropriate kind.[5] There are troubles with this approach, sure, but we mention it because it is regaining currency among philosophers of quantum mechanics, mainly among those defending that quantum particles are weakly discernible (we shall discuss this issue soon; for more on weak discernibility, identity, and Leibniz's laws, see Muller and Saunders 2008 and Caulton and Butterfield 2012). Another eliminativist approach may be advanced in which identity is understood as not being fundamental, but only as a projection of our cognitive apparatuses on reality, in a Humean sense of projection, just as Humeans do for causality. In this sense, identity is the result of a kind of mental construction, not a condition for the understanding of concepts or a metaphysical feature of reality conjoined with individuality.

Also, the relation between identity and individuality may be resisted. Even if identity is fundamental, it needs not have any metaphysical content; as Bueno (2014) contends, an empiricist may adopt the thesis that identity is fundamental *and* metaphysically deflated. In particular, identity by itself needs not confer individuality. Notice also that, on the other hand, typical Leibnizian reductions of identity are involved on a metaphysical view of individuality according to which items are individuated by their properties. However, it is not clear that a Leibnizian reductionist must adopt such a relation with metaphysics, and, also, for those willing to avoid such a

[5] 'The appropriate kind' here means that distinct versions of the principle are obtained according to the kind of properties allowed in the range of F. Three distinct versions are more prominent: (1) F ranges over every property and relation; (2) F ranges over every property and relation, except for spatio-temporal ones; (3) F ranges only over non-relational properties. See French (2015) for a discussion.

direct relation, there remains also the projectivist view, where identity may be seen as directly related to epistemology, not with any substantial metaphysical thesis on individuality.

In brief: a fundamentalist about identity may be deflationist about the metaphysical content of identity, or else have it playing a relevant role in individuality. On the same issue, the eliminativist about identity may have identity playing a role in individuality, or else hold that identity is unrelated to it.

Individuality

We have now briefly discussed what we shall mean by identity. That is a kind of logical relation, which may be separated from the metaphysical issue of individuality. Focusing on individuality now, to answer the question of what confers individuality for an item *a*, is to provide for another entity *b* and a relation between these two entities that accounts for what *a* is, and, obviously, that it does so only for *a*. As Lowe (2003, p. 75) put it, an individuation principle for an object is "whatever it is that makes it the single object that it is—whatever it is that makes it *one* object, distinct from others, and the very object that it is as opposed to any other thing". In this sense, individuality may be related with a substantial role for identity, but it need not (it will all depend on how one is framing the principle of individuality). This brings a whole bunch of questions which we shall try to make clear here in a rather schematic way, and which we shall address in our further discussion of quantum non-individuality.

First of all, 'individual' is not to be confused neither with 'particular item' nor with 'object'. Although most philosophers deal with the question of individuality for particular concrete items, such as Socrates and umbrellas, there may be issues about the individuality of universals, for instance, or about the individuality of particular items that are not objects, such as tropes. We shall be concerned here only with the issue of individuality of the particular items, generally called concrete particular objects, not to be confused with so-called abstract particulars, such as tropes. In this sense, given a principle of individuality, it may make complete sense to ask whether a particular concrete item is an individual or not. If the item is a particular but is not an individual, that may be understood as meaning that the item is a non-individual.

As a result of this first clarification, particulars may be individuals or not, and we leave it open whether the notion of object will coincide with the notion of particular concrete object (see Lowe 1998, chap. 3 for a discussion of particulars and a classification that does not colapse particulars with individuals and with objects). Whenever we use 'object' for a particular concrete item, with no further clarification, it is much in a neutral sense of the word that it is being used, much as the same as 'item' or 'entity'.

The second point that will be relevant for us is that a theory on the individuality of something needs not be the same as a theory of the constitution of the particular items. As Demirli (2010, p. 2) puts it:

> In answering the internal constitution question, we may begin an inquiry about the various categories that go into the composition of individual substances and hope that at the end of this inquiry we will come up with a list of ingredients that constitute various individuals. Just as a certain recipe in a cook book provides us with a list of ingredients and instructions for mixing these ingredients together, we may maintain that the list or the recipe of individual substances — God's recipe book, so to say — will tell us what items from various categories are used, and how these items are combined.

In this sense, a theory of the constitution or composition requires that we understand particular items as composed of other particulars, much in the same way as their ingredients; the ingredients may be understood as parts of the particular in a mereological sense, but they need not be so understood. Good examples are the so-called bundle theories, according to one version of which a particular is a bundle of universals (the properties it instantiates), and also the typical understanding of the bare particulars or substratum approach, according to which a particular entity is composed not only of universals, but also of another ingredient which is a particular, the bare particular or substratum, which works as a peg on which properties are hanged. In both cases, instantiation of a property P is understood in terms of the property P being one of the ingredients composing the entity.

These two theories (bundle and substratum) are also typically understood as theories of individuality: they conflate the constitution with that which attributes individuality. What makes a particular item precisely that item that it is? The bundle theorist answers: the specific universals that are bundled together by a co-presence relation! Distinct bundles are distinct individuals, and vice versa, distinct individuals must *be* distinct bundles, so that identity may also be understood in a reductive manner, and enter into the equation too. Alternatively, the substratum theorist would say: the individuality of an individual is accounted for by the substratum involved in its composition. It is the fact that each individual has its own substratum that accounts for numerical diversity and the fact that numerically distinct individuals are present.

In order to achieve such an identification between individuality and constitution, it is typically admitted a thesis on the identity of components implying identity of entities, the *Constitutional Identity Thesis* (CTI) (see Demirli 2010, p. 2):

CTI: If two entities have the same constituents, then they are numerically the same.

Of course, one may deny the CTI, while still embracing a thesis of constitution. We may well have that numerically distinct entities be constituted by the same components (see the discussion in Demirli 2010). That would require, of course, that their numerical distinctness be grounded by something other than its constituents, and would also require that we give an explanation on how to account for their individuality (if those items are thought of as individuals) in such a case. One good option would be to consider them as non-individuals: numerically distinct entities having nothing to account for their individuality, given that it is possible for two entities to have the same constituents (see the discussion in Arenhart 2017). Perhaps the main idea gets clearer when we consider the CTI in contrapositive form, let us call it CTIC:

CTIC: If two entities are numerically distinct, then they have at least one distinct component accounting for this difference.

The suggestion that CTIC may fail could prove very useful as an account of non-individuality, given that it allows for items being somehow indiscernible (on what concerns their components) but still not being numerically the same, a situation that bears close resemblance with the one described by quantum statistics and by standard chemical elements.

Besides avoiding confusion between composition and identity, and composition and individuality, here we shall follow Lowe (2003) in claiming that individuality is a metaphysical relation of explanation. A principle of individuality explains what is it that makes the other entity an individual. For instance, bare particulars are individuality principles that make precisely this. An individual is precisely the individual it is in virtue of the bare particular it has. The bare particular of Socrates explains why that entity is Socrates, and not any other individual. The same kind of reasoning could be employed using haecceities, or for bundle theories of individuality. In this sense, one cannot claim that a symmetric relation between entities a and b may be employed to individuate them. Neither does a explain $b's$ individuality, nor does b explain $a's$ individuality. This will be relevant when we discuss weak discernibility relations.

Before we proceed, let us get once again clear on one of the explanatory tasks of an individuality principle: it explains the numerical diversity of individuals. This, notice, goes only in one direction, from individuality to numerical distinctness (IN):

IN If a and b are numerically distinct individuals, then their individuality principle may be employed to ground their difference.

If two entities are individuals, then their individuality principle may be used to explain their numerical diversity. Notice that we do not mean that only the individuality principle does that explaining. For instance, one could well explain the difference between the individuals Socrates and Plato by the color of their T-shirts, but that is certainly not what accounts for their individuality.

What is also relevant for our purposes later is that the implication from individuality to an explanation of numerical distinctness does not work the other way in an even stronger sense, from numerical distinctness to an individuality principle (NI), at least not necessarily:

NI If two entities are numerically distinct, then, there is an individuation principle to explain their distinctness.

This may fail because, depending on which principle of individuality is chosen, two entities may well be numerically different without even being individuals. This should open the possibility of numerically distinct non-individuals, of course. So, the relation between identity and individuality is a delicate one. What is relevant for us is that once this implication is clear, then, it is open for us to have distinct objects without it being required that an individuality principle be there to account for such a diversity. Of course, some approaches, like that of Dorato and Morganti (2013), which we commented on earlier, also assume this more controversial direction of dependency, making a closer relation between numerical difference and individuality. We shall not assume that this more controversial relation holds, unless it is precisely stated when it comes to discuss it.

Individuation

Now that the metaphysical notion of individuality has been distinguished from many associated and related notions, such as identity and composition, and it was seen to be an explanatory relation, it is time to address another, closely related issue, now dealing with our ability to separate or single out objects in our sensory field for the sake of sensorial focus or reference. 'Individuation' is the word we shall use to refer to this purely epistemic correspondent of individuality. It concerns not the metaphysical ingredient doing the job, but rather our abilities to separate things from their environment, discern them as a unity, make them the object of our attention or of our reference.

The adaptation of an example from Lowe (2003, p. 75) will help us clarifying what is at stake: consider the Margin-winged stick insect,[6] an insect that looks very much like an Eucalyptus twig. One of the greatest features of this insect, of course, is its similarity with the eucalyptus. Most of us would not be able to identify any particular Margin-winged stick when looking at an eucalyptus, even if there is one such insect there. However, a trained specialist is able to identify the insect even on such situations. We say that the specialist has individuated the insect, by managing somehow to isolate it from its environment, discerning it from the leafs of the eucalyptus. Individuation is this epistemic act. Of course, independently of how well we fare on individuation, individuality, in the case of the insect, is granted by a metaphysical principle of individuality. Any failure in individuation is only an epistemological shortcoming, not a metaphysical shortcoming. This illustrates that the metaphysics may be separated from the epistemology: there may be cases where we cannot individuate some individuals.

Our main claim will be that the other direction also holds: there are some cases where we may have individuation without the item in question being an individual. The epistemology does not fail, but there is no metaphysical principle to account for the individuality. To elucidate the distinction of the two notions, we begin with a very interesting example provided for by Dalla Chiara and Toraldo di Francia in (1992, p. 163), where there is a process of individuation going on (singling out two objects as the focus of our perceptual attention, the attribution of names and attempt of reference) without the *existence* of two individuals accounting for that attribution. In short, the case is as follows: some years ago, what looked very much like two quasars were discovered in the sky. They had all the same features, and although they were very close one to the other, there were clearly two spots seen at the same

[6] Ctenomorpha marginipennis.

time. Names were even attributed—$Q33$ and $Q34$—to them, until it was suggested that they were in fact one and the same. The explanation was that due to relativistic effects on light, rays coming from one quasar were arriving to us as two spots. So, individuation is going on without the need of individuality.

Closer to our case in this paper, something similar (individuation without individuality) may happen under some accounts of non-individuality in quantum mechanics, as we shall see. For instance, consider that we are making a measurement on a single particle in a quantum experiment. By seeing the effect of the entity on a bubble chamber, or on a photographic plate, we have indirect access to the particle (this is what Dalla Chiara and Toraldo di Francia called ‘mock individuality’ in Dalla Chiara and Toraldo di Francia 1992, p. 266). It also happens in cases of trapped particles, such as Astrid and Priscilla, famously trapped by Hans Dehmelt. Although we can somehow individuate Priscilla, point to it, label it temporarily, and confer a kind of unity for it, there is no need for all of these acts to be accounted for by an underlying individuality principle; all that may be involved, so far as the situation goes, is individuation, granted by the context where the particle is trapped (see also Krause 2011 for further discussion and references).

Something similar occurs in cases of counting the electrons of an Helium atom through a process of ionization. Here is how Domenech and Holik (2007, p. 862) explain it:

> Put the atom in a cloud chamber and use radiation to ionize it. Then we would observe the tracks of both, an ion and an electron. It is obvious that the electron track represents a system of particle number equal to one and, of course, we cannot ask about the identity of the electron (for it has no identity at all), but the counting process does not depend on this query. The only thing that cares is that we are sure that the track is due to a single electron state, and for that purpose, the identity of the electron does not matter. If we ionize the atom again, we will see the track of a new ion (of charge $2e$), and a new electron track. Which electron is responsible of the second electron track? This query is ill defined, but we still do not care. Now, the counting process has finished, for we cannot extract more electrons. The process finished in two steps, and so we say that an Helium atom has two electrons [...].

Notice: the process of counting may be performed without mention to identity or individuality of the electrons (Domenech and Holik seem to conflate identity and individuality). One may provide for an individuation of the extracted electrons, by referring to the first and the second electron extracted, but that, by itself, means nothing about individuality yet. If the electron is an individual, it is due to an individuality principle, not due to any epistemic feature of individuation during the extraction process. On the other hand, if the electron is not an individual, as we mentioned, the individuation also does nothing to prevent such non-individuality. It remains a non-individual (see Arenhart and Krause 2014 for further discussions on this specific case).

The result is that we are able to understand the notion of individuation independently of the metaphysical notion of individuality. Both concepts may be kept separated. One may have a metaphysics of non-individuals, while still being able to account for individuation. As Lowe (2003, p. 92) remarked, that which accounts for the identity or difference of two items needs not be the same as that which accounts for their individuality. The same may be said about individuation: that which accounts for the individuation of an item (for instance, its discernibility from other items, its separability, or the fact that it is a unity), needs not be the same ingredient that confers individuality for it.

Schrödinger's Problem

With all those distinctions made, let's get started in getting the idea of a non-individual clear. The first way to metaphysically dress the RV is the one based on a direct reading of Schrödinger's claim that identity does not make sense for quantum entities. Given that it is the view that is encapsulated in the so-called non-reflexive systems of logic (more on these soon), let us call this one the "non-reflexive" approach to non-individuality. We shall divide this section in two short subsections, one dealing with the metaphysics, the other dealing with the logic. The logic is important here, because it is felt that it provides for the clarity needed for such a radically revisionary thesis, on both metaphysical and conceptual grounds (see French 2015, Sect. 5).

The Metaphysics

Metaphysically, this view will require a close relation between identity and individuality. Identity, being a logical concept, will be loaded with a metaphysical role in individuality. Before we proceed, let us get clear on what kind of principle of individuality is being employed here.

We have already briefly discussed compositional theories, and among them, the theories of substratum. As we mentioned, those theories are commonly employed as theories of the composition and of the individuality of an item. What is attractive about those theories, at least for its defenders, is that it allows for items being numerically distinct while also being qualitatively indiscernible. That is, while *a* and *b* may be composed of the same properties, they still may count as two entities because of their distinct substrata, which is a further ingredient.

Now, of course, a substratum as an extra ingredient poses difficult challenges. It must not have properties, although it bears the properties, it must be individuated somehow, although not by anything else, it must be empirically inaccessible, and so on; many other questions add to the mystery (see Lowe 2003). In order to avoid those difficulties, while retaining the possibility of having numerically distinct items being qualitatively indiscernible, a distinct approach proposes that particulars are individuated not by a particular substratum, but by another property, a haecceity. Haecceities are understood as non-qualitative properties (they do not contribute for the discernibility of a particular), which are uniquely instantiated by a particular. This allows for a particular being individuated by a property, uniquely possessed by that individual, while leaving the issue of discernibility to be decided by the qualitative properties.

For an illustration, according to this approach Socrates is an individual in virtue of instantiating his haecceity. A haecceity is understood as the property of being that particular individual; in the case of Socrates, it is the property of 'being Socrates'. So, each individual has its own haecceity. It has many similarities with the substratum approach, so that French and Krause (2006, chap. 1) follow Heinz Post's usage and call those theories of individuality 'Transcendental Individuality' (TI). They have their name precisely because their individuality is attributed by something transcending the qualities of the individual.

Differently from the substratum theory, however, a haecceity needs not (and generally is not) involved in a theory of composition. Most adherents of haecceities do not believe that a haecceity is another ingredient composing the particular, so that the view needs not be wedded to a theory of composition.[7] Also, due to its formulation as a property, 'the property of being precisely this individual', it seems to allow for a specific relation with identity: Socrates's haecceity, for instance, would be 'to be identical with Socrates'. The claim that everything is an individual would amount to the claim that everything is self-identical. So, here we have a very strong relation between identity and individuality.

Given this stage setting, if we are to have this theory of individuality allowing also for non-individuals, we will have to provide for two restrictions: (1) on a metaphysical level, to grant that some concrete particulars do not bear haecceities (so, they are not individuals), (2) that the relation of identity does not apply to every thing (so that haecceities are not applied to every thing). Of course, both restrictions are completely related on the view we are discussing, given that haecceities are understood in terms of identity.

Here is how French and Krause (2006, pp. 13–14) express both the relation between haecceities and identity, and the prospects of the failure of the relation:

> ...the idea is apparently simple: regarded in haecceistic terms, "Transcendental Individuality" can be understood as the identity of an object with itself; that is, '$a = a$'. We shall then defend the claim that the notion of non-individuality can be captured in the quantum context by formal systems in which self-identity is not always well-defined, so that the reflexive law of identity, namely, $\forall x(x = x)$, is not valid in general.

French and Krause go on:

> We are supposing a strong relationship between individuality and identity ...for we have characterized 'non-individuals' as those entities for which the relation of self-identity $a = a$ does not make sense. (French and Krause (2006, p. 248))

So, given that the principle of individuality is a form of TI, a haecceity, the non-individuals are the items having no such TI, they lack a haecceity (see also Arenhart 2017 for further discussion). Also, given that haecceity is expressed in general by the reflexive law of identity, non-individuals, consequently, will have to be entities

[7] Of course, one may try to spell the theory of substratum as a theory of individuality without being also a theory of composition.

without identity, capturing the Schrödingerian intuition presented in a quote in the beginning of the paper.

The connection with the physics, as explained in the statistics, is also rather direct. Recall that quantum particles obey permutation symmetry. The most common opposition between the classical case and the quantum case requires that we distinguish between what to do with permutations. While permutations of the labels of classical particles do give rise to a distinct state, permutations of quantum particles do not. The claim underlying this approach seems to be that quantum statistics require that quantum entities have no individuality, for otherwise a permutation would have to be regarded as giving rise to a distinct state. Of course, once the items in case have individuality, it seems to make sense that we speak of item a in state 1 and item b in state 2, or vice-versa. As a result, quantum statistics would not work, and quantum non-individuality seems required (this kind of analysis, of course, may be resisted, see French and Krause 2006, chap. 4; what matters for us is that there is a way to motivate the adoption of a metaphysics of non-individuality coming from the physics, even if this approach is not itself mandated by the physics).

Not having haecceity and identity, of course, will require distinct explanations for a whole bunch of ideas, including, perhaps, those closely associated with individuation, such as most prominently, counting and the trapped particles cases (for further challenges, see Bueno 2014). In general, counting a collection of entities involves the use of identity. Given that under this approach identity is metaphysically committed with individuality, counting will have to be explained in alternative ways. Something similar happens with the cases of trapped particles. It seems that we are able to distinguish a trapped particle from every other particle. So, what prevents us from attributing a kind of difference from every other item, thus involving also identity?

We shall not enter in the details of the discussion here (but see Krause 2011 and Arenhart and Krause 2014). It seems that a revision in individuality through haecceity, allowing for non-individuals, is compatible with revisions in those concepts too. The revision may be achieved through the reform of part of our logic, that is, by the adoption of non-reflexive logical systems. They allow us to give precise definitions of counting, for instance, that do not require the use of identity (and hence, according to this approach, of individuality, too).

Non-reflexive Logics

Here we informally present only the strongest system of non-reflexive logic: the quasi-set theory $\mathfrak{Q}$. This is a ZFU (Zermelo-Fraenkel with Urelemente) style set theory, but with two kinds of atoms.

Our system $\mathfrak{Q}$ will have all the usual logical vocabulary for first-order logic without identity: propositional connectives ($\neg, \rightarrow$), quantifiers ($\forall$), and a denumerable collection of variables $x, y, z, \ldots$. It is important to emphasize that there is no identity symbol, for identity will be a defined notion, whose definition will have limited applicability, as the view under discussion requires. The list of primitive non-logical symbols of $\mathfrak{Q}$ is the following one:

(i) the binary relations $\in$ for pertinence and $\equiv$ for indistinguishability;
(ii) the unary predicates m, M and Z, meaning m-atoms, M-atoms and classical sets respectively.
(iii) the unary function symbol qc, for quasi cardinal.

The intended interpretations of m-atoms are the quantum non-individuals, items for which identity must not make sense. M-atoms represent usual objects (classical Urelemente), items for which identity applies, and things to which the predicate Z applies are the sets in $\mathfrak{Q}$ that represent classical sets of ZFU. Terms are individual variables and expressions of the form $qc(t)$, where t is a term. Formulas are defined in the usual way. Now some useful definitions are in order:

(i) $Q(x) := \neg(m(x) \vee M(x))$ (x is a qset)
(ii) $D(x) := M(x) \vee Z(x)$
x is a *Ding*, a "classical object" in the sense of Zermelo's set theory, namely, either a set or a M-atom.

So, quasi-set theory has two kinds of atoms and qsets, collections having these atoms and other collections as elements. In this aspect, $\mathfrak{Q}$ is just like the usual theories with *urelemente*. The main difference concerns the behavior of the m-atoms: since this system is intended to capture the idea of a lack of haecceity for m-atoms, in the formal system we shall build, statements of the form $x = y$ or $x = x$ are simply not available when x and y are m-atoms; that is, the items that denote quantum particles (in our intended interpretation) are not relata of identity. To achieve this, we advance the following definition of identity:

Definition 1 $x =_E y := (Q(x) \wedge Q(y) \wedge \forall z(z \in x \leftrightarrow z \in y)) \vee (M(x) \wedge M(y) \wedge \forall z(x \in z \leftrightarrow y \in z))$ *(Extensional identity)*

Notice, the definition does not apply to m-atoms. There is nothing to be said about their identity or diversity. This is so in order to capture Schrödinger's claim that identity makes no sense for quantum entities, and also the intended metaphysical understanding of non-individuality through the associated claim that haecceities do not apply for everything.

The next important point we would like to mention concerns a relation of indistinguishability. Permutation symmetry implies that quantum entities are not discernible by any properties whatsoever. An obvious strategy for introducing the relation of indiscernibility would be to require that whenever $\alpha(x)$, where α is any formula with x free, would imply that $\alpha[y/x]$, that is, full substitution is allowed for indiscernible items (with the usual care to avoid clash of variables). Also, of course, we would like to have the indiscernibility relation reflexive, because everything is supposed to be indiscernible from itself. So, indiscernibility would be a reflexive relation allowing for full substitutivity. This sounds nice but there is one major problem with this idea: what is being introduced is precisely the same set of postulates for first-order identity! In order to avoid that indistinguishability coincide with identity for these items, endow the indiscernibility relation with a formal restriction: given that identity is an equivalence relation compatible with every other relation, indistinguishability, in our system, will lack this last property. Let us begin with some postulates:

($\equiv_1$) $\forall x(x \equiv x)$

($\equiv_2$) $\forall x \forall y(x \equiv y \to y \equiv x)$

($\equiv_3$) $\forall x \forall y \forall z(x \equiv y \land y \equiv z \to x \equiv z)$

These postulates ensure us that indistinguishability is an equivalence relation. As we commented above, this relation is not necessarily compatible with the other primitive predicates or relations. It seems plausible that indistinguishable objects should not necessarily be elements of the same qsets, so that indistinguishability is not compatible with membership. In $\mathfrak{Q}$ this holds for m-atoms, and it also grants that indistinguishability does not coincide with identity for these items, that is, if x and y are indistinguishable m-atoms, then being z a qset, we have that $x \in z$ does not entail that $y \in z$, and conversely. On the other hand, for other kinds of objects, identity and indistinguishability do coincide, and then indistinguishability is compatible with every relation, and in particular, with membership.

Once these basic ideas are in order, $\mathfrak{Q}$ just follows the usual set theories with atoms when it comes to grant the existence of collections. Postulates grant that a form of the pair axiom hold, the power set axiom, separation axiom, empty set axiom, and so on (the details may be seen in French and Krause 2006, chap. 7). In particular, what is relevant for us is that the M-atoms and the collections that do not include m-atoms, those satisfying the predicate Z, may be employed to develop inside $\mathfrak{Q}$ a classical system that behaves just like ZFU. So, as part of $\mathfrak{Q}$, we have classical set theory with atoms. In particular, inside the classical part of $\mathfrak{Q}$ we may develop the classical theory of cardinals. It is through these existing cardinals that we may attribute a cardinal also to qsets having m-atoms. This is achieved with a postulate:

(qc_1) $\forall_Q x(\exists_Z y(y = qc(x)) \to \exists! y(Cd(y) \land y = qc(x) \land (Z(x) \to y = card(x))))$.

(qc_2) $\forall_Q x(x \neq \emptyset \to qc(x) \neq 0))$.

The basic idea is that every qset has a quasi-cardinal, which is a classical cardinal, attributed by the function qc. When this qset is a classical set, the quasi-cardinal coincides with the classical cardinal. When the qset has m-atoms, then, the attribution must be made respecting the behavior of the operations over qsets. In particular, we are able to prove that singletons exists, that is, collections of objects indiscernible form x, which we represent by $[x]$.[8] Each qset $[x]$ has a sub-collection whose quasi-cardinal is 1, denoted by $[[x]]$. We call it the *strong singleton* of x. It has just one element, and we may think of this element *as if* it were x, but in fact, it follows from the definition that all we can know about it is that $[[x]]$ contains one item indistinguishable from x. In the cases in which x is not an m-atom we obtain the usual singleton, and we can prove that its element is x itself.

[8] Care must be taken here in order to separate $[x]$ from an already given collection z, so that $[x]$ is the collection of items indiscernible from x in z. This prevents singletons from being too big. For a full discussion see (French and Krause 2006, chap. 7) and (French and Krause 2010).

With these notions we are able to prove in $\mathfrak{Q}$ the theorem expressing the invariance by permutations:

Theorem 1 *(Invariance by Permutations). Let x be a finite qset such that $\neg(x = [z])$ and let z be an m-atom such that $z \in x$. If $w \equiv z$ and $w \notin x$, then there exists $[[w]]$ such that*

$$(x - [[z]]) \cup [[w]] \equiv x$$

In words: two indiscernible elements z and w, with $z \in x$ and $w \notin x$, expressed by their strong-singletons $[[z]]$ and $[[w]]$, are 'permuted' and the resulting qset x remains indiscernible from the original one.

Non-individuals with Identity

Now, while the non-reflexive approach to non-individuality and the Received View is the most traditional and well-known one, it is not the only possibility. Recall that in general lines the RV is the thesis that quantum entities are not individuals, and that the very idea of a non-individual needs not be articulated in terms of a lack of identity. In fact, given our previous discussion, there is a variety of options which are open for us to understand non-individuals, while still preserving the use of identity.

These approaches all benefit from the fact that that which confers numerical diversity to items needs not be the same thing that confers them individuality; that is, items may be different for reasons unrelated with their individuality principles (for a discussion with further options, see Arenhart 2017). That is, the approaches we shall deal with benefit from the fact that numerical distinctness does not imply individuality: facts about identity and diversity need not be facts about individuality. The advantage of separating an account of individuality and an account of identity relies precisely in the fact that we may revise the applicability of the theory of individuality without having to revise the applicability of identity. This allows for a much less revisionary approach than the non-reflexive one, given that the logic of identity and the mechanisms of individuation (discernibility, unity, separation, reference, and so on), may still be available.

The first approach to the RV (that is not also a revision of logic) which we would like to present benefits from the weak discernibility approach, and is suggested, although not directly, by Muller and Saunders (2008). Before we present it, a brief introduction in the terminology may be useful. As to the possible ways to discern two entities a and b, we have:

Abs a and b are absolutely discernible when there is an intrinsic property that one of them has, while the other does not have.

Rel a and b are relatively discernible when there is a relation R such that aRb or bRa holds, but not both.

Wea *a* and *b* are weakly discernible when there is a relation *R* that is symmetric (*aRb* implies *bRa*) and irreflexive (*xRx* fails for every *x* in the domain of the relation).

Entities that are not discernible in any of the former ways are said to be 'indiscernibles' (see Muller and Saunders 2008, and also Caulton and Butterfiled 2012). Quantum entities of the same kind certainly are not absolutely discernible and also not relatively discernible. However, they may be seen as weakly discernible. Two electrons in the singlet state, for instance, are weakly discernible by the relation "has spin in the opposite direction to". In fact, no electron has spin in the opposite direction to itself, and if the spin of *x* is opposite to the spin of *y*, then, certainly the spin of *y* is opposite to the spin of *x*. That allows us to ground the claim of numerical diversity: if a weakly discerning relation obtains, then, certainly we must have two objects.

However, the obtaining of weakly discerning relations, by itself, is not enough to ground individuality. Recall that individuality is an explanatory relation, which cannot be symmetric. In this sense, one item cannot be used to explain the other's individuality in a weakly discerning relation. We only go as far as numerical difference, without being able to attribute individuality to the items. Notice also that a weakly discerning relation does not allow for individuation: we cannot determine, in the case of the electrons in the singlet state, which electron is up, and which is down. The best we can say is that one is up and one is down.

The fact that a weakly discerning relation holds between quantum entities saves a version of the Principle of Identity of Indiscernibles. Recall that the principle requires that numerically distinct objects must have their distinctness grounded in some quality (this fits well with a reductive account of identity, thus, but is compatible with other non-reductive approaches to identity). While properties and asymmetric relations cannot ground such diversity in quantum mechanics, weakly discerning relations can. So, even if we cannot point to the particles (that is a problem of individuation, not of identity), we may have a numerically grounded difference. In this rather weak formulation, the PII resists in quantum mechanics.

Muller and Saunders' (2008) suggestion enters precisely here.[9] They combine these ingredients in a proposal which, we believe, is compatible with non-individuals. First, they define individuals as *absolutely discernible* objects, that is, objects having a property that allows them to be discerned from everything else. Now, this property is generally understood as being a physical property, not a haecceity or some non-qualitative property. That accounts for the explanatory role of individuality, and given that the approach to identity is reductive, also for the identity. Now, quantum particles are not absolutely discernible from other items, there is nothing in them allowing for such a discernibility. As a result, they are objects, but they are not individuals. Their numerical difference is accounted for, but their individuality is not.

Notice that although the PII is involved, this approach needs not be conjoined with a theory of composition. The particles in case need not be composed of relations and

[9] We are not suggesting that Muller and Saunders see themselves as providing a theory of non-individuals; our suggestion is that their definitions may be understood as a rendering of the RV.

properties (although one could, perhaps, try to provide for such a theory of composition too; see the discussion in Arenhart 2017). The relations just hold between them, as a quantum mechanical fact. This view puts discernibility in the center of the stage. Also, it meshes well with the statistics, because permutation symmetry may be understood as the fact that the relations obtaining between the quantum particles are all symmetric. There is also no property to provide for a difference before and after a permutation.

For another option on how to frame the RV, consider Bueno's approach to individuality in Bueno (2014). According to Bueno, identity is a fundamental, primitive relation, that is metaphysically deflated, thus, not by itself contributing to the individuality of anything. An individual is a particular object satisfying two much stronger conditions than mere numerical diversity:

Dis the item is discernible from every other individual;
Id the item is re-identifiable over time.

Notice that while one could appeal to weakly discerning relations to account for discernibility, the condition [Id] puts a heavy epistemological ingredient on the definition of individual, an ingredient which is not satisfiable in standard quantum mechanics. With this definition, identification is comprised in the epistemic notion of individuation: it is required that we identify something (single it out somehow), and then, at later instants of time, that we are able to re-identify it as being the same item. Bueno is not explicit about what is involved in re-identification, so, we shall take that what is meant is something along the lines we have described, which are very plausible demands for an empiricist.

However, for unobservable objects such as quantum particles, re-identifiability is not directly available. One could understand re-identification as the demand that we could, at least in principle, follow the trajectory of the individual at any given instant of time. As it is known, that is not available in standard quantum mechanics. In fact, even in Bohmian mechanics, where a trajectory is always present, it is a hidden ingredient (a hidden variable), so that the epistemic flavor of Bueno's definition would be lost. In fact, the trajectory in this context would work as a metaphysical ingredient unrelated to the available epistemology. So, on what concerns identity over time, there are also important distinctions between the metaphysics and the epistemology involved; under our interpretation, Bueno brings precisely the epistemic ingredient to be conflated with the metaphysical one: failing the epistemic ingredient, there is nothing else to be employed in order to grant individuality. In this sense, then, these theories would comprise non-individuals.[10]

[10] We are not claiming that it was Bueno's original goal to defend a theory of non-individuals; in fact, in Bueno (2014) he identifies the RV with the non-reflexive approach, and argues against it.

Conclusion

In this paper, we have provided for distinct ways to give some metaphysical flesh to the heuristic bones of the RV. Typically presented in rather vague terms, the RV merely says that quantum entities are not individuals, that they have lost their identities. However, nothing is said about how to formulate the metaphysically complex notion of individual and its failure in quantum mechanics. Hints are merely provided by the new statistics.

Here, to address these issues, we have distinguished the three core notions involved in attempts to provide for a theory of individuality: the concepts of identity (logical), individuality (metaphysical) and individuation (epistemological). These notions were provided with a rather minimal content, so that they could be employed in distinct combinations in order to provide for distinct theories of individuality and non-individuality as well. Some approaches put more weight in the identity, others in individuation. The whole point is that having those concepts clear in mind allowed us to provide for some examples of how to provide for metaphysical content for the notion of non-individual.

As we have seen, the most widely articulated and defended approach for the concept of non-individual is the non-reflexive one, as presented by French and Krause (2006). It originates on an interpretation of Schrödinger's claim that identity makes no sense for quantum particles; a close connection is provided for between identity and individuality through the use of haecceities as an individuality principle and its expression as self-identity. Now, while this provides for a clear determination when something is an individual (at least formally), that approach requires that failure of individuality should be accompanied by a corresponding failure of identity, which on its turn requires a revision of logic and many associated notions (think of naming, counting, quantification, isolating one entity, among others, which are typically related with identity). Non-reflexive systems of logic are presented in order to render the view with solid foundations, or, "philosophically respectable", as French puts it (see French 2015, Sect. 5).

While the association between identity and individuality, and lack of identity with non-individuality, has been widespread, and almost always identified with the RV itself, it is by no means the only option. In fact, some reject the RV due to its largely revisionary character on what concerns identity. In order to make clear that the non-reflexive approach is not the only one, and to dissociate the RV from one of its possible articulations, we have provided for two alternatives which keep identity intact, but define individuality in such a way that it is possible for quantum entities to fail them. One such approach was suggested by Muller and Saunders' (2008) definition of an individual. It requires that individuals be absolutely discernible from other entities. Quantum entities do not meet this condition, although they satisfy weaker forms of discernibility that grant them numerical diversity.

For another approach, relating the metaphysical notion of individuality with the epistemic notion of individuation, we have also briefly presented a proposal that may be found in Bueno's (2014). By relying heavily on the epistemic requirement of re-identification over time, Bueno puts such strong requirements on what can be an individual. Notice also that while his approach does not identify the metaphysical notion of individuality with the logical notion of identity, it makes identity an important ingredient of individuality, and what is more important, brings the metaphysical concept of individuality closer to the epistemic concept of individuation. By incorporating the re-identification clause for individuality, Bueno leaves open the door for items that do not meet this condition. That allows for non-individuals to come in, while identity still applies.

As a result of these distinct articulations, the RV may be seen as providing only for a kind of general guidance on what quantum entities should be, without providing for no specific metaphysical approach to non-individuals. All that is required is that the demands put by quantum mechanics (a form of indiscernibility by permutation invariance) be respected. Of course, those approaches have distinct merits, and the proper examination of which is better (if any), is an issue that we shall discuss in another place.

References

J.R.B. Arenhart, The received view on quantum non-individuality: formal and metaphysical analysis. Synthese **194**, 1323–1347 (2017)

J.R.B. Arenhart, D. Krause, Why non-individuality? A discussion on individuality, identity, and cardinality in the quantum context. Erkenntnis **79**, 1–18 (2014)

O. Bueno, Why identity is fundamental. Am. Philos. Q. **51**(4), 325–332 (2014)

A. Caulton, J. Butterfield, On kinds of indiscernibility in logic and metaphysics. Brit. J. Philos. Sci. **63**, 27–84 (2012)

N.C.A. da Costa, O. Bueno, Non-reflexive logics. Revista brasileira de filosofia **232**, 181–196 (2009)

M.L. Dalla Chiara, G. Toraldo di Francia, Individuals, kinds and names in physics, in *Bridging the Gap: Philosophy, Mathematics, and Physics*, ed. by G. Corsi, M.L. Dalla Chiara, G.C. Ghirardi. Lectures on the Foundations of Science (Kluwer Academic Press, 1992), pp. 261–284

S. Demirli, Indiscernibility and bundles in a structure. Philos. Stud. **151**, 1–18 (2010)

G. Domenech, F. Holik, A discussion on particle number and quantum indistinguishability. Found. Phy. **37**(6), 855–878 (2007)

M. Dorato, M. Morganti, Grades of Individuality. A pluralistic view of identity in quantum mechanics and in the sciences. Philos. Stud. **163**, 591–610 (2013)

S. French, Identity and individuality in quantum theory, in *The Stanford Encyclopedia of Philosophy*, ed by E.N. Zalta (Fall 2015 edn.) (2015), http://plato.stanford.edu/archives/fall2015/entries/qt-idind/>

S. French, D. Krause, *Identity in Physics. A Historical, Philosophical, and Formal Analysis* (Oxford University Press, Oxford, 2006)

S. French, D. Krause, Remarks on the theory of quasi-sets. Studia Logica **95**(1–2), 101–124 (2010)

W. Hodges, Elementary predicate logic, in *Handbook of Philosophical Logic—Vol. I: Elements of Classical Logic*, ed. by D.M. Gabbay, F. Guenthner (D. Reidel, Dordrecht, 1983) pp. 1–131

D. Krause, Is Priscilla, the trapped positron, an individual? Quantum physics, the use of names, and individuation. Arbor (Madrid. Internet) **187**, 61–66 (2011)

E.J. Lowe, *The Possibility of Metaphysics: Substance Identity and Time* (Clarendon Press, Oxford, 1998)

E.J. Lowe, Individuation, in *The Oxford Handbook of Metaphysics*, ed. by M.J. Loux, D.W. Zimmerman (Oxford Un. Press, Oxford, 2003), pp. 75–95

E. Mendelson, *Introduction to Mathematical Logic* (Chapman & Hall, Cornwall, 1997)

F.A. Muller, S. Saunders, Discerning fermions. Brit. J. Philos. Sci. **59**, 499–548 (2008)

E. Schrödinger, *My View of the World* (Cambridge Un. Press, Cambridge, 1964)

E. Schrödinger, *Nature and the Greeks and Science and Humanism, with a Foreword by Roger Penrose* (Cambridge Un. Press, Cambridge, 1996)

E. Shumener, The metaphysics of identity: is identity fundamental? Philos. Compass (2017). https://doi.org/10.1111/phc3.12397

H. Weyl, *The Theory of Groups and Quantum Mechanics* (Dover, 1950)

Chapter 16
Quantum Mechanics as a Semantic Problem

Hans Herlof Grelland

Introduction

Physics, like all science, has grown out of our desire understand the world. Today, physics has obtained a special position by providing a basis for other natural sciences like chemistry and biology. The most fundamental theory of present-day physics is quantum mechanics, including a variety of quantum field theories. Thus, quantum mechanics can be considered as one of the most important intellectual tools for exploring and understanding the physical world. This is true even if one holds that there exist other aspects of the wold which are beyond the methods and concepts of the natural sciences.

Quantum mechanics was formulated in 1925 as a theory of the microworld of the atoms, but it soon developed into a highly successful general theory, and it has been applied to an almost endless variety of problems, so far always giving answers that are in accordance with observation. Generally, the physicists have no difficulty in setting up the right equations, solving them, although sometimes with a considerable calculational effort and by introducing approximations, and finally comparing the calculated results with observations. One is also able to construct devices based on the quantum nature of matter, like solar cells and rudimentary quantum computers. We can safely say that the theory is well understood and extensively used *instrumentally*. However, the situation is quite different when it comes to quantum mechanics as a description and a means for understanding physical reality. This problem is demonstrated by the existence of a variety of competing interpretations of the theory, with no decisive argument or general agreement deciding which of them is the correct one. Even the most generally accepted interpretation, the Copenhagen interpretation, has been shown to be a common name for many dif-

H. H. Grelland (✉)
University of Agder, Kristiansand, Norway
e-mail: hans.grelland@uia.no

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_16

ferent positions (Camilleri 2009). The pioneers who developed the theory in the first place, Albert Einstein, Niels Bohr, Werner Heisenberg, Erwin Schrödinger, and Wolfgang Pauli, all disagreed more and less on the theory's physical interpretation. The agreement was on the mathematics of the theory and how to apply it instrumentally to experiment and observation. The problem was, and is, what this mathematics *means.*

There has since then accumulated a vast literature on this question, far beyond the capacity of one individual to read trough. However, there exists only a finite number of interpretations, each with highly competent followers, which have survived after ninety years of continuous discussion. Many of these were described by Jammer (1974), and some more recent attempts are added by Pykacz (2015). We can roughly divide these interpretations into two categories.

One category is the attempts at finding some "physical" description behind the mathematical formalism, trying to regain some place for visual imagination or physical intuition in this abstract mathematical theory. Early attempts of this kind are Schrödingers "matter waves", in which quantum systems are visualisable waves in space, or the notion of the wave-particle duality, in which a system in some situation can be visualised as particles, in other situations as waves, or hidden variable theory, in which quantum systems are classical systems with some peculiar additional features. I will classify these interpretations as *intuitionistic*. Intuitionism is roughly the view that the meaning of the mathematical expressions in a physical theory depends on the possibility of associating these mathematical entities with mental images of the physical reality they represent and that somehow can be related to the world of human experience. The ability to do so is often called physical intuition. Such images can be those of particles or waves moving in three-dimensional space.

The second category of interpretations are usually called instrumentalist or operationalist. It takes as its starting point the fact that we know how to connect quantum mechanics to experimental observations. Thus, the mathematical formalism is considered as nothing but a formal description of the correlations between observations or operational procedures with a measurable outcome. In this kind of "interpretation" one does not consider the theory as a tool for describing the physical world, but rather as tool for handling quantum systems by doing experiments or constructing quantum devices. Examples are the most common versions of the quantum logic approach.

My aim is to present an approach to the interpretation problem which do not belong to any of these categories, and to try to give an answer to how we can let quantum mechanics become a tool of understanding the world without associating it with some "physical interpretation" in the sense of visualisation and mental images.

This position corresponds closely to the one we can find in the writings of Paul A. M. Dirac. Dirac was one of the three physicists who got the Nobel Prize for the discovery of Quantum Mechanics. The other two were Erwin Schrödinger and Werner Heisenberg. Dirac contributed to the theory in many ways, he developed its mathematical formulation and he extended it to include the insights, and hence the mathematical structure, of the special theory of relativity through the so-called

Dirac equation. As early as in 1930 he published what became a famous textbook with the title *The Principles of Quantum Mechanics*. This book reflects Dirac's understanding of the theory and is characterised by putting mathematics in the forefront, and by *letting the mathematics of the theory speak for itself*. It was therefore of crucial importance for him that the mathematical notation was as clear and transparent as possible, in which he succeeded to an outstanding degree.

The book went through four editions, the last one published in 1958. Already in the first edition he used a mathematical language which was more general, but also more abstract, than the presentations which were common at that time. In the third edition, this was developed into a new, mathematical notation. Of the first edition, Einstein (1973) commented that *Principles* was "the most logical perfect presentation of quantum mechanics". In 1955, after the publication of the third edition, John von Neumann (1955) wrote that "Dirac, in several papers, as well as in his recently published book, has given a representation of quantum mechanics that is scarcely to be surpassed in brevity and elegance …" (p. viii). Furthermore, Eugene P. Wigner and Abdus Salam (Salam and Wigner 1972) wrote:

> Posterity will rate Dirac as one of the greatest physicists of all time. The present generation values him as one of its great teachers – teaching both through his lucid lectures as well as his book *Principles of Quantum Mechanics*. This exhibits a clarity and spirit similar to those of *Principia* written by a predecessor of his Lucasian Chair in Cambridge [i.e. Isaac Newton]".

Finally, Freeman Dyson wrote that "He [Dirac] presents quantum mechanics as a work of art, finished and polished" (Farmelo 2009, p. 428).

These praises of the book are philosophical significant, because they are all about the clarity of expression of Dirac's mathematical notation. Thus, they imply that the mathematics *expresses*, that the mathematical expressions really are expressions in a linguistic sense. Thus, they more than hint that mathematics is a language, a device for meaningful expression.

This, I think, is the reason why many students of quantum mechanics, myself included, have felt that the reading of *Principles* made them understand quantum mechanics better, although it did not attempt to explain any visualisable "physical content" behind the mathematics.

In the introduction to the first edition of *Principles*, Dirac (1930) discusses the question of how we can understand quantum mechanics. He explicitly rejects intuitionism, which was the traditional way of understanding classical physics:

> The classical tradition has been to consider the world to be an association of observable objects (particles, fluids, fields, & c.) moving about according to definite laws of force, so that one could form a mental picture in space and time for the whole scheme… It has been increasingly evident in recent time, however, that nature works according to a different plan. Her fundamental laws do not govern the world as it appears in our mental picture in any very direct way, but instead control a substratum of which we cannot form a mental picture without introducing irrelevancies.

In what follows, we will show how this statement goes into an extensive philosophical debate on the meaning of language in general. In this discussion, Edmund Husserl, founder of the philosophical school of phenomenology, is the spokesperson for intuitionism. We will study closely Jacques Derrida's general arguments against the intuitionism in general. What Dirac says we can learn from quantum mechanics, Derrida suggests that we can learn by analysing language itself. For Derrida, there is "nothing outside the text", as for Dirac, there is nothing (in quantum mechanics) outside the mathematical expressions and equations, which is the "text" of this theory. And, I would like to point out, still both Dirac and Derrida are realists. Both believe in a reality outside us a reality that language is *about*, it being the natural language or mathematics. The question is only how language represents this world.

But Dirac says more. If quantum mechanics cannot provide us with pictures of reality on the traditional sense of visual images, it can give us another kind of picture:

> One may extend the meaning of the world 'picture' to include any *way of looking at the fundamental laws which make their self-consistency obvious*. With this extension, one may gradually acquire a picture of quantum phenomena by becoming familiar with the laws of quantum theory (Dirac 1935, p. 10).

This "picture theory" of the mathematical language (remember that "fundamental laws" always refer to mathematical equations) is strikingly close to what the philosopher Ludwig Wittgenstein had been trying to formulate in *Tractatus Logico-Philosophicus*, published in 1922. Wittgenstein grew up in Vienna, studied engineering in Berlin, but lived his philosophical life in Cambridge, in the beginning with Bertrand Russell as his mentor. Also for Wittgenstein "there is nothing outside the text", all intentionality is representative. But the linguistic expressions, in addition to stating facts, are also in some abstract sense pictures of reality. Wittgenstein's picture theory is very similar to Dirac's idea of "a new kind of picture". This is probably due a common influence: Heinrich Hertz's classic from 1899, *The Principles of Mechanics Presented in a New Form* (Hertz 2003). After having discussed Derrida's arguments against intuitionism, which I consider as valid, we will therefore turn our attention to Wittgenstein's picture theory. This theory provides us with what we are searching for, an intentional meaning of language, including the language of mathematics. By intentional meaning I mean, how the language says something about physical reality.

So, what I will attempt to do, is to put these pieces together like a mosaic and thus construct what I suggest calling the "Dirac-Derrida-Wittgenstein interpretation" of quantum mechanics. An interpretation which does not seek images but meaning, not to go behind the mathematical language to see in a direct intuition what it means "physically", but to stay within the mathematics and ask for its meaning "from the inside". And this meaning corresponds to pictures of a "new kind". Thus, this interpretation sees the problem in a linguistic sense as a problem of semantics.

And, this is our starting point: to consider the mathematics of quantum mechanics as a language, and ask in which way this language makes meaningful statements, and how it can be read as a linguistic expression of how the world is.

Although I reject Husserl's intuitionism, I adapt Husserl's main philosophical method, phenomenology, to conceptualise the problem at hand. That Husserl's intuitionism can be separated from phenomenology is today generally accepted, and Derrida is included among modern phenomenologists.

Husserl's Intuitionism

What Derrida calls Husserl's *intuitionism* (a terminology which I will adapt) is most clearly expressed in Husserl's late work *The Crisis of European Sciences and Transcendental Phenomenology* (Husserl 1970) and its associate article *Origin of Geometry* (Derrida 1962). The main idea is that the mathematization of science and the associated abstraction has led to a loss of meaning content. In European science, this trend starts with Galilei, but already in Greek geometry we see a similar development. The case of geometry is described in *Origin*. Derrida translated *Origin* into French, and equipped it with a critical introduction three times as long as Husserl's text. In this introduction, Derrida developed his own philosophy in dialogue with Husserl.

Husserls analysis of the origin and development of geometry is not based on historical research, he rather used geometry as an example of how the development of a mathematical theory based on abstract concept from its origin in the practices of humans must have been. Thus, it is mainly an analysis of what a mathematical theory essentially is.

In *Crisis* Husserl contrasts the "immediate experiencing intuitions" and the "experiential knowledge" of the "prescientific life-world" with the abstract scientific knowledge consisting of "correlation between mathematical idealities", as we find it in geometry. This contrast depends on the idea of immediate intuition [*Anschauung*] and the corresponding immediate presence of the object intuited. Geometry derives from this kind of life-world experience through abstraction, which leads to the establishment of ideal objects. An example of such an ideal object is the concept of a straight line, which is abstracted from experiences like that of the edge of a rectangular table. By going from the visual and possibly tactile experience of such an edge to the ideal object of a straight line, something is lost, according to Husserl. The table edge has a *meaning*, because it is a part of our life-world and the object of experience, as something which can stand before us in its immediate presence.

The ideal straight line, however, has lost most of this meaning, but is gaining in something else: it is an eternal or timeless object, and what we can say about it will be eternal truths. Now, even a straight line has some of its meaning retained, since the ideal figures of geometry still are similar to some features of real objects, although it is stripped for the depth of meaning provided by being part of our life-world. But modern geometry is subject to a further development, to an increased mathematization leading to further loss of visual meaning, and this is what Husserl calls the arithmetization of geometry “liberated from all intuited actuality” (Husserl 1970, p. 44). And here we come to the crucial point:

> The arithmetization of geometry leads almost automatically, in a certain way, to the emptying of its meaning. … one lets the geometric signification recede into the background as a matter of course, indeed drops it all together … One thinks, one invents, one makes great discoveries – but they have acquired, unnoticed, a displaced, “symbolic” meaning” (Husserl 1970, p. 44-45). We end up with a “science … which can be constructed in pure thought and in empty, formal generality (Husserl 1970, p. 45).

The correlations between ideal entities are expressed through mathematical formulae, and Husserl contrasts what he calls its formulae-meaning to the true being obtained from the presence in the intuition of a real object.

The apparent eternal content of abstract geometry, as Husserl describes it in *Origin of Geometry*, is not really about eternity, but about repetitivity; the truths of this abstract science does not recide in a permanent world of ideas, as Plato thought, but can be repeated in human minds, they can be thought and written again and again at any time. And (and this fact will turn out to be of crucial importance to Derrida), this repetitivity is secured by *writing* and *written symbolism*, the mathematical notation.

Now, if we trace the development of geometry back to its original application to table edges and similar tings, we rediscover a meaning which has been hidden behind the mathematics like geological sediments. This is what Husserl calls the “sedimentation of meaning” in the development of science. His thesis is that science in general, and physics in particular, has been subject to such a development, leading to a sedimentation and eventually to a loss of meaning, and hence a “crisis”.

Derrida: There Is Nothing Outside the Text

Husserl’s intuitionism is the starting point for Derrida, who wrote a 130-page’s introduction to *Origin*, a 24-page’s article, and later developed the subject further in his own book *Voice and Phenomena* (Derrida 1973). A main argument for Derrida is that not only the mathematics of science, but all language is about idealization, since it is about repeatability. This idealization implied by the use of language also concerns the description of an immediate experience, which, in Derrida’s analysis turns out to be less simple than Husserl thought. According to him, all experience is constituted by language, and hence idealization, and there is no such thing as pure immediate experience or immediate presence. Thus, there is no pre-given life-world, since the world, as it appears to us, is already interpreted by our language and our theories about it, including geometry and other sciences.

Consequently, we can go back to mathematical science and claim that what mathematics does in physics is what language does in every-day life. And this is just the starting point for Derrida: the idealities of geometry are analogous to the ideality of any concept. In geometry, such an ideality is established by written mathematical symbolism and equations, in ordinary language, by written letters, words, and sentences. *The meaning of language is constituted through the constitution of linguistic signs*. This constitution give rise to both ideal objects, *and* meaning, since the meaning of a linguistic sign is, in itself, an ideal object.

In his introduction, Derrida notes Husserl's understanding of the mathematical notation and hence the writing as a precondition for establishing the ideal objects of geometry. The ideal objects are established as permanent structures through the written symbols of the mathematical notation. Thus, idealities are constituted as meaning of written signs. But to Derrida the difference between speech and written language is inessential, they both consist of physical conventional signs. Thus, also for speech it is the case that "speech is no longer simply the expression (*Aüsserung*) of what, without it, would *already* be an object: caught again in its primordial purity, speech *constitutes* the object" (Derrida 1989 p. 77). The word-sign, and we talk now of the spoken as well as the written word, is not just a means for expressing an already existing meaning, an ideality, but the establishment of a word implies the constitution of such an ideality. And word-meanings are already such idealities. Thus, with reference to Husserl, his idea of an intuitive meaning without ideality would mean meaning prior to language, even prior to every-day language, as if the experienced object can stand there before us cleansed of linguistic meaning, equipped with some more original pre-linguistic meaning content. This is a position which is, in my view, as well as in Derrida's, impossible to defend.

Thus, Derrida does not accept Husserl's premise that the idealizing acts leading to geometrical truths draws is its meaning from a pre-existing structure of purely experiential meaning appearing in an immediate and pre-linguistic presence. This idea is it that Derrida spots as the weak point in Husserl's phenomenological enterprise in general. Against it, Derrida holds that the idealization itself implies a leap which cannot be given a causal explanation. This leap is the meaning-giving act itself. This leap is what we do when we see a tree and recognize it as a tree, and the ideal meaning of a "tree" permeates our whole experience of it. We can have hunch that there could be another way of experiencing it if we did not have the concept, so that in any experience we experience something which is in a way different from the impossible, un-experientable pure object. So, our way of experiencing the world always introduces such a difference—which, however, is not a difference in an ordinary sense, which is always a difference between something given meaning by our linguistic constitution. Thus, we are in the need of neologisms, and Derrida provides them. According to him, this leap is similar but not identical with something characterized by words like "difference", "shift", or "delay", giving rise to the neologism "différance"—This idea of a leap makes impossible the idea of ideal entities as representing a pre-given simple "living" presence in intuition:

> The impossibility of resting in the simple retention of a living present, the sole and absolute origin of the actual and the regular, to being and sense, but always other in its self-identity; the inability to live enclosed in the innocent undividedness of the primordial Absolute, because it is not present except by being deferred-delayed [*différant*] without respite, this

> powerlessness and this impossibility are given in an original and pure consciousness of the *Différance*. (Derrida, 1989, p. 153, the translation slightly altered by the present author).

This means, that there is no such thing as the immediate, unmediated intuition in the sense of Husserl, only an intuition which is constituted through the *différance* implied by constitution of (ideal) meaning. Furthermore, there is no given experience which can function as an absolute foundation, an Archimedean point, for our understanding, as the empiricists have held. We are woven into the ideality of language already in having the experience, which, in itself, is invaded by language. Husserl's idea of a pure perceptive intuitive act, fundamentally different from the signitive act and thus from the intervention of language, is an illusion. The perceptive act is permeated by language. Signification generally implies idealization, this is not something limited to geometry or mathematical physics, and thus it interferes with experience in a way that Derrida tries to express through the word *différance*, which for him is not a concept in ordinary sense with a specific meaning, but something characterizing all constitution of meaning by idealization.

The main point here, having a bearing on modern physics, is that the idealization given by mathematical (and all linguistic) concepts *is also a constitution of meaning*. In the case of geometry, it means that the new idealized concepts, like that of a straight line, provides *new meaning*, which can make the description of spatial objects meaningful in a new way. This is the key to understanding mathematical physics.

For Derrida, his introduction to *Origin of Geometry* is a beginning, in which the course is plotted. In *Voice and Phenomenon* (Derrida 1973) he develops his ideas further. From the special case of mathematical idealization, he asks the question of the sign in general. And he does that by going back to Husserl's first main work in phenomenology, *Logical Investigations* (Husserl 2001).

Derrida starts here by considering Husserl's intention of avoiding so-called metaphysical thinking. For Husserl, instead of metaphysical speculation, he wants to base philosophy as well as science on the immediate experience or intuition, in which the outside reality shows itself as itself by being immediately present to consciousness. This is expressed in Husserl's slogan: back to the thing itself! This is for Husserl his "principle of principles", and Derrida asks rhetorically:

> Does not the phenomenological necessity, the strictness and subtlety of Husserlian analysis … nevertheless hide a metaphysical presupposition? Does it not hide a dogmatic or speculative attachment, which … wants to *constitute* phenomenology from the inside, in its critical project and in the value which institutes its own preconditions: in just this which itself soon will acknowledge as the source and guarantee of a value, the "principle of principles", namely the originally given evidence, the present, or meaning as *presence* to the fullness and originarity of an intuition (Derrida 1973, p. 3).

Derrida makes us aware of that already the experience of an immediate presence depends on ideality and "the ideal object". Only as an ideal entity can something be something that becomes repeatable and hence recognizable and thinkable, and it *is only as an ideal entity it can be an object to consciousness*: "This ideality is even the form where the presence of an object altogether can be repeated as the same"

(Derrida 1973, p. 8). The timelessness of the ideal entities is not founded on a being outside time, but on the possibility of a being *repeated* in time, as Husserl in fact already points out in *Origin of Geometry*, as mentioned earlier. *In the absence of ideality and hence of repeatability, the presence is lost as well.* If we imagine a pre-expressive presence, nothing can be recognized or identified, nothing can be seen *as something*, an object can never be *the same* through the passing of time. A presence without ideality cannot be repeated, it can also not be described, or remembered, or imagined, or even thought of: "The ideality is the salvation or the mastery of the presence in the repetition." (Derrida 1973, p. 8).

The presence is secured through ideality, but ideality itself is constituted by writing (including speech or any physical signification). This applies to words in language as well as mathematical symbols in geometry and physics. This is the meaning of the statement "there is nothing outside the text", where text in most applications is spoken or written in ordinary language, but in physics written in the language of mathematics.

This means that it is not simple to talk plainly about experiences as such because when we talk, the speech itself *constitutes*, it is not neutral and purely expressive. According to Derrida, it is not possible to answer directly the question of the essence of the sign, because the question itself assumes the possibility of placing oneself outside any sign system.

However, neither is the sign itself a new experiential foundation for a linguistic metaphysics; also the sign is ideal; it is not a unique concrete event:

> A sign is never an event, if an event means something irreplaceable and irreversible empirically unique. A sign which happens just "once" would not be a sign. A purely idiomatic sign would not be a sign. A signifier must be recognizable in its form, in spite of and across differences in empirical features which may modify it. It must always stay the same and be available for repetition as such in spite of and across those changes which necessary follows from what we call an empirical event (Derrida 1973, p. 55).

Even if speech or writing is subject to the variation of the concrete world, they are only speech and writing to the extent that they are repeatable and can be recognized, i.e. as ideal entities:

> A phoneme or a grapheme is necessarily to a certain extent always different, each time it appears in an act or a perception, but it can only act as a sign and a language in general if a formal identity allows its repetition and recognition. This identity is necessarily ideal. It implies therefore necessarily a representation: as *Vorstellung*, the general place for ideality, as *Vergegenwärtung*, the general possibility of the reproductive repetition, as *Repräsentation*, to the extent as every meaningful event is a substitute for (for the ideal form of) the signified as well as the signifier. *Since this representative structure is the meaning itself*, I cannot involve myself in actual speech without being enrolled in an unlimited representativity [my italics] (Derrida 1973, p. 55-56).

Where there is meaning, we already have ideality, and the ideality is constituted by being represented, by the sign. But the sign is not a sign without referring to a meaning. Thus, we do not have meaning first, and then sign, or sign first and then meaning, but both things at the same time. This gives meaning to Derridas concept of trace and archi-scripture:

> The trace is not an attribute about which one might say that the living presence itself is originary. One must think originarity from the trace and not the other way around. *This archi-scripture is the origin of meaning* (Derrida 1973, p. 95).

In his elaboration of this in his later work *On Grammatology,* Derrida expresses this by the statement "there is nothing outside the text", or, in French, the untranslatable "*il n'y a pas de horse-texte*". There is nothing we can experience, remember, think about or talk about which are not ideal objects which are meanings of signs and hence text.

In the mathematics of a physical theory the *texte* is the mathematical expressions and equations. It does not mean that this mathematics, like any language in use, does not deal with the external reality, but *it deals with it as "text"*; it is understood by the mathematical description *from within*. Now, a new language does not only open for us new concepts, but a new logic and structure of reasoning, of description and understanding. Such is the case also for the mathematics of quantum mechanics. As we search for the meaning of this strange language, we have again to search for the meaning which is constituted *inside* this *texte*, not to expect the meaning to be given from the outside, by some kind of pure pre-linguistic original intuition.

Wittgenstein and the Picture Theory

Given Derrida's statement that meaning is something constituted within language, including the mathematical language of geometry, and hence also the mathematics of physics, one question becomes urgent: How can this linguistic structure represent the external world? After all, language is about something, language a way of directing our awareness, not towards language itself, but towards the world outside. This is an implicit assumption in Derrida, and it also develops to be an important concern of his, but in this article I want to turn our attention towards another philosopher, Wittgenstein, to look for the clarification needed.

As we have observed, Dirac expressed this concern by what he calls an extended meaning of a picture.

A picture theory of this kind, we can find in the thinking of the early Wittgenstein of *Tractatus*. Thus, Wittgenstein can help us to understand *in which way* this mathematical language can have meaning with respect to physical reality. This should not be unexpected, considered that *Tractatus* was inspired by the philosopher-physicists Heinrich Hertz and Ludwig Boltzmann. Wittgenstein often referred to Hertz, and he also planned to study under Boltzmann, but was prevented from doing so by Boltzmann's sudden suicide. The problems facing those who want to understand modern physics are, however, goes much deeper than those of the classical physics of Hertz and Boltzmann, and we will need a more general and abstract approach, which is exactly what Wittgenstein tries to formulate. In *Wittgenstein's Vienna*, Janik and Toulmin (1973) sees Wittgenstein's early philosophy as developing out of the Viennese intellectual's critique of language in the Kantian spirit, combined with the approach to physics of Hertz and Boltzmann and the writings on logic by Gottlob Frege and Bertrand Russell. This lead to a general philosophy of language and logic which found its expression in the terse statements of *Tractatus*. Wittgenstein's early philosophy has been generally overlooked by both physicists and philosophers of physics. However, we will see that some of the philosophical insights found in *Tractatus* may be just what we need for dealing with the strange epistemological situation created by the physics of the 20th Century.

After the completion of *Tractatus* at the end of the First World War, Wittgenstein was retreating from philosophy for some years, spending most his time as an elementary schoolteacher in the countryside of Austria. In this period *Tractatus* was published. Wittgenstein returned to philosophy and to Cambridge in 1929, first as a fellow and then as a professor from 1939. Thus, he was at Cambridge simultaneously with Dirac, but there are no signs of any intellectual contact. At that time Wittgenstein had both changed his field of interest and some of his philosophical views since *Tractatus*. So, I do not suggest any influence from one to the other, only a similarity of ways of thinking. Some of this similarity may be traced back to a common influence from Hertz and Boltzmann, who they both may have read.

In the introduction to *The Principles of Mechanics Presented in a New Form* from 1891 (Hertz 2003), Hertz considers mathematical models as they are applied in physics. He refers to them as pictures or images (*Bilder*). He lists four criteria to be satisfied for a model to be such a picture. The first two criteria are that it has to be logically consistent and that it must be empirically verified. The third criterion is subtler, and it concerns the inner quality of the model itself. Among given alternatives one should chose the one which is most distinct and clear (*deutlich*), which Hertz saw as an indication that the model is the most appropriate (*zweckmässig*), including into it the more essential relation to the object of which the model is a picture. The fourth criterion is that of simplicity, the simplest model should be preferred if one still has a choice. Clarity and simplicity is often associated with beauty in physics, and later we will see that this becomes an important notion in Dirac. Note that even Hertz, who is still considering classical physics, in which it is still possible to visualise the objects and processes modelled, does not look at the models as plain pictures. Rather, they are—in a Kantian spirit—models produced by our minds and necessarily affected by the way we construct our models. In fact, Hertz even compares the structure of the picture to a grammar, consider his own presentation of mechanics as a systematic grammar, in contrast to e.g. a grammar devised for the purpose of making the language easy to learn.

Following Janik and Toulmin, we can consider *Tractatus* as an attempt to solve the general problem of language by generalising the model theory Hertz and Boltzmann into language in general. *Tractatus* itself is a mosaic, put together by elements to construct a philosophical system. It also implies a paradox, since it through its own logic leads to the conclusion that such a system is impossible. Thus, Wittgenstein calls his own statements nonsensical. However, he compares the book to a ladder which should be thrown away after one has ascended it. Thus, this ladder consisting of nonsense obviously will help us to ascend to a higher level of understanding. This idea makes the whole work enigmatic. In addition, some of the notions of the book have gradually become rejected as untenable, e.g. the idea of elementary sentences fulfilling criteria which makes them impossible to construct. This is the core idea of the atomic theory of logic (all sentences are logical combinations of "atomic" or elementary sentences) which Wittgenstein got from Russell.

Here, we will focus mainly on one piece of the mosaic of *Tractatus*, the picture theory. Wittgenstein's picture theory of language is an abstract generalisation of Hertz's picture theory of mathematical models. And we need an abstract generalisation to deal with non-classical physics.

The picture theory seems to enter Wittgensteins thinking independent of many of the other elements of *Tractatus*. We can see this in his *Notebooks 1914–1916* (Wittgenstein 1961), where the idea gradually develops. The main idea is that sentences (not the mathematical equations of Hertz) are pictures of a kind of reality:

> That a sentence is a logical portrayal (*Abbild*) of its meaning is obvious to the uncaptive eye. (p. 5).

Note that Wittgenstein talks of a *logical* picture or portrayal, not a visual one. To illustrate this statement of Wittgenstein's, it is easy to choose a visualisable example, like "The cup is on the table", but this example is also misleading, for it conceals the important point that the kind of picture a sentence is, is logical, not visual. This is clearly pointed out later in the text:

> It can be said that we are not certain of being able to turn all situations into pictures (*Bildern*) on paper, we are still certain that we can portray all logical properties of state-of-affairs (*Sachverhalte*) in a two-dimensional writing (*Schrift*) (p. 7).

Thus, Wittgenstein proceeds:

> We can say straight away: Instead of: this proposition (Satz) has such and such sense (Sinn): this proposition represents (*stellt dar*) such an such state-of-affair.… Only in this way can the proposition be true or false: It can only agree or disagree with reality by being a picture (Bild) of a state-of-affair. … *The proposition only says something in so fare that it is a picture*. (p. 8). … Logic is only interested in reality. And thus in sentences only in so far that they are pictures of reality. (s.9)

Then he introduces the important difference between what can be said and what can be shown: "The agreement between to complexes is obviously *internal* and for that reason cannot be expressed but can only be shewn" (p. 9). This is elaborated in *Tractatus*.

Applied to quantum mechanics, we can say that what the mathematical expression *says* is their instrumental meaning, but what it *shows* is the logical structure of the quantum world, i.e. the world as described by quantum mechanics. Thus, we cannot *say* how the quantum world is, we can only show it, and we show it by setting up the right mathematical expressions. Because we cannot say how the quantum world is, we must be silent about it.

The *Notebooks* are where Wittgenstein develops the system put together in *Tractatus*. *Tractatus* consists of short statements, numbered after sophisticated system which strictly defines the logical orders of the statements. I will refer to the statements by their number in this system.

Inspired by Hertz' model for the presentation of physics, Wittgenstein considered the sentences of language to be models or pictures of *state-of-affairs*. Hertz' concept of a mathematical model is already an abstract notion, which does not concern the visualising intuition of a physical system, but rather something that provides a means to perceive the logical form or structure of the theory by observing the logical form of the model. Thus, as we have seen, the picture theory is generally *not* about visual pictures or visualisation, except in a few cases. Wittgenstein made a new step in abstraction when he generalised this notion to language in general. The sentence is a kind of picture, but this picture cannot be directly compared with the state-of-affairs of which it is the picture, since the state-of-affairs is only available to us *through* the picture. But this is nothing but a reformulation of the statement that there is nothing outside the text. So, we have to have two things in our mind at the same time. The state-of-affairs is something real, something out there in the real world. But it can only be a state-of-affair by corresponding to the linguistic model. In other words. If you change language, you have other state-of-affairs. Never the less, the state-of-affairs is out there in reality, it is not a linguistic entity. This is the essence of the *intentionality* of language. Language does not exist only in itself, closed into itself, *qua language*, but always in an intentionality relation to the world, which it is about. *Only through the linguistic expression, i.e. only through the picture we can grasp and express the state-of-affairs*. To get a clear understanding of some fact, it is necessary to have a clear picture, i.e. a clear expression in language, as already Hertz pointed out in his limited context.

Wittgenstein tried to specify the limits of language "from within" by specifying what can be *said*, and only by implication, what cannot be said. He was therefore able to perform a critique of language, yet save it for use whenever appropriate. It is appropriate when we deal with state-of-affairs; therefore, language is adequate for "science", when we take "science" in the widest possible sense to mean any field of knowledge which is concerned with facts. Thus, Wittgenstein operates with such a wide concept of science that it includes many subjects of study which other philosophers than Wittgenstein would call philosophy.

In classical mechanics, the mathematics is assumed to be translatable, not only into ordinary language, but into a system of images, where we imagine stones that falls, planets orbiting the sun, water flowing in a pipe. This imaginative translatability is in fact an extraordinary situation, different from language in general, and significantly different from the situation in quantum mechanics. Therefore, quantum mechanics needs the general language problem to be solved before it can be properly understood. It needs the *tractarian* notion of a picture which is the only means to grasp some facts and the logical structure of their relations. The fact itself is said, and the logical form or structure is displayed by the form of the sentence.

If we turn to Dirac and *Principles* (Dirac 1930), he interprets quantum mechanics in a similar manner by giving clear expression of what can be *said* within his theory, and the clearest possible display of its logical structure. Thus, he exhibits its logical structure, which cannot be said, only shown, and even what can be said is exclusively expressed in mathematical language. Consequently, the mathematical symbols should be treated as analogies to ordinary word-signs, and the mathematical equations to sentences. In the same way that a sentence is a picture (in Wittgenstein's notion) of a state-of-affairs, so is the mathematical equation. Furthermore, in the same way that we are left with the linguistic picture to grasp the content or meaning of a state-of-affairs, we are also left with the mathematical symbols and equations of quantum mechanics.

Wittgenstein also rids the physicist of the apparently unanswerable question of meaning of each single symbol. Applied to quantum mechanics: instead of asking the meaning of "position" in a theory where one no longer can imagine a particle as something confined to a single place, we must accept that "Only propositions have sense; only in the nexus of a proposition does a name have meaning" (*Tractatus* 3).

In quantum mechanics, the nexus have changed. The new meaning of the old word in quantum mechanics is the meaning it acquires through its logical relationship to other symbols (names) in the logical (mathematical) form or structure of the theory itself.

Nonetheless, quantum mechanics is not only a propositional structure, but is supposed to be true about the world, even empirically verifiable. Does this coupling to the external world prevent us from thinking from the "inside"? It is true that traditional experimental equipment has, historically speaking, often been described by classical mechanics to such an extent that Niels Bohr thought that this was a necessity. However, it is not. The simplest and perhaps most widespread kind of quantum measurement is spectroscopy. This has traditionally been thought of as an interaction between a quantum object (such as an atom) and a classical electromagnetic field. However, although inconvenient, the electromagnetic field can be described within an extended quantum theory, and the interaction as well as the outcome can be described within the quantum language. Such a measurement serves as an empirical confirmation of both quantum mechanics and quantum electrodynamics, representing the "end points" of Wittgenstein's measurement gauge:

> Only the end points of the graduating lines actually touch the object that is to be measured (*Tractatus* 2.15121).

Today, there is great activity in developing models for experimental measurement within the quantum theory. Models are even being developed for understanding why classical mechanics can be derived from within quantum theory (the so-called "decoherence phenomenon"). This how we see it today: quantum theory is the universal description of nature, while classical mechanics is a specific case found within this theory.

Finally, what about people—such as philosophers—who do not have a sufficient mathematical education to understand quantum physics "from within"? Also here, Wittgenstein clarifies the situation. In order to deal with this problem, we need to put Wittgenstein's notion of "nonsense" into use. Outside the proper mathematical language, we are left to talk nonsense; however, nonsense is, in the philosophy of Wittgenstein, far from meaningless. E.g. "the particle is in many places simultaneously" or "have many velocities simultaneously" are useful and informative nonsensical statements. I find Wittgenstein's notion of nonsense very illuminating in cases like this. Such statements are both correct and slightly incorrect at the same time, and are used with a slight uneasiness by physicists. Nevertheless, they remain the best way of expressing the strangeness of the quantum world. Informed nonsense brings linguistically inaccessible truths about nature back to "the man in the street", including the philosophers.

The first scientific revolution was caused by the discoveries of Copernicus, Kepler, Galileo, and Newton. The new world view became universally known and accepted, and has since become a part of both our cultural heritage and what we now consider to be common knowledge.

The next revolution in physics took place in the twentieth century and consisted of three main steps. The first and second steps were the special and general theories of relativity. The third (and even more revolutionary) step was quantum physics, starting with quantum mechanics in 1925. Except for a short period of newspaper headlines in 1919–20 making Einstein the most famous scientist ever, one can safely say that the man in the street never noticed that any change had taken place. The reason was not that these new developments were less revolutionary than the Copernican Revolution. The main reason is the fact that the new theories are inaccessible to people without a solid background in mathematics. Moreover, experts—including the creators of the theory—have been discussing throughout an entire century what it truly says, without reaching any conclusion upon which all parties can agree. Nonetheless, the theory's mathematical structure has been established beyond discussion as being consistent, highly developed, and, according to physicists, beautiful.

Should philosophers care about these questions? Yes, for at least two good reasons. One compelling reason is that new physics challenges some of our most obvious and fundamental assumptions about the material world. For instance, we take it as obvious that what is present at this moment of time exists, while neither past events nor future ones exist. Likewise, it is obvious that when a physical object moves in space, at each instant it has one and only one position there, and only one velocity. Theories that challenge these assumptions are certainly of philosophical interest. The second reason for philosophers to be interested in the new physics is that it may be the case that philosophy can contribute to the interpretation of the theories, that proper philosophical theory may be needed as the key to the correct understanding of modern physics. There are good reasons to suspect that physicists trying to develop a correct interpretation have in their thinking built-in philosophical assumptions of which they remain unaware and which may prevent them from arriving at the right answers.

This paper focuses on quantum mechanics as a philosophical case, although a similar reasoning may be completed concerning the theories of relativity. My hope is that other philosophers than myself will find this to be an interesting and challenging case, allowing them to catch a glimpse into a surprising, strange and beautiful part of the world into which, according to Heidegger, we are thrown.

Quantum mechanics is perhaps the most consistent application of Galileo's thesis that the book of nature is written in the language of mathematics. Quantum mechanics is exclusively written in mathematics, and is in fact not translatable into any ordinary language. We therefore have a very clear-cut situation for examination of mathematics used as a language.

In quantum mechanics, some of the properties of a physical object have names which are known from classical mechanics. A quantum particle has properties like position, velocity and energy. However, in quantum mechanics, an object does not have these properties in a straightforward sense. For instance, they do not always have specific numerical values; they may instead be associated with mathematical distributions, indicating that the quantity in a sense has many values simultaneously. Thus, a particle may be in a state where it has several positions, it is at several places, at one time, or it may have many velocities simultaneously (even velocities pointing in opposite directions). Such a particle is impossible to imagine, and the existence of such strange objects is not easy to accept. Nevertheless, things like this are roughly what quantum mechanics says about the physical world.

To deal with the apparent abstractness of quantum mechanics, physicists generally have felt compelled to add an interpretation to the mathematical structure or formalism in which all this is expressed. One takes our inherited conception of matter and intuition of material objects for granted, and tries, in some way or another, to explain the strange mathematical symbolic relations in terms of such concepts and intuitions. I call these attempts to make an interpretation from the outside.

Turning to Dirac, one striking feature of Dirac's approach to quantum mechanics is that it is not an interpretation made "from the outside". In the spirit of *Tractatus*, Dirac is silent about whereof we cannot speak. Thus, he does not talk about interpretation, but instead he lets the theory express itself "from the inside", and in the clearest possible way, even developing his own mathematical notation that is notable for its brevity, elegance, and transparency. In this way, he let the logical structure of the theory be displayed, lets it "show itself" as clearly as possible, and then adds nothing about it.

References

K. Camilleri, Constructing the myth of the copenhagen interpretation. Perspect. Sci. **17**(1), 26–57 (2009)

J. Derrida, *Edmund Husserl's Origin of Geometry. An Introduction*. (Including Husserl, E: *Origin of Geometry*) (University of Nebraska Press, 1989). French original: *Edmund Husserl's*

L'origine de la géométrie, traduction et introduction par Jacques Derrida. Epiméthée, Essais Philosophiques, Collection fondée par Jean Hyppolite (Presses Universitaires de France, 1962). German original of Husserl's *Origin of Geometry*: Husserl, E.: *Der Ursprung der Geometrie als intentional-historisches Problem* (*Ursprung*). Revue internationale de philosophie **1**, 2, 1939. Extended version: *Die Frage nach dem Ursprung der Geometrie als intentional-historisches Problem*. I: *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie*. Husserliana VI. Den Haag, 1962

J. Derrida, *Edmund Husserl's Origin of Geometry An introduction* (University of Nebraska Press, London, 1962)

J. Derrida, *Voice and Phenomena*, (Nortwestern University Press, Evanston, 1973). French original: *La voix et la phénomène* (Presses Universitaires de France, 1967)

P.A.M. Dirac, *The Principles of Quantum Mechanics*, 1st edn. (Oxford University Press, Oxford, 1930)

P.A.M. Dirac, *The Principles of Quantum Mechanics,* 2nd edn. (Oxford University Press, Oxford, 1935)

A. Einstein, I *James Clerk Maxwell: A Commemorative Volume 1831–1931* (Cambridge University press, Cambridge, 1973)

G. Farmelo, *The Strangest Man* (Faber & Faber, London, 2009)

H. Hertz, *The Principles of Mechanics Presented in a New Form* (Dover, New York, 2003). German original: *Die Prinzipien der Mechanik in neuem Zusammenhange dargestellt*. Drei Beiträge 1891–1894. Verlag Harry Deutsch

E. Husserl, *Logical Investigations* I-II (Routledge, London, 2001). German original: *Logische Untersuchungen* I-II, 1900–1901

E. Husserl, *The Crisis of European Sciences and Transcendental Phenomenology* (Northwestern University Press, Evanston, 1970). German original: *Die Krisis der europäischen Wissenschaften und die tranzendentale Phänomenologie*, Haag: Nijhoff 1954

M. Jammer, *The Philosophy of Quantum Mechanics* (Wiley, New York, 1974)

A. Janik, S. Toulmin, Wittgenstein's Vienna (Simon and Schuster, New York, 1973)

J. Pykacz, *A Brief Survey of Main Interpretations of Quantum Mechanics*. Chapter 2 in: *Quantum Physics, Fuzzy Sets and Logic* (Springer, 2015)

A. Salam, E.P. Wigner (ed.), *Aspects of Quantum Theory* (Cambridge University Press, Cambridge, 1972)

J. von Neumann, *The Mathematical Foundations of Quantum Mechanics* (Princeton University Press, Princeton, 1955). German original: *Mathematische Grundlagen der Quantenmechanik* (Dover, New York, 1943)

L. Wittgenstein, *Notebooks 1914–1916*. (English and German edition) (Blackwell, Oxford, 1961)

L. Wittgenstein, *Tractatus Logico-Philosophicus*. First German edition: *Logisch-Philosophische Abhandlung*. Wilhelm Ostwald (red..), *Annalen der Naturphilosophie*, 14 (1921). First English translation (an edition with both the Germaan and the translated text) by C.K. Ogden og F. P. Ramsey (Kegan, Paul, Trench, Trubner & Co, London, 1922).

Chapter 17
Mapping Quantum Reality: What to Do When the Territory Does Not Make Sense?

J. Acacio de Barros and Gary Oas

Introduction

To many physicists, the goal of science is to explain nature. To do so, as Galileo pointed out about 500 years ago, one needs first to describe the natural world: it is counterproductive, he argued, to ask *why* things happen the way they do; instead, one should ask *how* events occur. Thus, according to Galileo, the first task of a physicist is to *map* nature. Here we give the word map a more general meaning, referring to the construction representing any type of relationship between observable physical events, and not referring necessarily only to spatial relationships.

In regular maps the representation is straightforward and (hopefully) one-to-one. The same is not that clear when we are thinking about processes or time evolution. This is because some of the observable physical elements we want to represent are not all simultaneously present, meaning that we can only see parts of the territory at a time, and how we piece these parts together is not a trivial matter. For example, in physics we cannot design an experiment that tests the frequency of a photon *and* its position simultaneously. We can do one after the other, but not at the same time. Analogously, when we ask a person two questions, we cannot simply utter them simultaneously: this would be unintelligible. Instead, we need to choose an order for the questions and then ask them.

So, given that the mapping of processes is not as straightforward as the mapping of a territory, we need to set some ground rules for it. The first thing we note is that one of the purposes of a map, a geographical one that is, is to give a representation of the real territory that allow us to glance at it and thus derive some understanding of the relationships between the whole and its parts. The same must be true of our mapping of processes: it needs to provide relationships that make sense. Making sense, of course, means to frame such relationships in a way that does not violate the

J. Acacio de Barros (✉)
School of Humanities and Liberal Studies, San Francisco State University,
San Francisco, CA, USA
e-mail: barros@sfsu.edu

G. Oas (✉)
Stanford Pre-Collegiate Studies, Stanford University, Stanford, CA, USA
e-mail: oas@stanford.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_17

classical rules of inference, codified in classical Aristotelian logic. So, the idea goes, we impose a logical structure to our representations, and such representations allows us to construct from them a narrative that is consistent with the observed phenomena.

Alas, nature does not seem to like the above plan. Attempts to create consistent detailed representations of certain (quantum) phenomena in a way that allows us to map all properties consistently are problematic to say the least. This is the core of what we want to present in this paper: the idea that mappings that give us a consistent narrative for certain empirical situations are not always possible. We will do so in the following way. First, in section "Mapping Dynamics: Random Variables" we describe the constraints that using classical logic imposes on our constructions of representations of physical systems. Then, in section "When the Map Fails: Contextuality" we discuss an important situation in quantum physics where properties of certain systems behave contextually, and defy a classical-logic-based representation of them. Finally, in section "Alternative Mapping: Using Negative Probabilities" we introduce modifications to the classical descriptions, and show how they can be used to map systems that are not describable via the classical tools.

Mapping Dynamics: Random Variables

How do we represent the outcomes of experiments that may change with time? There are many different was to do this. Say, for example, we want to represent changes of temperature in a given physical system. We can think of a variable denoting the temperature $\mathscr{T}$ (taken by a thermometer) that we measure at equal time intervals Δt. With this we can create a function $\mathscr{T}(t)$ from the set $\{t_0, t_1, t_2, \dots\}$ of all possible times into the set of possible answers, e.g. $\{0.0, 0.5, 1.0, 1.5, 2.0, \dots, 120.0\}$ for a thermometer that measures temperatures between 0 and 120 °C. If I started measuring the temperature outside my house today at 1 AM in intervals of 1 hour, my function would be something like this: $\mathscr{T}(1) = 10.5$, $\mathscr{T}(2) = 10.0$, $\mathscr{T}(3) = 9.5$, $\mathscr{T}(4) = 9.5$, $\mathscr{T}(5) = 9.5$, $\mathscr{T}(6) = 11.5$, and so on. So, if we want to map change with time, a reasonable mathematical tool is a function.

However, functions do not necessarily take into account one important feature of natural phenomena: their uncertainty. What we mean by uncertainty is the following. Imagine we have a device that throws a six-sided die every one minute. These outcomes could indeed be represented by a function $D : T \to O$, where the actual outcomes of the die could be recorded, using $T = \{0, 1, 2, 3, \dots\}$ as the set of times and $O = \{1, 2, 3, 4, 5, 6\}$ as the set of possible outcomes of the die. However, because the outcomes of a die show a great deal of randomness, this description seems inadequate, as it does not capture the randomness of the phenomenon.

Where does this unpredictability come from? Consider the simpler case of tossing a fair coin (with two outcomes) instead of a die (which has six outcomes). When we toss the coin, regardless of how carefully we control its launch, if it is thrown with enough speed and rotation, its outcomes are unpredictable. This unpredictability comes from small uncontrollable variations on the coin's initial conditions that lead to large variations in outcomes (Ford 1983; Keller 1986; Vulović and Prange 1986). Of course, the underlying dynamics for this system is assumed to be the classical Newtonian mechanics, which is deterministic. However, because of uncertainties in the initial conditions, it has, for all practical purposes (FAPP), completely unpredictable dynamics. Something similar also happens to the die.

Another example of unpredictable outcomes generated by deterministic dynamics is the three-body problem. This problem was extensively studied by many researchers since Newton, including Alekseev (1968, 1969, 1986). The three-body problem is notoriously unsolvable, but Alekseev provided a case with enough symmetries that allowed him to prove interesting theorems. Take the situation where we have two bodies of equal masses, M, orbiting around the barycenter; this is a two body problem, and its solution is quite simple, with the motion of the planets defining a plane P. Now, let us take a small body of mass $m \ll M$, where m is so small as to not affect the motion of the two orbiting bodies M, and let us place m on an imaginary line passing through the barycenter and perpendicular to the plane defined by the orbits of M's. Because of the system's symmetry, the mass m will oscillate around the barycenter, and this oscillation is, unsurprisingly, quite chaotic.

How chaotic? Imagine we observe m at equal time intervals, and only write down not where m is, but whether it is in one side of the plane P or the other; say we write $+1$ if it is in one side, and -1 if it is in the other side. Then, the "motion" of m would be described in a course-grained way by a "trajectory" in time given by a sequence of ± 1; these sequences are called symbolic trajectories. Alekseev proved that for the simplified three-body problem above, the system is so sensitive to initial conditions that regardless of how long we observe this system, the symbolic trajectory generated by the deterministic Newtonian laws is completely indistinguishable from a completely random trajectory generated by the tossing of a coin (say, where we write 1 for heads and -1 for tails).

So, we come back to our initial question: how do we represent the outcomes of experiments? For deterministic *and* predictable cases, trajectories (even symbolic trajectories like the ones used by Alekseev) are a great way to do it. But trajectories do not include all the desirable characteristics of the three-body system; in a certain sense it does not faithfully represent the inherent unpredictability of it. We need to have a way to describe general processes, such as the motion of a planet, the tossing of a coin or die, or the changes in temperature in my house. To do so, we need to use a language that allows us to describe unpredictability and uncertainty: probabilities.

Modern probability arose from a question posed to Blaise Pascal, a devout Jansenist, by his friend, Chevalier de Mére. The problem was a simple one: how to figure out fair payoffs for each player in a game of chance if the game was interrupted before ending? To tackle this, Pascal had to developed the concept of probabilities. Probabilities are well-known to most of us, and it is not our goal to present a self-contained exposition here. Instead, we will focus on certain aspects of it that are relevant for our discussions in the next section.

Pascal's idea of probabilities was for a *normative* theory, and not *descriptive*. Descriptive theories attempt to simply describe how things are, whereas normative theories try to say how things ought to be. According to Boole (1854), "Probability I conceive as to be not so much expectation, as a rational ground for expectation." Or, perhaps more clearly stated by De Morgan (1847), the probability that an event has three times the probability to happen as to not happen means "that in the universal opinion of those who examine this subject, the state of mind to which a person ought to be able to bring himself is to look three times as confidently upon the arrival as upon the non-arrival." In other words, probabilities tell us what we should do if we are rational people, but do not tell us what most people do (in fact, many people violate basic tenets of probability theory; see de Barros and Suppes 2009; de Barros et al. 2016 and references therein).

In a nutshell, we can think of probability as a measure of belief for a rational agent.[1] Let an agent start with a set of propositions P_i and attach to them values $p(P_i)$ corresponding to their belief on the truth value of such propositions. We assume that this rational agent sets a higher value p to propositions thought to be more likely (e.g. "The Sun will rise in the East tomorrow") and lower values to less likely propositions (e.g. "A large earthquake will shake New York in the next few minutes"). Starting with a set of initial propositions and (hopefully justifiable) beliefs about them, it is natural for this rational agent to inquire as to the logical consequences of such propositions. This requires a calculus of belief values consistent with this logical structure (i.e., with an underlying Boolean logic of propositions), which is exactly what probability theory is (Jaynes 2003; Cox 1961).

To focus on the main concepts of probability theory that are relevant to us, let us consider a definition of probability put forth by Kolmogorov (1956). Imagine we have a set of *all* possible events (or propositions) $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, called the sample space or the set of elementary events. Given that Ω is the set possible outcomes, Kolmogorov defined his probability measure p to be zero for the empty set, i.e. the probability that nothing happens is $p(\emptyset) = 0$, and one for a certain event, i.e. the probability that *something* happens is $p(\Omega) = 1$, since Ω contains the all possible events.[2] So, we have that probabilities are valued between 0 and 1, with numbers in between 0 and 1 given different degrees of belief (the closer to one, the more certain we are about the proposition).

Given Ω and its elements, we can form other sets, such as, say, $\{\omega_1\}$, $\{\omega_3\}$, $\{\omega_2, \omega_4\}$, etc. In fact, we can even think of the set of all subsets of Ω, denoted 2^Ω. This set of all subsets forms what mathematicians call an algebra, which here we will refer to as $\mathscr{F} = 2^\Omega$. Kolmogorov defined a probability as a measure over elements of the algebra[3] $\mathscr{F} = 2^\Omega$. The idea here is intuitive. Imagine we have Ω as the set with four elementary events $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. If we have beliefs over the occurrences

[1] We will not give a detailed survey of different interpretations of probability, but only consider here one possible view, that espoused by Jaynes (2003). For details on the history of interpretations, the interested reader should refer to the wonderful book of Galavotti (2005).

[2] Other values for certain and impossible events may be given, e.g. $p(\Omega) = 100$ or $p(\emptyset) = -1$, but the choice of 1 and 0 for them makes the overall expressions for the calculus of probabilities simpler (Jaynes 2003).

[3] To be precise, he used an algebra over Ω which did not necessarily include all the subsets. As importantly, later on it was discovered that for certain sets we need to require the algebra to be countable, i.e. a σ-algebra, but such nuances are not relevant for our discussions.

of each one of the elements of Ω, then we should also have beliefs over their disjunction, e.g. we should be able to assign a probability to, say, $A = \{\omega_1, \omega_3, \omega_4\}$. And if we can assign a probability to A, then we should also be able to assign it to its conjunction with other elements of $\mathscr{F}$, e.g. $A \cap \{\omega_1, \omega_2, \omega_3\} = \{\omega_1, \omega_3\}$. More importantly, as Cox and Jaynes emphasized (Cox 1961; Jaynes 2003) and as we mentioned above, such assigned values of probabilities need to be consistent with this algebra. Therefore, probabilities are functions from a (Boolean) algebra of events (perhaps made up of elementary events) to the interval $[0, 1]$.

The final piece of the puzzle to get probabilities is to determine how we assign the values for different sets in a consistent way. Cox (1961) showed that under some reasonable assumptions of rationality, the (Boolean) structure of the underlying algebra of events impose as rule that, for two sets A and B in $\mathscr{F}$, probabilities must satisfy $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, which in the special case where A and B are disjoint reduces to $p(A \cup B) = p(A) + p(B)$. With those simple constructs at hand and the rules for computing p, we have a consistent way of assigning values to beliefs that are rational, a necessary step in providing a consistent narrative representation for natural phenomena.

However, probabilities are only part of the story. We started this section talking about mapping outcomes of experiments, and all that probabilities do is tell us what we can say about such outcomes in a coherent and rational way.[4] The tool that is needed are random variables.

Formally speaking, random variables are (measurable) functions that take elements from Ω into a set of numbers representing experimental outcomes. Let us look at a simple example: the game of craps. In this game two dice are thrown and payoffs are determined mainly by the sum of their values. For two dice, we can represent the set of possible outcomes by $\{\omega_{11}, \omega_{12}, \omega_{13}, \dots \omega_{66}\}$, where ω_{23} is the outcome of one die being 2 and the other 3. Since in this game we are only interested in their sum, we can create a random variable $\mathbf{X}$ that maps from the elements of Ω into the set $\{2, 3, \dots, 11, 12\}$, such that $\mathbf{X}(\omega_{23}) = \mathbf{X}(\omega_{32}) = 5$. If we now assume that each element of Ω is sampled with equal probability (representing our belief that the dice are not biased), it is straightforward to compute that $p(\mathbf{X} = 2) = 1/36$, $p(\mathbf{X} = 3) = 1/18$, and so on, which coincides with the standard expected probabilities for such outcomes in a game of craps.

Random variables can be used to account for a variety of phenomena. For example, we can use them to model games of chance, as shown above. We can also use them to model behavior, as has been done in stimulus-response theory (Suppes 1959), and in fact random variables are widely used in mathematical psychology, as well as sociology and economics.

So, to summarize, outcomes of experiments that have some randomness associated to them can be modeled by random variables. Notice that the only imposed

[4] i.e. if I have the belief that tomorrow it will rain with probability 0.5 and that a movie I want to watch has probability 0.3 of being interesting, and so on, I can rationally decide whether I will go to the beach or to the movies. Of course, probabilities only offer a measure of belief, and not whether I would extract more pleasure from one activity or another. To model this, economists use what is known as utility, a part of rational choice theory (Anand et al. 2009).

constraint to this representation of experimental outcomes is that they need to satisfy the simple rules of reasoning, i.e. classical logic. This was done through the use of a (Boolean) algebra of sets (see Jaynes 2003 for details). This requirement comes from our desire to be able to think of the processes involved in the description as part of a rational and consistent narrative. If, for some reason, our description violated the rules of simple reasoning, the told story represented by the model would not make sense. As we shall see in the next section, sometimes such rules are to restrictive to allow for a description of all experimental conditions.

When the Map Fails: Contextuality

As we indicated in the previous section, there are examples of situations where the description of possible experimental outcomes are not clearly representable with random variables. Let us examine some examples.

Imagine the case where Bob and Carlos are two good friends, and tend to have very similar views about arbitrary topics, almost always agreeing with each other when a question is posed to them. Now imagine an experimenter who provides Bob and Carlos random subjective questions and asks them whether they agreed with them or not. This experimenter is *only* interested in their answers, so she codes them as $+1$ for true and -1 for false, and models their responses in terms of random variables $\mathbf{B} : \Omega \rightarrow \{-1, 1\}$ and $\mathbf{C} : \Omega \rightarrow \{-1, 1\}$. The random variable $\mathbf{B}$ is, thus, a representation of the proposition "Was Bob's answer to my question yes or no?" with $\mathbf{B} = 1$ corresponding to "yes" and $\mathbf{B} = -1$ to "no." We do not know what Ω is in this case, but the experimenter verifies that the expectations of those random variables are the following, since the original questions were completely random: $E(\mathbf{B}) = E(\mathbf{C}) = 0$, as $p(\mathbf{B} = \pm 1) = p(\mathbf{C} = \pm 1) = 1/2$. Additionally, because they always agreed with each other, the experimenter also observed that $E(\mathbf{BC}) = 1$, since, for example, when Bob said "yes" Carlos also said "no."

Now let us imagine that the experimenter modified the situation slightly. Instead of asking the question to Carlos and Bob when they were alone in the room, she invited Alice to ask the question. Alice is a very smart friend of both Bob and Carlos, and whenever she is around, they became competitive, trying to outsmart each other, and start to always disagree. So, for this case, the experimenter still observes $E(\mathbf{B}) = E(\mathbf{C}) = 0$, but the joint expectation changes to $E(\mathbf{BC}) = -1$.

It is clear from the above example that two random variables $\mathbf{B}$ and $\mathbf{C}$ are *not* appropriate to represent the experimental outcomes. This is because the outcomes of the random variables are directly affected by the context. The presence of Alice in the room changes the behaviors of Bob and Carlos. This is a case where the experimental settings directly influence the random variables. What to do? Simply create different random variables for different situations, and index them accordingly (see Dzhafarov and Kujala 2017 and references therein). For example, we can call the Bob and Carlos only experiment "Context 1" and the experiment where Alice is there "Context 2," and then label the random variables $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{C}_1$, and $\mathbf{C}_2$, with the index referring to the context.

The above example is what some in the literature call "explicit contextuality" (de Barros et al. 2016). It is explicit because the actual expectations of a random variable (or, in this case, their joint expectation) change from one context to another. When this is the case, the cause for changes in the random variables are clear: in our example, the introduction of another person in the setup of the interview.[5]

Another type of contextuality is more subtle, and we call it "hidden contextuality.[6]" In contrast with explicit contextuality, here the observable expectation values do not change with the context. However, the expectation values can not be obtained from a proper joint probability distribution, implying that the random variables can not be considered to have definitive values sans context.

Consider the following example, imagine that Alistair, Briana, and Carol are philosophy students who always love to disagree with each other. Because of their schedule this semester, we never find all three of them in the same class, but we often find two of them. In other words, in Class 1 we will find Alistair and Briana, but not Carol, in Class 2 we will find Alistair and Carol, and in Class 3 we will find Briana and Carol. A researcher, not knowing that they always disagree between themselves, observes one class, and tries to model the answers of students with random variables, say $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively, in the same way as Bob and Carlos's setup. After observing several classes, the researcher tabulates the following expectations: $E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) = 0$, and $E(\mathbf{AB}) = E(\mathbf{AC}) = E(\mathbf{BC}) = -1$. At first sight, there does not seem to be any problem with those expectations, but if we analyze them closely, we see that this is not the case.

Let us take the joint expectations $E(\mathbf{AB})$, $E(\mathbf{AC})$, and $E(\mathbf{BC})$, and assume the researcher observed the same question in all three classes. Let us say that Alistair's answer was recorded as -1, which means that Briana's was 1. But in Class A, if Alistair answered the same thing, Carol would have answered 1 as well, which leads to a contradiction for the last class observation. What is happening here? Similarly to the explicit contextuality case, the random variables must be different in one context than in another, because it is not logically possible for three people to disagree on the same question. Once again, we could then propose different random variables for each observational condition, and no contradiction would arise.

The hidden contextuality example above seems a little contrived, not only because it requires unreasonable behavior from the three subjects, but also because each question is not the same, so it seems silly to record everything in terms of the same random variable. However, hidden contextuality appears also in important examples in physics, and we will explore an important one proposed in quantum mechanics by Greenberger et al. (1989).

[5] Some authors, most notably Dzhafarov and Kujala (2017), label this type of contextuality as "direct influences," to highlight the idea that the experimental context directly influences the variables in question. They reserve the label "contextuality" only for what we call here "hidden contextuality.".

[6] The term "explicit" and "hidden" contextuality was, as far as the authors know, proposed by Kurzyński (2017).

It would go beyond the scope of this paper to present the whole quantum mechanical formalism, but we will simply provide a set of experimental setups, and discuss their outcomes. Imagine we have three scientists, Angela, Brenna, and Clara, who have set up their labs very far away from each other. Each set up an experiment where every second three particles would be sent to each of their labs, and they agree to always either measure property X or Y of the particle entering their own lab. Measuring X (or Y) is equivalent to asking the question "Does the particle have property X?" Whether they measure X or Y would depend only on their own free choice, and they would record this property in their lab books. Since they are talking about the same property of different particles whose properties are behaving in a probabilistic way, it is natural to try to model this by random variables. Let us use $\mathbf{X}_A$, $\mathbf{Y}_A$, $\mathbf{X}_B$, $\mathbf{Y}_B$, $\mathbf{X}_C$, and $\mathbf{Y}_C$ as our random variables, with the subscript representing whose lab the particle was measured.

Once Angela, Brenna, and Clara analyze their data, the first thing that they notice is that the outcomes of their variables seem completely random, e.g. they look like a coin toss, with $E(\mathbf{X}_i) = E(\mathbf{Y}_i) = 0$, $i = A, B, C$. However, when they put all their data together, they see that some relationship exists between their outcomes when *all three of them* are combined. More specifically, they see that when, by chance, Angela measures X, Brenna measures X, and Clara measures Y, the product of their outcomes is always -1, or $\mathbf{X}_A\mathbf{X}_B\mathbf{Y}_C = -1$, a deterministic outcome! They also notice this for other products, namely that $\mathbf{X}_A\mathbf{Y}_B\mathbf{X}_C = -1$, $\mathbf{Y}_A\mathbf{X}_B\mathbf{X}_C = -1$, and $\mathbf{Y}_A\mathbf{Y}_B\mathbf{Y}_C = 1$. However, they also notice something strange with these outcomes. Imagine you take the product of all four perfect deterministic correlations above, namely

$$G = (\mathbf{X}_A\mathbf{X}_B\mathbf{Y}_C)(\mathbf{X}_A\mathbf{Y}_B\mathbf{X}_C)(\mathbf{Y}_A\mathbf{X}_B\mathbf{X}_C) = -1. \tag{17.1}$$

where the -1 comes from each term in parenthesis being -1. If we expand Eq. (17.1), we obtain

$$G = \mathbf{X}_A^2\mathbf{X}_B^2\mathbf{X}_C^2\mathbf{Y}_A\mathbf{Y}_B\mathbf{Y}_C, \tag{17.2}$$

which, from $\mathbf{X}_A^2 = \mathbf{X}_B^2 = \mathbf{X}_C^2 = 1$, and that $\mathbf{X}_i$ can be either 1 or -1, results in

$$G = \mathbf{Y}_A\mathbf{Y}_B\mathbf{Y}_C. \tag{17.3}$$

But we then reach a contradiction, namely that $G = \mathbf{Y}_A\mathbf{Y}_B\mathbf{Y}_C$ is 1 and -1, a mathematical impossibility.

Clearly something is wrong in our argument, as outcomes of experiments cannot yield a contradiction. A careful analysis show what the culprit is: when we say that $\mathbf{X}_A^2 = 1$, this is true if $\mathbf{X}_A^2$ is the same for two different experiments. Clearly, this assumption may not be true, and in fact the value of the random variables *must* change from experimental condition to experimental condition or context.

There are many other examples of contextuality in quantum physics, and we will not survey them here (the interested reader is referred to de Barros and Oas 2015). However, what we wanted to point out in this section is that the idea of mapping certain properties in a direct way, by creating a rational probabilistic account, does not work, as it does not fit the experimental data. In the next section we will see how we can extend probability theory to allow a description of this type of phenomena.

Alternative Mapping: Using Negative Probabilities

In section "Mapping Dynamics: Random Variables" we saw that if we impose a mapping bringing some phenomena into a structure that satisfies some simple criteria of rationality, we are lead to probability theory and random variables. Such a scheme works well for most phenomena, but fails when the quantities represented by random variables are context-dependent. In section "When the Map Fails: Contextuality" we saw some examples of when this map fails, i.e. examples of contextuality. We also saw that there are two types of contextuality: explicit contextuality and hidden contextuality.

Both types of contextuality show that there is some influence from the context into the outcomes of random variables. In both cases examined above, we saw that the assumption that one variable had the same value in a different context lead to a clear contradiction. Explicit contextuality can only appear as a direct consequence of some causally related event. A clear example is the Mach-Zehnder interferometer, shown in Fig. 17.1. If the interferometer is left undisturbed, the probability of a detection in D_2 is zero. However, if we attempt to measure which-path information, say by placing some type of detector in one of the arms of the interferometer thus determining whether a photon reflected off the bottom mirror or the top one, the probability of detection in D_2 jumps to 0.5, and the same for D_1. In other words, attempting to obtain which-path information affects the probabilities of outcomes for the output of the interferometer. Though this result is, in a certain sense, disturbing

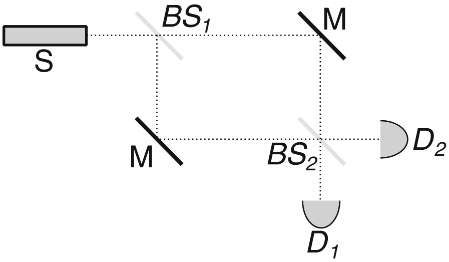


Fig. 17.1 Mach-Zehnder interferometer. A light source S emits photons that impinge on the beam splitter BS_1. The photon then has a 50% probability of going through to the upper mirror or being deflected to the lower mirror. The two possible paths are then mixed again by the beam splitter BS_2, resulting in two possible outcomes that are measured by detectors D_1 or D_2. If the interferometer is adjusted appropriately, due to interference effects, the probability of detection at D_2 is zero whereas the probability of detection at D_1 is 1

(see the discussion in Scully and Druhl 1982), it is not unreasonable, in the sense that the act of measuring may be causally related to the observed changes, since it precedes the detections in D_i.

The same cannot be said about the GHZ example. If we assume (which is what Einstein's theory of relativity imposes) that we cannot have interactions that travel with speed greater than light, then it becomes puzzling that the simultaneous measurements done by Angela, Brenna, and Clara seem to affect the correlation. How is this possible? For instance, if we were to try to come up with a "mechanism" for such correlations (what in the literature is called a hidden-variable model), this "mechanism" would have to rely on superluminal interactions (see the firefly example in de Barros et al. 2016).

So, does QM violate relativity? No. If we only look at the experimental outcomes, i.e. what Angela, Brenna, and Clara actually measure, all we can say is that they are related to each other. But to see this relation, we need to actually see all measurements at the same time. We cannot, for example, know what Angela measured by looking at Brenna and Clara's outcomes, simply because we cannot know what settings they decided to measure. Another way to think about this is the following: Angela, Brenna, and Clara cannot use this setup to send any information to each other. In other words, superluminal communication is not possible, nor is transfer of matter or energy (a type of superluminal communication).

So, is there a way to describe physical systems that are contextual but that do not violate relativity in terms of probabilities? One possible way (among many others) is to relax the rules of probabilities by allowing them to take negative values.[7] If we do so, all the rules of probability given by Kolmogorov still hold, such as defining probabilities over elements of the algebra $\mathscr{F} = 2^{\Omega}$, having $p(\Omega) = 1$ and $p(\emptyset) = 0$, and having the rule that for two disjoint sets in $\mathscr{F}$, the probability of their union is the sum of the probabilities, i.e. for $A, B \in \mathscr{F}$, $A \cap B = \emptyset$, we have

$$p(A \cup B) = p(A) + p(B). \tag{17.4}$$

However, we can relax the requirement that probabilities are defined over the interval $[0, 1]$, and allow them to take negative values. We emphasize here that when we loosen this constraint, we are not claiming that we can observe events with negative probabilities: this would be strange, to say the least, and inconsistent with the current view of probability, in the worst case. Instead, this only means that unobservable events can have negative probabilities, whereas observable events must always have non-negative probabilities.

To understand this, let us take a look at the simple example of three random variables $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ with expectations: $E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) = 0$, and $E(\mathbf{AB}) = E(\mathbf{AC}) = E(\mathbf{BC}) = -1$, examined in section "When the Map Fails: Contextuality".

[7] Another way is to use upper probabilities (Suppes and Zanotti 1991; de Barros and Suppes 2010). For example, Holik, Saenz, and Plastino showed that if we relaxed the requirement for a Boolean algebra and instead allowed for orthomodular lattices, one would get upper probabilities instead of standard probabiilty theory (Holik et al. 2014).

In this example, we never observe all three random variables simultaneously, but only in pairs. As such, the probability of having $\mathbf{A} = 1$ and $\mathbf{B} = 1$ is zero, whereas the probability of having $\mathbf{A} = 1$ and $\mathbf{B} = -1$ is 1/2, and so on. It is only unobservable events that have negative probabilities. For instance, let us try to compute the probabilities associated with the above expectations. From the rules of probability theory, we know that

$$p(\mathbf{A} = 1, \mathbf{B} = -1) = p(\mathbf{A} = 1, \mathbf{B} = -1, \mathbf{C} = 1) + p(\mathbf{A} = 1, \mathbf{B} = -1, \mathbf{C} = -1). \tag{17.5}$$

Notice that we cannot observe the three random variables simultaneously, and therefore cannot observe directly the probabilities on the right hand side, but we can still try to compute their probabilities from the expectations. For instance, from $E(\mathbf{AB}) = -1$ we have

$$p(\mathbf{A} = 1, \mathbf{B} = -1) = p(\mathbf{A} = -1, \mathbf{B} = 1) = \frac{1}{2}$$

and

$$p(\mathbf{A} = 1, \mathbf{B} = 1) = p(\mathbf{A} = -1, \mathbf{B} = -1) = 0.$$

By writing all those equations, and then using the idea expressed in Eq. 17.5, we can write a set of equations for the probabilities, namely

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} + p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.6}$$

$$p_{abc} + p_{\bar{a}bc} - p_{a\bar{b}c} + p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.7}$$

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} - p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.8}$$

$$p_{abc} - p_{\bar{a}bc} - p_{a\bar{b}c} + p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -1, \tag{17.9}$$

$$p_{abc} - p_{\bar{a}bc} + p_{a\bar{b}c} - p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -1, \tag{17.10}$$

$$p_{abc} + p_{\bar{a}bc} - p_{a\bar{b}c} - p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -1, \tag{17.11}$$

where we are using the simplifying notation that $p_{abc} = p(\mathbf{A} = 1, \mathbf{B} = 1, \mathbf{C} = 1)$, $p_{a\bar{b}c} = p(\mathbf{A} = 1, \mathbf{B} = -1, \mathbf{C} = 1)$, $p_{a\bar{b}\bar{c}} = p(\mathbf{A} = 1, \mathbf{B} = -1, \mathbf{C} = -1)$, and so on. In addition to the expectations, probabilities (even negative probabilities) also need to add to one (from $p(\Omega) = 1$), and we require

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} + p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = 1. \tag{17.12}$$

If we were to solve the above equations for the probabilities of simultaneously observing **A**, **B**, and **C**, we would obtain

$$p_{abc} = -p_{\overline{a}bc} = 2\alpha - 1, \tag{17.13}$$

$$p_{a\overline{b}c} = p_{ab\overline{c}} = \frac{1}{2} - \alpha, \tag{17.14}$$

$$p_{\overline{a}\overline{b}c} = p_{\overline{a}b\overline{c}} = \alpha, \tag{17.15}$$

$$p_{\overline{a}\overline{b}\overline{c}} = -p_{a\overline{b}\overline{c}} = \frac{1}{2} - 2\alpha, \tag{17.16}$$

where α is a parameter that is not fixed, since we had seven equations and eight variables. However, it is straightforward to see that we do not have non-negative solutions. For example, since $p_{abc} = -p_{\overline{a}bc}$, it follows that $\alpha = 1/2$ if we were to require them to be non-negative. But then, from the last equation, it follows that $p_{\overline{a}\overline{b}\overline{c}} = -1/2$, a negative probability!

As we can see from the above probabilities, none of the observable experimental outcomes have negative probabilities, as we mentioned above. So, what is the point of using negative probabilities? Additionally, what do they mean?

Let us first tackle the issue of meaning. There are some proposals to give meaning to negative probabilities, such as reinterpreting them in terms of violations of the principle of stability for probabilities (Khrennikov 1993, 2007, 2009a, b), thinking of them as events that can erase entries in a data table (Abramsky and Brandenburger 2014; Burgin and Meissner 2016; Burgin 2016), using an analogy to a half-coin through convolution coefficients (Ruzsa and Székely 1983; Székely 2005), or perhaps even coming from more ontological principles such as the indistinguishability of fundamental particles (de Barros et al. 2017).

However, in the absence of what we consider a satisfactory way to think about physical systems in light of negative probabilities, we take the rather pragmatic point of view that negative probabilities are simply an accounting tool, similar to what negative numbers mean. For example, it makes no sense to say that we have a negative number of apples, in the same way that it does not make any sense to say that we have a negative number of events in our frequency count of probabilities.[8] However, we can allow ourselves to start with a basket with 3 apples, tell John that we will give him 10 apples, and think of having a −7 number of apples right now, considering that the three we have and the next seven we get our hands on will go to John, and not to us. So, why not allow similar accounting processes in probabilities?

We note that historically negative numbers met with lots of resistance. For instance, as late as the 19th Century, the famous mathematician Augustus De Morgan (1910) wrote the following. “Above all, he [the student] must reject the definition still sometimes given of the quantity $-a$, that it is less than nothing. It is astonishing that the human intellect should ever have tolerated such an absurdity as the idea of a quantity less than nothing; above all, that the notion should have outlived the belief in judicial astrology and the existence of witches, either of which is ten thousand times more possible.” However, because negative numbers became an important tool in

[8] Of course, here we mean negative numbers as counting of actual objects. As is well known, other useful interpretations of negative numbers such as placement on the number line were introduced that make sense.

"bookkeeping" and essential in certain representations of mathematical ideas, the concept became more acceptable. Perhaps this will happen with negative probabilities.

Let us now tackle the other question raised above: what is the point of using negative probabilities, since we are neither computing probabilities that can actually be measured nor are we giving any meaning to negative probabilities? First, we point out that even though we are not giving meaning to a particular event, say p_{abc}, having negative probabilities, it does not mean that negative probabilities are completely meaningless. An interesting example comes from the fact that Eq. 17.12 has elements that all add to one, but some of them are negative. If we were to take the absolute value of each of the factors in Eq. 17.12, their sum would be greater than one. More importantly, the lowest bound of how much this sum exceeds one is a measurement of how contextual the system is (de Barros et al. 2015). To see this, let us modify our example with three random variables, and set them to the following expectations: $E(\mathbf{A}) = E(\mathbf{B}) = E(\mathbf{C}) = 0$, and $E(\mathbf{AB}) = E(\mathbf{AC}) = E(\mathbf{BC}) = -\varepsilon$, where ε is a number between 0 and 1. We already know that for $\varepsilon = 1$ we do not have a proper probability, and we need negative probabilities. It is also easy to see that for $\varepsilon = 0$ a standard non-negative probability exists. It is also possible to solve the system of linear equations below

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} + p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.17}$$

$$p_{abc} + p_{\bar{a}bc} - p_{a\bar{b}c} + p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.18}$$

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} - p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} - p_{\bar{a}\bar{b}\bar{c}} = 0, \tag{17.19}$$

$$p_{abc} - p_{\bar{a}bc} - p_{a\bar{b}c} + p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -\varepsilon, \tag{17.20}$$

$$p_{abc} - p_{\bar{a}bc} + p_{a\bar{b}c} - p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} - p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -\varepsilon, \tag{17.21}$$

$$p_{abc} + p_{\bar{a}bc} - p_{a\bar{b}c} - p_{ab\bar{c}} - p_{\bar{a}\bar{b}c} - p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = -\varepsilon, \tag{17.22}$$

$$p_{abc} + p_{\bar{a}bc} + p_{a\bar{b}c} + p_{ab\bar{c}} + p_{\bar{a}\bar{b}c} + p_{\bar{a}b\bar{c}} + p_{a\bar{b}\bar{c}} + p_{\bar{a}\bar{b}\bar{c}} = 1. \tag{17.23}$$

When we do so, we see at once that negative probabilities are not necessary when $\varepsilon \leq 1/3$, but that for $\varepsilon > 1/3$ there is no standard probability, but only negative probabilities, and the system is contextual (a result that agrees with Suppes and Zanotti 1981). When we compute the sum or the absolute values of the probabilities, let us call it M, and take its minimum value possible (which will be a function of ε), we see at once that $M_{min} = 1$ for $\varepsilon \leq 1/3$ and that M_{min} starts to increase as ε increases. In other words, the further the description deviates from a logical probabilistic one, the stronger are the underlying contradictions in this system, and the more contextual it is. Thus, M_{min} can be thought as a measure of contextuality (de Barros et al. 2015).

Knowing how contextual a system is, or rewriting certain properties in terms of negative probabilities can be valuable. For instance, in quantum computation contextuality is responsible for its "magic" (Veitch et al. 2012), and it is possible that the more contextual a quantum system is, the more computational resources it may be able to provide.

Additionally, being able to re-write contextuality in terms of an easily computable set of negative probabilities may provide some insight into the inner works of such systems. For instance, a long-standing question in physics is what makes quantum mechanics special, i.e. what are the physical principles that define quantum mechanics. Perhaps writing such principles in terms of negative probabilities, because of their simplicity, may help gain insight towards figuring out such principles (Oas and de Barros 2015).

Finally, there might be applications of negative probabilities outside of physics. For example, in recent years much debate has been happening about using the mathematical (contextual) tools of quantum mechanics to describe social phenomena. The idea is that certain observations of human behavior may be better modeled by probabilities defined over vector spaces, as it is done in quantum mechanics, and not by standard probability theory. Therefore, because of their inherent contextuality, human behavior may be thought as being *quantum-like*, a term used to differentiate them from actual quantum systems that exhibit contextuality because of its underlying quantum dynamics.

Examples of quantum-like human behavior abound (see the reviews by Khrennikov 2010; Busemeyer and Bruza 2012; Haven and Khrennikov 2013 for a more comprehensive survey). For example, human decision-makers violate Savage's Sure-Thing Principle (STP). STP states a simple idea in probability theory: given only two possible realizable scenarios A and $\neg A$, if B is preferred over $\neg B$ in scenario A, and if B is also preferred over $\neg B$ in scenario $\neg A$, then we should prefer B over $\neg B$ regardless of the outcomes of A (or our knowledge of such outcomes). It so happens that humans consistently violate this rule. Furthermore, violations of STP seem to be better described by quantum models than by models using classical probabilities. Examples of quantum models are the dynamical models found in the works of Busemeyer and collaborators (Busemeyer et al. 2009; Pothos and Busemeyer 2009), the contextual probability models of Khrennikov (2004, 2009a), Khrennikov and Haven (2009), or the use of quantum Bayesian networks by Moreira and Wichert (2016a, b, 2017), to name a few.

A more skeptical reader may object to using quantum mechanics to describe social phenomena, arguing that surely humans are classical objects, therefore subject to the laws of classical physics. First, we point out that the above models do not claim that humans are quantum, and not classical; they simply claim that they are better described by using the modified probability theory of quantum physics. Second, humans do not satisfy all the constraints of a quantum system; for example, quantum systems cannot have contextuality for the three random variable example we used above, but human decisions can (de Barros 2012a, 2015). Finally, as emphasized in de Barros and Suppes (2009), the type of non-classicality entailed by the quantum-like descriptions are not really spooky, like the true quantum ones. Whereas quantum systems cannot be modeled classically, the quantum-like contextuality that we see in social systems can (Khrennikov 2006; de Barros 2012b; Busemeyer et al. 2017).

Final Remarks

When we want to describe a process, we want to do so in a coherent way. This coherent description can be thought of as a mental map or model of the relationship between concepts and properties that we are trying to describe. Having a coherent map of such processes is one of the basic tenets of science: this is what scientists mean when they say that the world is understandable. However, we saw that sometimes it is difficult to produce such a description, at least with reasonable global coherency. This is true when systems are contextual.

In this paper we described in a somewhat non-technical way the issue of contextuality, which is pervasive in quantum mechanics but also shows up in social phenomena (Khrennikov 2010; Haven and Khrennikov 2013; de Barros and Oas 2015; Cervantes et al. 2017). Contextuality, as we showed, presents challenges to a random variable description. Contextual systems need to either have an increased sample space, indexing all random variables according to context (the approach of Contextuality by Default), or need to have some other way of describing them.

There are several different ways of describing contextual systems, such as quantum probabilities (which are defined over a lattice structure that is more general than a Boolean algebra), upper probabilities (which may add to more than 1), and negative probabilities, to name a few. Here we presented negative probabilities in more detail. Our goal has been to show that the idea of representing a contextual phenomena could be achieved without requiring a different logic of events and propositions. Negative probabilities provide a nice computational tool that may help us tell a little more to the story, to create a better map if you will.

However, keeping this algebra has a cost. Negative probabilities are non-monotonic. This means that counter-intuitive results arise from negative probabilities (for example, a subset of a set A may be more probable than A). Perhaps if we keep them bounded to actually observed non-negative marginal probabilities, such issues may be tamed, but this is still an open question. It is also unclear how negative probabilities may help understand certain fundamental questions, such as what are the physical principles that define quantum mechanics. For example, indistinguishability may be related to contextuality (de Barros et al. 2017; Kurzyński 2017), and perhaps the miscounting that happens when we confuse one property with another might give rise to negative probabilities, but none have demonstrated such thus far.

Negative probabilities also open up other research questions. For example, in conjunction with standard probability theory, negative probabilities may be used to create random variables that model contextual narratives, such as those including violation of STP. Perhaps even more interesting, negative probabilities may go beyond simple description, and possibly provide a normative answer to situations where contextuality plays an important rule (for an example, see de Barros 2014).

Regardless of the approach, be it indexation of random variables (Contextuality by Default), negative probabilities, upper probabilities, quantum logics, or quantum probabilities, they all offer different means to talk about mapping elements of reality into (at least partially) coherent narratives. Those means provide distinct insights into the difficulties of describing territories that are seemingly inconsistent, and may highlight the relevance of Alfred Korzybski's famous dictum that "the map is not the territory."

References

S. Abramsky, A. Brandenburger, An operational interpretation of negative probabilities and no-signalling models, in *Horizons of the Mind. A Tribute to Prakash Panangaden*, ed. by F. van Breugel, E. Kashefi, C. Palamidessi, J. Rutten, number 8464 in Lecture Notes in Computer Science (Springer International Publishing, 2014), pp. 59–75

J.A. de Barros, Beyond the quantum formalism: consequences of a neural-oscillator model to quantum cognition, in *Advances in Cognitive Neurodynamics (IV)*, ed. by H. Liljenström (Advances in Cognitive Neurodynamics (Springer, Netherlands, 2015), pp. 401–404

J.A. de Barros, Joint probabilities and quantum cognition, in *AIP Conference Proceedings*, vol. 1508, ed. by A. Khrennikov, A.L. Migdall, S. Polyakov, H. Atmanspacher (American Institute of Physics, Vaxjo, Sweden, 2012a), pp. 98–107

J.A. de Barros, Quantum-like model of behavioral response computation using neural oscillators. Biosystems **110**(3), 171–182 (2012b)

J.A. de Barros, F. Holik, D. Krause, Contextuality and indistinguishability. Entropy **19**(9), 435 (2017)

J.A. de Barros, J.V. Kujala, G. Oas, Negative probabilities and contextuality. J. Math. Psychol. **74**, 34–45 (2016)

J.A. de Barros, P. Suppes, Quantum mechanics, interference, and the brain. J. Math. Psychol. **53**(5), 306–313 (2009)

J.A. de Barros, P. Suppes, Probabilistic inequalities and upper probabilities in quantum mechanical entanglement. Manuscrito **33**(1), 55–71 (2010)

J. A. de Barros, *Decision Making for Inconsistent Expert Judgments Using Negative Probabilities*. Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2014), pp. 257–269

J. A. de Barros, E.N. Dzhafarov, J.V. Kujala, G. Oas, Measuring observable quantum contextuality, in *Quantum Interaction* ed. by H. Atmanspacher, T. Filk, E. Pothos, number 9535 in Lecture Notes in Computer Science, July 2015 (Springer International Publishing), pp. 36–47. https://doi.org/10.1007/978-3-319-28675-4_4

J. A. de Barros, G. Oas, Some examples of contextuality in Physics: implications to quantum cognition, in *Contextuality, from Quantum Physics to Psychology*, ed. by E.N. Dzhafarov, R. Zhang, S.M. Jordan (World Scientific, 2015), pp. 153–184

V.M. Alekseev, Quasirandom dynamical systems. I. Quasirandom diffeomorphisms. Sb.: Math. **5**(1), 73–128 (1968)

V.M. Alekseev, Quasirandom dynamical systems. II. One-dimensional nonlinear oscillations in a field with periodic perturbation. Sb.: Math. **6**(4), 505–560 (1968)

V.M. Alekseev, Quasirandom dynamical systems. III Quasirandom oscillations of one-dimensional oscillators. Sb.: Math. **7**(1), 1–43 (1969)

P. Anand, P. Pattanaik, C. Puppe (eds.), *The Handbook of Rational and Social Choice: An Overview of New Foundations and Applications* (Oxford University Press, Oxford, England, 2009)

G. Boole, *An Investigation of the Laws of Thought: On which are Founded the Mathematical Theories of Logic and Probabilities* (Dover Publications, Mineola, New York, 1854)

M. Burgin, An introduction to symmetric inflated probabilities, in *Quantum Interaction*. Lecture Notes in Computer Science, July 2016 (Springer, Cham), pp. 206–223

M. Burgin, G. Meissner, Extended correlations in finance. J. Math. Finance **06**(01), 178–188 (2016)

J.R. Busemeyer, P. Fakhari, P. Kvam, Neural implementation of operations used in quantum cognition, in *Progress in Biophysics and Molecular Biology*, May 2017

J.R. Busemeyer, P.D. Bruza, *Quantum Models of Cognition and Decision* (Cambridge University Press, Cambridge, 2012)

J.R. Busemeyer, Z. Wang, A. Lambert-Mogiliansky, Empirical comparison of Markov and quantum models of decision making. J. Math. Psychol. **53**(5), 423–433 (2009)

V.H. Cervantes, E.N. Dzhafarov, Advanced analysis of quantum contextuality in a psychophysical double-detection experiment. J. Math. Psychol. **79**, 77–84 (2017)

R.T. Cox, *The Algebra of Probable Inference* (The John Hopkins Press, Baltimore, 1961)

A. De Morgan, *Formal Logic: Or* (The Calculus of Inference, Necessary and Probable (Taylor and Walton, 1847)

A. De Morgan, *On the Study and Difficulties of Mathematics* (Open Court Publishing Company, 1910)

E.N. Dzhafarov, J.V. Kujala, Contextuality-by-default 2.0: systems with binary random variables, in *Quantum Interaction: 10th International Conference, QI 2016* ed. by J. A. de Barros, B. Coecke, E. Pothos, volume 10106 of Lecture Notes in Computer Science (Springer International Publishing, 2017). arXiv:1604.04799

J. Ford, How random is a coin toss? Phys. Today **36**, 40 (1983)

M.C. Galavotti, *Philosophical Introduction to Probability*, vol. 167 (CSLI Lecture Notes (CSLI Publications, Stanford, CA, 2005)

D.M. Greenberger, M.A. Horne, A. Zeilinger, Going beyond Bell's theorem, in *Bell's Theorem, Quantum Theory, and Conceptions of the Universe*, ed. by M. Kafatos, volume 37 of Fundamental Theories of Physics, (Kluwer, Dordrecht, Holland, 1989), pp. 69–72

E. Haven, A. Khrennikov, *Quantum Social Science* (Cambridge University Press, Cambridge, 2013)

F. Holik, M. Saenz, A. Plastino, A discussion on the origin of quantum probabilities. Ann. Phys. **340**(1), 293–310 (2014)

E.T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, Great Britain, 2003)

J.B. Keller, The probability of heads. Am. Math. Mon. **93**(3), 191–197 (1986)

A. Khrennikov, *Interpretations of Probability* (Walter de Gruyter, 2009a)

A.Y. Khrennikov, *Contextual Approach to Quantum Formalism* (Springer Science & Business Media, 2009b)

A. Khrennikov, Contextual approach to quantum theory, in *Information Dynamics in Cognitive, Psychological, Social and Anomalous Phenomena*. Fundamental Theories of Physics (Springer, Dordrecht, 2004), pp. 153–185. https://doi.org/10.1007/978-94-017-0479-3_9

A. Khrennikov, *Why so negative about negative probabilities?, in Derivatives: Models on Models* (Wiley, West Sussex, England, 2007), pp. 323–334

A. Khrennikov, p-Adic probability theory and its applications. The principle of statistical stabilization of frequencies. Theor. Math. Phys. **97**(3), 1340–1348 (1993)

A. Khrennikov, Quantum-like brain: "Interference of minds". Biosystems **84**(3), 225–241 (2006)

A. Khrennikov, *Ubiquitous Quantum Structure* (Springer, Heidelberg, 2010)

AYu. Khrennikov, E. Haven, Quantum mechanics and violations of the sure-thing principle: the use of probability interference and other concepts. J. Math. Psychol. **53**(5), 378–388 (2009)

A.N. Kolmogorov, *Foundations of the Theory of Probability*, 2nd edn. (Chelsea Publishing Company, Oxford, England, 1956)

P. Kurzyński, Contextuality of identical particles. Phys. Rev. A **95**(1), 012133 (2017)

C. Moreira, A. Wichert, Exploring the relations between quantum-like Bayesian networks and decision-making tasks with regard to face stimuli. J. Math. Psychol. **78**(Supplement C), 86–95 (2017)

C. Moreira, A. Wichert, Quantum-like Bayesian networks for modeling decision making. Front. Psychol. **7** (2016a)

C. Moreira, A. Wichert, Quantum probabilistic models revisited: the case of disjunction effects in cognition. Front. Phys. **4** (2016b)

G. Oas, J. A. de Barros, A survey of physical principles attempting to define quantum mechanics, in *Contextuality From Quantum Physics to Psychology*, ed. by E. Dzhafarov, R. Zhang, S.M. Jordan (World Scientific, 2015)

E.M. Pothos, J.R. Busemeyer, A quantum probability explanation for violations of ‘rational’ decision theory. Proc. Roy. Soc. B: Biol. Sci. **276**(1665), 2171–2178 (2009)

I.Z. Ruzsa, G.J. Székely, Convolution quotients of nonnegative functions. Monatshefte für Mathematik **95**(3), 235–239 (1983)

M.O. Scully, K. Druhl, Quantum eraser: a proposed photon correlation experiment concerning observation and "delayed choice" in quantum mechanics. Phys. Rev. A **25**(4), 2208–2213 (1982)

P. Suppes, A linear learning model for a continuum of responses, in *Studies in Mathematical Leaning Theory*, ed. by R.R. Bush, W.K. Estes (Stanford University Press, Stanford, CA, 1959), pp. 400–414

P. Suppes, M. Zanotti, When are probabilistic explanations possible? Synthese **48**(2), 191–199 (1981)

P. Suppes, M. Zanotti, Existence of hidden variables having only upper probabilities. Found. Phys. **21**(12), 1479–1499 (1991)

G.J. Székely, Half of a coin: negative probabilities. Wilmott Mag. **50**, 66–68 (2005)

V. Veitch, C. Ferrie, D. Gross, J. Emerson, Negative quasi-probability as a resource for quantum computation. New J. Phys. **14**(11), 113011 (2012)

V.Z. Vulović, R.E. Prange, Randomness of a true coin toss. Phys. Rev. A **33**(1), 576–582 (1986)

Part III
Mathematics/Computer Science

Chapter 18
Mathematics, Maps, and Models

Ian Stewart

The map is not the territory, but, as Alfred Korzybski wrote in 1933, that's why we make maps. In *Science and Sanity* (Korzybski 1933) he pointed out that 'a map *is not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness.' He traced these remarks back to the mathematician Eric Temple Bell's statement that 'the map is not the thing mapped'.

Bell is best known for his popular books on mathematics, such as *The Last Theorem*, *The Development of Mathematics*, and *Mathematics, Queen and Servant of Science* (Bell 1961, 1940, 1951). The most popular of them all is *Men of Mathematics* (Bell 1937), which inspired several prominent mathematicians to take up the subject. Historians criticise it for romanticising many stories, but that was part of its appeal. Science fiction buffs also know that Bell wrote fifteen science fiction novels under the pseudonym 'John Taine', starting with *The Purple Sapphire* (Taine 1924) and ending with *G.O.G. 666* (Taine 1954).

It's standard—though to some extent misleading—to carve mathematics up into two main categories: pure and applied. Pure mathematics, it is often claimed, is an abstract logical game that pays no attention to reality. Applied mathematics is practical and solves real-world problems. In actual fact, there's no clear-cut distinction of this kind, except as an administrative convenience in some institutions. Mathematics doesn't split neatly into two disconnected areas. Instead, ideas constantly flow from theory to applications and back again, enriching both. It's possible to distinguish two different *attitudes*: some mathematicians take their main inspiration from internal questions within the subject, while others are motivated by questions about the outside world. But these attitudes are two extremes of a far richer spectrum, not the only possibilities.

In these terms, Bell was a pure mathematician, but his historical works pay equal attention to applications. Applied mathematicians formulate mathematical models of real-world systems, and analyse the resulting mathematical questions to gain insight into the natural or human world. For example, in order to predict the effects of climate

I. Stewart (✉)
Mathematics Institute, University of Warwick, Coventry, UK
e-mail: I.N.Stewart@warwick.ac.uk

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_18

change and to understand its causes, we have to use sophisticated mathematical models. It's not possible to re-run the climate again under different conditions and discover what happens; we can observe only the reality in which we live. A model, on the other hand, can be run many times, we can choose the data we feed into it, and we can 'observe' any feature we wish without encountering practical obstacles.

However, a model can never be completely accurate. So applied mathematicians have their own version of Bell's and Korzybski's epigrams: *the model is not the reality*.

Even pure mathematics has its own arguments about map and territory, on a more philosophical and foundational level. In what sense does mathematics exist? What sort of 'thing' is it? Platonists assert that (in some mysterious sense that they never explain) the abstractions of mathematics have a genuine existence independent of human agency—territory. Others (notably Reuben Hersh in *What is Mathematics, Really?* (Hersh 1999)) see mathematics a collective human mental construct—a map. For Platonists, the mathematical map *is* the territory; for Hersh the map is a convenient way to present the territory, which resides in the minds and records of humanity.

In 1921 Albert Einstein addressed the Prussian Academy of Sciences in Berlin, on *Geometry and Experience* (Einstein 1921). A few sentences into his talk, he made a famous statement: 'As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality.' Critics of mathematical modelling often quote this remark to support claims that mathematics is useless. But the context for Einstein's remark offers no justification for rejecting mathematical modelling. He said:

> An enigma presents itself which in all ages has agitated inquiring minds. How can it be that mathematics, being after all a product of human thought which is independent of experience, is so admirably appropriate to the objects of reality? Is human reason, then, without experience, merely by taking thought, able to fathom the properties of real things?
>
> In my opinion the answer to this question is, briefly, this: As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality. It seems to me that complete clearness as to this state of things first became common property through that new departure in mathematics which is known by the name of mathematical logic…

It make little sense to interpret a remark preceded by 'mathematics… is so admirably appropriate to the objects of reality' as a rejection of mathematical models of the real world. Einstein was merely warning would-be users of mathematics, yet again, that *the map is not the territory.*

He went on to illustrate what he meant by discussing three different approaches to lines and points in geometry. One approach is the actual geometry of real space. A second is the traditional view of Euclid, in which lines and points are idealisations of the (presumed) geometry of real space. The third is the axiomatic approach to geometry pioneered by David Hilbert: what matters is not what the concepts *are*, but how they relate to each other. As Hilbert put it: 'One must be able to say at all times—instead of points, straight lines, and planes—tables, chairs, and beer mugs.'

Einstein's hidden agenda was to persuade his audience that the notion of curved space embodied in his theory of General Relativity can be reconciled with human visual intuition—in effect, by devising the correct map.

This is deep stuff. It involves the nature of mathematics, the status of its 'truths', how we arrive at those truths, and how the formal concepts of mathematics relate to real objects in the physical universe. What is the map? How credible is it? How is it drawn? What are its limitations as a description of the territory?

It used to be thought—indeed, assumed without question—that mathematics consists of absolute truths. It may or may not be raining today, but $2 + 2 = 4$ now and forever, throughout all time and space. In many cultures, the only truths that could compete with mathematical ones were those of the prevailing faith. But mathematics has the edge even on the Word of God, because mathematical truths have logical proofs.

We no longer see mathematics that way (and since the Enlightenment, religion has also been viewed somewhat differently in many circles). Mathematics is not about truth, but about deduction. A mathematician can reliably assert 'if A then B', but not just 'B'. In particular, a statement such as Pythagoras's Theorem is no longer seen as an unassailable property of real space, but something that's valid *provided* Euclid's axioms for geometry apply. Or, better, Hilbert's axioms, since Euclid's list omits many tacit assumptions.

One of Hilbert's axioms is: 'Of any three points on a line, at most one point lies between the other two.' This is obvious in a diagram, which is why Euclid failed to state it, but it's far from obvious from Hilbert's 'tables, chairs, and beer mugs' perspective. If three beer mugs are situated on a chair, does at most one of the mugs lie between the other two? Avoid the trap of thinking of actual chairs and mugs. The mathematical question is: *does this property follow logically from the axioms that Euclid stated?* The answer, as Hilbert showed, is 'no'. No blame should be attached to Euclid for missing logical fine points like this one. For the period in which he lived, he did a remarkable job. He even realised that his famous Parallel Axiom is necessary, because it can't be proved from his other axioms. It took roughly two thousand years for mathematicians to take this issue on board, culminating in the invention of two distinct types of non-Euclidean geometry.

The internal logical structure of mathematics provides consistent models for the true geometry of space—if such a thing exists—but mathematics alone cannot determine which model best fits reality. Mathematics presents us with a variety of possible maps, but no amount of investigation of the logical principles of geometric map-making can tell us which map best represents the territory. Only some kind of reality-check can do that.

It turns out that in real space, Pythagoras's Theorem is *false* in the neighbourhood of a massive gravitating body. So is the theorem that 'the angles of a triangle add up to 180 degrees.' Carl Friedrich Gauss was among the first to understand that there's no logical necessity for the geometry of space to be that of Euclid, opening up the intriguing possibility that it isn't. He tried to determine whether space is Euclidean by measuring the angles of a large triangle formed by three mountain peaks. However, his apparatus was insufficiently precise to come to any firm conclusion.

Einstein's General Relativity holds that space is definitely not Euclidean. Indeed, its geometry varies from one region to another, an effect that we call 'gravity'. The apparent attractive force of gravity is a manifestation of the curvature of space—its departure from the Euclidean model. General Relativity was initially confirmed by measuring the positions of stars in the sky during an eclipse, when the curvature of space caused by the gravitation of the Sun could be measured.

In 1893, Charles Lutwidge Dodgson, another mathematician who dabbled with history and the literature of the fantastic, wrote *Sylvie and Bruno Concluded* (Carroll 1893) under his usual pseudonym Lewis Carroll. One passage reads:

> 'What do you consider to be the largest map that would be really useful?'
>
> 'About six inches to the mile.'
>
> 'Only six *inches*?' exclaimed Mein Herr. 'We very soon got to six *yards* to the mile. Then we tried a *hundred yards* to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of *a mile to the mile!*'
>
> 'Have you used it much?' I enquired.
>
> 'It has never been spread out, yet,' said Mein Herr: 'the famers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well.'

It's widely understood that if the map is the same size as the territory, it's not much use. Later, I'll examine this belief more closely, but for the moment, let's accept that a map that's too large and too detailed isn't terribly useful, because the purpose of a map is to present important information compactly and comprehensibly.

For similar reasons, the aim of a mathematical model is not to represent reality *exactly*. Reality is too complicated. A model is useful if it *simplifies* reality without losing too much, and gives comprehensible insights. That's what mathematical models are for. The days when mathematical models were seen as Laws of Nature, exact descriptions of 'the system of the world', as Isaac Newton put it in his *Principia*, are long gone. When Relativity displaced Newton, it became clear that mathematical models only approximate reality.

It might be assumed that the better the approximation, the better the model. That's perhaps true in the deep philosophical sense of improving the *rules* of the model: we want the best possible description. For example, Einstein's original motivation for General Relativity was a minuscule anomaly in the motion of the planet Mercury. Its orbital ellipse slowly rotates, under the influence of the other planets in the solar system; the technical jargon is 'precession of the perihelion'. But Newton's Law of Gravity predicts a rotation rate that is very slightly different from what's observed.

Einstein used Relativity to calculate the rate of rotation of the elliptical orbit, and after a few mistakes, his result was almost exactly the observed rate. This was an impressive vindication of Relativity. The previous theory held that the discrepancy was caused by an unknown planet, dubbed Vulcan, orbiting between the Sun and

Mercury, and that Newton's Law of Gravity would then explain the observed value of the precession of the perihelion of Mercury. The great astronomer Urbain Le Verrier, who had discovered Neptune by exactly this method, was awarded the prestigious Legion d'Honneur on the basis of this prediction. But no one could find Vulcan, and when Einstein showed it wasn't needed, Relativity won the day.

For some purposes, Einstein's improvement to Newton is essential. SatNav systems have to use Special Relativity to compensate for the motion of the satellites that transmit navigation signals, and General Relativity to account for the effect of the Earth's gravitational field. But there's a downside to Relativity: the calculations are more complicated. The elliptical orbits spotted by Johannes Kepler and calculated by Newton solve, exactly, the problem of the motion of two bodies under gravity. But in General Relativity, this two-body problem seems not to have an explicit solution, so numerical solutions found by computer are used instead. Most of today's space missions are planned using *Newton's* model of gravity. It may be less accurate in a deep philosophical sense than Einstein's, but in a purely practical sense it gets sufficiently accurate answers more simply. Quite literally, it's 'good enough for government work'.

At the moment, even our best physical theories of reality are imperfect approximations to whatever it is that the universe actually does. I confess that I occasionally get the impression that some fundamentalist physicists don't recognise this when it comes to Quantum Mechanics, because it predicts many physical quantities to remarkable accuracy. But since in some circumstances Quantum Mechanics and Relativity contradict each other, there are reasons to believe that even that very successful model is imperfect. Albeit by a tiny amount. But then, the precession of the perihelion of Mercury was very tiny too…

Mathematical modelling is far from straightforward, and a classic example is fluid flow. Mathematical models of fluid flow go back to Euler and Bernoulli, who continued where Newton left off, formulating models of the physical world in the form of partial differential equations. Later, Claude-Louis Navier and George Gabriel Stokes modified the equations to take account of fluid viscosity (stickiness). The resulting 'Navier-Stokes Equations' are widely used by engineers to understand how water flows past ships and submarines, and how air flows past aircraft, road vehicles, and racing cars.

Until recently, the main tool for aerodynamic design was the wind tunnel. Scale models of aircraft or cars were constructed, and placed in a large tube through which air was blown at high speed by powerful fans. Then various features of the flow, in particular the amount of air resistance (known as 'drag' in Formula 1 and Indycar racing) was measured. Then another slightly modified model was built and analysed in the same way. The process was slow and expensive.

However, it has been found that solutions of these equations resemble real flows so closely—quantitatively as well as qualitatively—that in many areas of engineering there's no longer any need for wind tunnels. Computer Fluid Dynamics (CFD) has taken over almost completely. Large numbers of alternative designs can be analysed rapidly and (after the initial investment on hardware and software) cheaply. Moreover, virtually any desired measurement can be extracted from the computer simulation, and the flow can be visualised in many different ways. This is in stark contrast to wind-tunnel experiments, where measurements are difficult, often impossible, and may interfere with the flow.

Here the mathematical map—the Navier-Stokes Equations—is certainly a very useful representation of the territory—physical fluid flows, in complicated and realistic geometry. However, the map is definitely not the territory, for two contradictory reasons. The Navier-Stokes Equations are continuum models, the mathematician's name for equations that assume the physically world is infinitely divisible. In effect, the equations describe the motion of a fluid that can be divided into infinitely many infinitesimal 'fluid elements'. (Technically, into as large a number of elements as we please, each as small as we please, in such a manner that the more elements we use, and the smaller they are, the more accurate the equations become.)

The territory is not like that. If we subdivide a fluid, eventually we get down to the atomic scale, and it becomes indivisible. Except perhaps by nuclear reactions, but at that scale atoms also cease to behave like fluids. So for some reason, the idealised model of a continuum behaves like a real fluid, despite being based on an assumption that is *false* for real fluids. One reason is that although fluids are not infinitely divisible, they can certainly be subdivided into extremely small regions. But mathematically, the idealisation to infinitely small regions is simpler, more convenient, and works amazingly well.

That's one difference between map and territory. The other difference points in exactly the opposite direction, and it involves a second map. Namely, the method used on the computer to solve the Navier-Stokes Equations. The equations are complicated and 'nonlinear', which among other things means that they can't be solved by an explicit formula—unlike the elliptical orbits of two bodies in Newtonian gravitation. The standard remedy, now that we have fast computers with huge memories, is to use numerical methods to *approximate* the solution. Typically the fluid is divided into a fairly coarse grid, like a fishing net, and the fluid velocity is calculated only at points of the grid—where the strings of the net join together. The jargon is that space is 'discretized'—the continuum assumption is deliberately broken. The same goes for time, which in the Navier-Stokes Equations is also assumed to be continuous. In the computer approximation, time clicks on in tiny but finite steps.

Finding effective discretisations is something of a black art, especially if the geometry of the moving object—such as a passenger jet—is complicated. The squares of the mesh are chosen to be smaller wherever the shape of the object changes on a shorter length scale—corners, or places where it's strongly curved, for example. Over the wings, the grid squares might be quite large, but at the edges of the wing they have to be refined into a finer grid. Of course they're not always squares, and in adaptive grids the regions of the grid can change as time passes. The grid may even flow with the fluid rather than be attached to the aircraft. It all depends on what you want to know and how much computation you can afford to carry out.

So here we have a 'map of the map' which breaks the continuum assumption—the main difference between the Navier-Stokes map and the real territory. However, the map of the map (the numerical grid method) is even less realistic than the map, because it breaks the fluid up into pieces that are *too big* to be realistic. Real fluids subdivide into much smaller regions than those of any practical grid, and within those regions they don't flow in the simple manner assumed by the numerical method.

And yet… *it all works.* So well that aircraft manufacturers use CFD almost exclusively. It's more accurate than wind-tunnel measurements. But what a strange procedure! First, we make a map of the territory that we know includes a feature that the territory doesn't possess: infinite divisibility. Having done that, we then make a new map *of the map* that throws that assumption away, but still fails to mimic the territory. Yet, when we do these contradictory things, we obtain a practical method for calculating fluid flows that represents the territory with remarkable felicity.

In some areas of engineering, the continuum modelling step is discarded from the start. The elastic stresses in, say, the metal frame of a tall building, are computed using 'finite element methods'. In this approach, a metal girder is represented as a large but finite collection of small rigid elements, joined to each other by springs. The springs model the elastic forces within the girders. Tracking how the elements move as time clicks past in discrete steps, engineers can calculate how the girder will respond to the forces that act on the building, such as gravity and wind pressure. Again, the result are astonishingly accurate if you set up the finite element model correctly.

As is so often the case with mathematical maps used by engineers, physicists, even biologists, the map is not just different from the territory: it's *better*. As any good map should be, for its intended purpose.

The most challenging territory in science is the human brain. We know from the inside of our own brain that it can do remarkable things. The area now known as neuroscience has been struggling for centuries to figure out even the tiniest part of how the brain works. The same goes for other parts of the nervous system: how do we see, hear, and walk? Today, two very ambitious international research programmes are aimed at understanding the brain: the European Human Brain Project and the American Brain Research through Advancing Innovative Neurotechnologies® (BRAIN) Initiative. For both projects, a difficulty arises. It's one thing to map the structure of the brain, either as an artificial construct or a vast and complex network of observed interconnections. It's quite another to figure out what the resulting structure does, and why, and how.

This is where simpler, older methods come into play. Until recently, the main way to gain insight into brain function was through models. I'll discuss one simple type of model that I've worked on myself, with various colleagues, because I *know* that it's far too simple to represent reality accurately. Not only isn't it a brain: it's not even a plausible *component* of a brain. Yet it makes testable predictions that seem to match many experiments, which is fascinating.

The context is the human visual system. How does the brain turn signals from the eyes into a vivid impression of the outside world? A common scientific technique for figuring out how something works is to interfere with it and make it go wrong. That often gives insights into what it does when it goes right. Human vision can be fooled in many different ways, often lumped together as ‘visual illusions’. There are many kinds. Ambiguous figures can have several different interpretations, with the brain switching rather randomly between them, unsure of which is correct. A famous one resembles a duck if you think it’s facing one way, but a rabbit if you think it’s facing the other way. Static shapes can appear to be moving. Parallel lines can seem to converge. Two regions of the visual field can appear to be dark and light when in fact they’re the same shade of grey. Regions that appear to have different colours can actually be coloured identically. Then there are impossible figures, in which small regions match possible real-world objects, but the overall shape is geometrically inconsistent. The artist Maurits Escher used these figures in his work—stairs that follow closed loops yet seem to ascend forever, a watermill in which water flows from the base of the wheel, goes downhill, and returns to the top. Each of these illusions sheds some light on how the visual system works.

In 1593 Giambattista della Porta discovered another type of visual effect, and reported it in his *De Refractione: Optices Parte: Libri Novem* (Della Porta 1593). He took two books and put one of them in front of one eye and the other in front of the other eye. He reported that he could read from one book at a time, and that changing from one to the other required withdrawing the ‘visual virtue’ from one eye and moving it to the other. We now call such an effect *binocular rivalry*. In its simplest form, the observer alternately switches perception from one of the presented images to the other. But in some experiments, observers report seeing additional images that weren’t shown to either eye.

A famous case is the ‘monkey-text experiment’ reported by Ilona Kovács and colleagues (Kovács et al. 1996). This begins with two images: one of a monkey (probably a juvenile ape, but it’s generally referred to as a monkey) and the other of some text. Each image is cut into six pieces, which are reassembled to create two ‘scrambled’ images with three regions of monkey and three of text. One scrambled image is shown to the left eye, and the complementary scrambled image to the other eye. Most observers report that their perception alternates between the two scrambled images. However, some see alternation between a complete monkey and a complete text image. This experiment is evidence that the visual system processes incoming images by breaking them down into components and then reassembling the parts, a process known as ‘binding’. Sometimes the binding process goes wrong, a clue that it must exist.

One way to investigate this kind of decision-making is to construct a simplified network of model neurons and analyse its dynamics mathematically. The idea goes back to the neuroscientist Hugh Wilson, who used it to model general features of decision-making in the brain. In 2013 Martin Golubitsky and coworkers specialised it to decision-making in the visual system (Diekman et al. 2013).

In this experiment, both images split into two distinct regions: call them A and B. One image has ‘monkey’ in A and ‘text’ in B. The other has ‘monkey’ in B and ‘text’ in A. The network has four model neurons corresponding to these possibilities:

monkey in region A

text in region A

monkey in region B

text in region B

A passive neuron just stays in the same rest state. An active neuron outputs a series of electrical spikes, and the more rapidly those spikes occur, the more active it is. The model assumes that in any competition between neurons, the more active one wins. That is, the network detects the interpretation represented by the more active neuron.

To present an image to this network we stimulate the corresponding neurons, making them more likely to become active. We train the network to recognise the two scrambled images by adding links between neurons. Some are inhibitory: they ensure that the network does not detect two different things in the same region at any instant. Others are excitatory: they join neurons that represent the components of a given scrambled image, so that if part of that image is detected, the network is more likely to perceive the rest of it. Briefly, ‘monkey in A’ links to ‘text in B’, and ‘monkey in B’ links to ‘text in A’. There are no other links.

The key mathematical question is: what does the model network ‘perceive’ when presented with the two scrambled images? General theory and computer simulations show that it can behave in two ways. One is to alternate between the two scrambled images. The other is: alternate between a complete monkey and complete text. Now both neurons for monkey are active, then they become passive while both neurons for text becomes active, and then the sequence repeats. The corresponding neurons are not linked, yet they sometimes behave as if they are, because of all the other links.

These two types of behaviour are robust predictions from the mathematics, and they are exactly what happens in experiments. No one seriously imagines that the brain really perceives these images using just four neurons—yet that model makes correct predictions. A more realistic model could use a population of neurons, linked in some sort of circuit, rather than just one neuron. Much the same mathematics leads to the same predictions: either alternation of the scrambled images, or of the unscrambled ones. The unscrambled images are a little surprising because they weren’t shown to the eyes as complete images. Somehow the visual system in the brain has reassembled parts of distinct images in a different way—presumably reflecting prior knowledge of what monkeys and text look like.

Similar models, often with more elaborate networks, make accurate predictions about other experiments on rivalry. The success of such simple models raises an intriguing question: is the visual system wired in a similar manner to the models? Does the visual system use connections between neurons to encode learned information about images, causing it to react more strongly to those images, thereby recognising them?

The part of the brain that processes images received by the eyes is the visual cortex, a structure composed of layers of neurons, wired together in complex ways. Recently, Maria Florencia Iacaruso and colleagues have reported intriguing discoveries about the wiring of the visual cortex, revealing the presence of encoded learned patterns (Iacaruso et al. 2017). Neurons in the top layer of the visual cortex detect boundaries between different parts of an image, and also the direction in which those boundaries are pointing. When we look at a window, some of those neurons detect the vertical edges of the window, and others detect the horizontal edges. This process is a first step in 'segmenting' the image into pieces that the brain can compare to things we have previously encountered. The recent discovery is that neurons that detect a particular direction are linked (by excitatory connections) to *some* of the other neurons that detect the same direction. Which ones? Those whose locations in the cortex are situated along the corresponding line. In effect, the visual cortex has learned to recognise straight lines. Some of these connections may be 'hard-wired' by evolution, others are modified during a child's growth. Even adult brains can change parts of their wiring diagrams, and the strengths of specific connections.

We conclude that a type of model that's obviously far too simple to represent the *real* territory of the visual system nevertheless agrees with many experiments. It provides useful hints about the organised complexities of the real system. Crucially, its simplicity makes it comprehensible. Just as a map of Helsinki makes the city comprehensible when you're looking for a restaurant.

Many commentators, from Lewis Carroll to Jorge Louis Borges, have poured scorn on maps that are the same size as the territories they represent. The presumption is that size alone obviously makes such maps useless. However, this belief stems from conventional paper maps, representing geographical information. An image of a bacterium in a microscope is also a map, and it's far larger than the territory. For many purposes, from scientific research to medical diagnosis, what counts is the map. So although the map is not the territory, it may be superior to the territory. In fact, the main reason for creating a map is that it's superior to the territory for some useful purpose. A paper map of Helsinki folds up and fits in your pocket. It's probably superfluous to point out that you can't do this with Helsinki itself. When you're heading for a museum or a restaurant, you consult the map, not the city. The city is the problem; the map is the solution.

However, a map that's the same size as the territory can be useful: for instance, when it's a computer representation that can be interrogated to find out how the real system is likely to behave, in fine detail. A familiar example is Google Earth. The closer it gets to reality, the *more* useful it is. Digital technology has now advanced to such a level that it's possible to carry in your pocket a map of an entire country with sufficiently fine resolution to show not just every house, but every car; indeed, every cat and dog that was visible to the satellite overhead when it took a picture of the area. In fact, it's often impossible not to carry such a map, because the operating system won't let you delete it. Much of the detail may exist up in the Cloud, to be downloaded as required—but the Cloud is part of the map too. The information might be five years old, but that's a temporary limitation based on cost. It's already technologically feasible to show such features in real time, and on a scale of one to one.

Indeed, this kind of computer map can actually contain *more* than the territory. A virtual map in a computer or on a phone can be annotated with additional information—names of towns and streets, restaurant reviews and menus, museum opening times, your own movements for the past decade… Some of this information also resides in the territory, but it's easier to access it on an app. Some of it can be found only in the app, because past history is automatically erased from the territory unless someone or something records it. This kind of 'enhanced reality' will soon become commonplace, especially when the computer doesn't live in your pocket, but in your spectacles, where it can interface directly with your eyes.

Mathematical models are becoming equally advanced. There exist highly detailed models of entire cities, down to the level of every individual inhabitant. Researchers use mobile phone data to track overall patterns of movement of people and vehicles. City authorities use such models to control traffic, design street layouts, and predict the effect of a football match. Companies sell software and operate services to predict crowd flow in airports, railway stations, shopping malls, and Olympic stadiums. The results can be used in many ways: to make retail services more obtrusive and therefore (allegedly) more profitable, or to stop crowds reaching dangerous densities.

There's now a vogue for collecting and storing 'big data'. At the moment this trend is mostly unmatched by any clear idea of how to extract useful information from the data—the attitude seems to be 'if you can't find the needle, build a bigger haystack'. But as mathematicians and statistician learn how to sift the needles from the straw, gigabytes of numbers may yet be turned into useful insights. In the rush to amass data, we should never forget that the central goal is not raw information, but refined understanding. No amount of fancy electronic gadgetry can change one basic point: a mathematical model, either on paper or programmed into a computer, is only as good as the assumptions that are built into it.

The app is not the territory.

References

E.T. Bell, *The Last Theorem* (Simon & Schuster, New York, 1961)

E.T. Bell. *The Development of Mathematics* (McGraw-Hill, New York, 1940)

E.T. Bell, *Mathematics, Queen and Servant of Science* (McGraw-Hill, New York, 1951)

E.T. Bell, *Men of Mathematics* (Simon & Schuster, New York, 1937)

L. Carroll, *Sylvie and Bruno Concluded* (Macmillan, London, 1893)

G. Della Porta, *De Refractione: Optices Parte: Libri Novem* (Io. Iacobum Carlinum & Antonium Pacem, Naples, 1593)

C. Diekman, M. Golubitsky, Y. Wang, Derived patterns in binocular rivalry networks. J. Math. Neuro. **3** (2013). https://doi.org/10.1186/2190-8567-3-6

A. Einstein, Geometry and experience. *Königlich Preußische Akademie der Wissenschaften* (Berlin) Sitzungsberichte (1921), pp. 123–130

R. Hersh, *What is Mathematics, Really?* (Oxford University Press, Oxford, 1999)

M.F. Iacaruso, I.T. Gasler, S.B. Hofer, Synaptic organization of visual space in primary visual cortex. Nature **547**, 449–452 (2017)

A. Korzybski, *Science and Sanity: an Introduction to Non-Aristotelian Systems and General Semantics* (Institute of General Semantics, Forest Hills, 1933)

I. Kovács, T.V. Papathomas, M. Yang, A. Fehér, When the brain changes its mind: interocular grouping during binocular rivalry. Proc. Natl. Acad. Sci. USA **93**, 15508–15511 (1996)

J. Taine, *The Purple Sapphire* (Dutton, New York, 1924)

J. Taine, *G.O.G. 666* (Fantasy Press, Reading, 1954)

Chapter 19
A View from Space: The Foundations of Mathematics

Jean-Pierre Marquis

Introduction

Suppose we were to meet with extraterrestrials and that we were able to have a discussion about our respective cultures. At some point, they start asking questions about that something which we call "mathematics". "What is it?", they ask. Tough question. How should we answer them?

We start with elementary examples. We illustrate how we count simple things and check that they understand that. We then move to arithmetic more generally and talk about natural numbers $1, 2, 3, \ldots$ and arithmetical operations. We then move to geometric figures, the most common and simple ones—triangles, squares, pentagons, circles, etc.—and how we measure them. To our dismay, they suddenly seem utterly puzzled. They understand numbers and geometric figures, but not how this makes *one* thing. It has to do with counting and measuring, we tell them. It gives answers to the questions "How many?" and "How much?". Counting and measuring are two different operations that yield different kinds of results, they reply. On the one hand, whole numbers and on the other, they use a word we don't understand, but it does not seem to belong to the category of numbers … Moreover, they tell us, these have to do with *our* operations. So, is mathematics about us? We are puzzled. That does not seem right. Or, is it? We are lost. We try to give them more examples, to illustrate the diversity inherent to mathematics. We introduce ideas about polynomial equations and how to solve them, then move on to differentials and integrals, probabilities and statistics and we do our best to give them a better, more complete picture of mathematics. They were able, apparently, to understand the specific details of each and every one of our examples. We did have to clarify certain points, but they were able to quickly come back and verify with us that they had understood. They are still not

J.-P. Marquis (✉)
Department of Philosophy, Université de Montréal, Montreal, Canada
e-mail: jean-pierre.marquis@umontreal.ca

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_19

satisfied. In fact, they are even more puzzled. The conversation goes back and forth, but to no avail.

After quite some time, they pause. Then, as hit by lightning, one of them asks a very simple question: what is this mathematics of ours *about*?

It is our turn to be puzzled. We do not know what to say. What *is* mathematics about? Should we bring them to a library and show them all the books we have on mathematics? We know that mathematics is rich, vast and complex and we are far from sure that a tour of the section on mathematics in a library will answer their question. Then we realize that we can tell them something. We do have an answer. And it is, in fact, quite nice.

Foundations of Mathematics: The Global Picture

Let us leave our extraterrestrial friends aside for a moment and focus on their question: what is mathematics about? Of course, the answer to that question changed over the centuries, at least in western civilization. For a very long time, mathematics was considered to be about quantity, discrete and continuous. It was the "science" of quantity.[1] In the early 20th century, motivated by the creation of a new logical machinery and mathematical results garnered in the 19th century, original proposals were put forward by mathematicians, philosophers and logicians. Some of them proposed that, in the end, mathematics was reducible to logic. Of course, this amounts to displacing the problem and ask what is logic about. Saying that the latter is about relations between universals does not seem to help much. At least, not on its own. (See, for instance, Hintikka 2009.) Others proposed that, in the end, mathematics was about a basic temporal intuition, a basic operation of consciousness in time, a breaking up of time in two distinct moments. According to this view, mathematics *is*, ultimately, about us and its adoption required that we change important aspects of mathematics as we knew it. (See van Dalen 2000; van Dalen and van Atten 2002.) Still others proposed that mathematics was akin to a game with symbols, the rules of which were constrained by social, psychological and natural constraints. Thus, in the latter, mathematics is simply *not* about anything. (For a more subtle discussion, see Simons 2009; Sinaceur 1998.) These descriptions are of course too coarse and too short to do justice to each of these positions. The fact is that none of them were accepted by the community has rendering justice to the richness of mathematics.

Amidst these philosophical discussions, arguments and theories, a new *mathematical* framework found its way: set theory. And it covers all of mathematics as we know it. According to this theory, mathematics is about sets or collections and the

[1] It should be pointed out that that there was also a distinction between pure mathematics and what was called for a long period of time 'mixed' mathematics. The latter distinction was replaced by the contemporary distinction between pure and applied mathematics in the 18th and 19th century. See, for instance, Stedall (2012) for an historical overview, and Maddy (2008) for the rise of the distinction between pure and applied mathematics.

operations on them.[2] That is it. More specifically, it is all based on a very simple relation, that is saying that something, call it *a*, belongs to a set, call it *A*, and written '$a \in A$', also read '*a* is an element of *A*'. It is understood that such a statement is either true or false, depending on what '*a*' and '*A*' refer to. The whole of mathematics is then based on a list of basic propositions, called 'axioms', which state the essential properties of sets and these axioms, together with the infinite logical consequences it generates constitute what is called 'set theory'. We will not give all the axioms of the theory here, for it would take us too long. We will present a few of the axioms to illustrate the theory.

The first axiom determines when two sets are identical. Two sets *A* and *B* are identical if and only if the have the same elements. Another axiom claims that there is an empty set, that is a set with no element. There is also an axiom stating that there is a set with an infinite number of elements. There are axioms asserting that the intersection of two sets exists as well as their union. And there are others, more technical axioms. We have written these axioms in ordinary language, but they are usually written in a formal language, based on first-order logic. The theory is called Zermelo-Fraenkel set theory, in honour of two mathematicians, Ernst Zermelo and Abraham Fraenkel, who wrote down these axioms in the early 20th century.[3] Since most of mathematics can be derived from the axioms of the theory with the help of logic, it is considered to constitute a foundation for mathematics. We can thus provide a partial, but instructive answer to our guests: mathematics is about sets or collections in *that* particular sense. Once one understands the axioms and knows how to reason with the help of first-order logic, one can see what mathematics is about.[4]

Let us come back to our extraterrestrial friends. They would most likely understand our set theory. Of course, we may be wrong about this and it would be fascinating to see *what* they would not understand and *why*. However, let us assume that they *do* indeed understand it. But after a while, they might be intrigued about a specific aspect of our answer. It is fine to say that mathematics is about sets, but surely, there must be some ideas, concepts and propositions, apart from the axioms, of course, that are more important than others. Since clearly there are infinitely many possible logical consequences to our theory, how do we determine which ones are worth pursuing? Do we deduce propositions form the axioms randomly? Are they all of equal importance? They would certainly point out that we seem to have suddenly forgotten about our whole numbers, our geometric figures, our probabilities, etc. Where have they gone?

For one thing, mathematicians do not randomly deduce consequences from the axiom of set theory. Our answer as to what mathematics is about was tailored to our friends' question. It does not represent what mathematicians do on a daily basis.

[2] One of the standard references on the subject is Jech (2003). For a philosophical discussion, see Maddy (2011a, b).

[3] For more on the history of set theory, see Ferreiró (2007).

[4] We could be even more specific here, for there is a picture of a universe of sets, called the cumulative hierarchy, that is used to interpret the axioms and understand the structure of these sets. It is not necessary to present it at this stage and we will refrain from doing it.

Far from it. But it provides mathematicians with a simple language, a uniform way to define and construct mathematical entities and it comes with certain proof techniques. Furthermore, mathematicians also developed in that period what they called the 'axiomatic method', or what might be better called the 'abstract method'. (See, for instance, Marquis 2015, 2016.) In a nutshell, the idea is that mathematicians identified certain concepts that play an important role in mathematics and wrote down the main properties of these concepts, the axioms for that concept, and developed the ensuing theories. Thus concepts like that of group, ring, field, vector space, topological space, partial order, lattice, Hilbert space, Banach space, measure space, etc., were all defined and developed using that method. Natural numbers, integers, rationals, reals, complex numbers, etc. can be introduced, as well as geometric spaces and their figures. The important point for us, however, is that underlying this method, one finds sets, the theory of sets and first-order logic. Thus, a group is usually defined to be a *set* together with… and here one writes the basic operations and/or relations and their properties defining the concept of group and similarly for the other concepts. This is very familiar to anyone who has learned advanced mathematics.

This global picture of what mathematics is about, namely sets or collections with relations and operations, was developed by mathematicians, logicians and philosophers at the beginning of the 20th century. It was made possible by the simultaneous development of logic itself and its formalization. (See, for more details, Ferreirós 2001.) It is important to understand how this works. Mathematics is done in ordinary language together with specific formalisms or symbolisms that we learn at school. It is a enriched language with specific features. Logicians were able to eliminate, in fact in theory completely, the use of ordinary language and replace it, in principle, by a purely formal language. Not that anyone does that on a daily basis. But it can in principle be done systematically. The real gain is that it provided logicians with the tools to construct maps of fundamental aspects of mathematics itself.

Simplifying greatly, one can say that mathematicians basically do five things: they *define* various concepts, they *state* various conjectures about these concepts, they *prove* assertions about these concepts and they *design* ways to *compute* various formulas associated with these concepts. These activities interact with each other constantly, but nonetheless mathematicians often characterize one another by saying that so and so is a theory builder, the other is great at posing problems, this one is a theorem prover and that one is well-known for devising ways of computing and her computational skills.

First, an analysis of definability and theory building in mathematics is provided by what is called 'model theory'. The name comes from the fact that one defines what it is to be a model, that is more or less a particular instance, of a given theory written in a specific formal language and it studies the properties of these models, as they are defined by that language. (See, for instance, Marker 2002.) Second, an analysis of the notion of logical proof in mathematics was developed and yielded a field now called 'proof theory'. It more or less studies properties of mathematical proofs in certain formal systems, for instance the logical strength required to prove certain mathematical theorems. (See Takeuti 1987.) Finally, an analysis of the notion of computation or algorithm lead to a better understanding of a central phenomenon of mathematical thinking, namely recursion and a whole theory underlying theoretical computer science, namely recursion theory. (See Enderton and Herbert 1977.) All of these, namely logic, set theory, proof theory, model theory and recursion theory constitute what is called *metamathematics*, or the formal study of mathematics and mathematical knowledge itself.[5]

We have to underline that the harvest of results in metamathematics is rich and philosophically significant. The most spectacular results are certainly those that establish *limitations* of certain theories. Gödel's incompleteness theorems are central to foundational studies. The first incompleteness theorem asserts that in any consistent formal theory T in which a sufficiently important part of arithmetic can be done, there are sentences in the language of T such that these sentences can neither be proved nor disproved in T. The second incompleteness theorem, which uses the first in its proof, states that such a theory T, if consistent, cannot prove its own consistency. In particular, the set theory we have been presenting to our extraterrestrial friends falls prey to these theorems. Thus, although that set theory can be taken as a foundation for mathematics, we know that there are statements that cannot be proved nor disproved in it. (See, for a thorough presentation of the theorems (Smith 2013). For a more general discussion about the theorems and their impact, see Franzén 2005.) That is a fact of life, here and everywhere.

Historical Aside

It often happens in mathematics that some idea developed in a certain framework for a certain purpose, i.e. solve a specific problem, can be transported, generalized or adapted to a different framework and yields a rich and unexpected outcome. What we are about to tell belongs to that type of development.

In the early 1940s, two mathematicians, Samuel Eilenberg and Saunders Mac Lane, came up with a new theory, namely category theory, in order to clarify an intriguing situation they had stumbled upon in the context of algebraic topology. Their work resulted in the definition of the notions of category, functor and natural transformation, to which we will turn in the next section. These notions were not conceived with the foundations of mathematics in mind. Their introduction did raise some issues with respect to set theory, however. Indeed, there was some technical glitch with the fact that one could consider the category of all sets and, strictly speaking, such a thing cannot exist. For if a category is ultimately a set, like all mathematical entities in the context of set theory, then the category of all sets would be the same as the set of all sets and the latter cannot be a genuine mathematical entity, as Russell had already shown in the early 20th century with his infamous paradox. Furthermore, as we will see, it immediately dawn on Eilenberg and Mac Lane that one could define the category of categories and although the concept was perfectly

[5] The locus classicus of metamathematics is still (Kleene 1952).

natural, it did not appear to be mathematically useful and it raised issues in the set theoretical framework.

In fact, the foundational issues were not new nor considered a serious problem and Eilenberg and Mac Lane were aware of these facts. After all, it was just Russell's paradox again and one can also consider the well-ordered set of all well-ordered sets as an example of a concept that applies to itself, yielding potential foundational problems. In both cases, set theorists had offered solutions to circumvent these difficulties. Not that the solutions are entirely satisfactory, but the introduction of category theory did not generate new problems in the foundations of mathematics, at least, not when it was created.[6]

The situation changed fifteen years later, around the end of the 1950s when two mathematicians, Daniel Kan and Alexandre Grothendieck, started to use systematically what are called functor categories in their work. The introduction of these constructions required to go slightly beyond, to extend, the known set theory. Thus, the territory of mathematics was transformed, but these changes did not suggest a radical modification of our representation of the foundations of mathematics.

Then, in 1963, a young mathematician named Bill Lawvere came along and launched a radical research program. He proposed three ideas that no one had dare consider before. First, that category theory be an integral part of metamathematics. Second, that set theory, more specifically ZF set theory, be replaced by a set theory developed in the context of category theory. Third and last, that the category of categories be taken as the foundations of mathematics. It should be said that category theory had a metamathematical flavour right from the start and Eilenberg and Mac Lane themselves were aware of it. It is worth quoting them on that point:

> In a metamathematical sense our theory provides general concepts applicable to all branches of abstract mathematics, and so contributes to the current trend towards uniform treatment of different mathematical disciplines. In particular, it provides opportunities for the comparison of constructions and of the isomorphisms occurring in different branches of mathematics: in this way it may occasionally suggest new results by analogy. (Eilenberg and Mac Lane 1945, p. 236)

What Eilenberg and Mac Lane understood is that categorical ideas could be used in all branches of mathematics, thus providing a unifying framework at a certain conceptual level and this was seen as being heuristically useful. Since it provides a way of reorganizing a given mathematical discipline, it is in this sense metamathematical. But it is not metamathematical in its primary, foundational sense.

The latter step was taken explicitly by Lawvere. It was a radical and bold idea. Lawvere showed in his Ph.D. thesis and in subsequent publications that it was not only possible to import category theory in metamathematics, but that it was fruitful and promising in many respects. His work indicated the possibility that the "uniform treatment" suggested by Eilenberg and Mac Lane could also be applied to the foundations of mathematics. It has to be said that between Eilenberg and Mac Lane's publications and Lawvere's work, certain central concepts of category theory were

[6] For the history and philosophy of category theory, see Krömer (2007), Marquis (2009).

discovered and developed, for instance the concepts of adjoint functors and representable functors, and that Lawvere's work relies heavily on these.

The first two proposals turned out to be extraordinarily fruitful and led to the rapid and rich development of what is now called categorical logic. The third one is still under construction, but not only for its relevance in the foundations of mathematics, but also for other purposes, for instance in algebraic topology, algebraic geometry, homological algebra, mathematical physics and theoretical computer science. Let us now turn to these ideas and see how they provide a different map of mathematics.

Categories: The Basic Ideas

From a purely iconic point of view, a category can be thought of as a network of nodes connected by arrows that satisfy some mild conditions. A useful informal interpretation of such a network is to think of an arrow $f : x \to y$ as a process that transforms the node x it starts from into something that is in the node y it ends in. Thus, an arrow is always attached to nodes: the starting node, also called its domain, and its ending node, also called its codomain. For a process is always a transformation of something into something. Note that the starting node can be the same as the ending node: loops are allowed. With this analogy in mind, it is easy to understand the conditions such a network has to satisfy to be a category. Processes should combine or compose and these compositions should ultimately all be "the same". Also, there should be a process that amounts to doing nothing. Thus, every node x has an identity arrow, denoted by 1_x, which can be thought of as the identity transformation: something always transforms into itself. Whenever there is a process f from x to y and a process g from y to z, then this the same as a process $f \circ g$ from x to z.[7] In words, this says that processes compose whenever they are defined. These are part of the data of the network. They have to satisfy two simple conditions. First, the composition of processes has to be associative, that is $f \circ (g \circ h) = (f \circ g) \circ h$. Second, the identities have to act as neutral elements with respect to the composition of processes: $f \circ 1_x = f$ and $1_y \circ g = g$. The formal definition amounts to writing these data and conditions into the appropriate mathematical language. Here is how it is often done.[8]

Definition 1 A *category* C is given by a collection $Ob(C)$ of objects together with, for $x, y \in C$ a set $Hom(x, y)$, called the *morphisms* of x into y of C and for three objects x, y, z of C, an operation, called the composition of morphisms,

$$Hom(x, y) \times Hom(y, z) \to Hom(x, z),$$

[7] We are not reversing the order of composition of f and g, as it is usually done in textbooks. We are doing it on purpose.

[8] For those who would like to know more about category theory, the standard reference is still (Mac Lane 1998). Two slightly different takes on the theory can be found in Leinster (2014), Riehl (2016).

which takes morphisms $f : x \to y$ and $g : y \to z$ and yield a morphism $f \circ g : x \to z$, which satisfy the following two conditions:

(1) Composition is associative: $f \circ (g \circ h) = (f \circ g) \circ h$;
(2) For x in C, there is a morphism 1_x in $Hom(x, x)$ such that $f \circ 1_x = f$ and $1_y \circ f = f$, for $f : x \to y$.

Since we have said that a category can be pictured as a collection of nodes and arrows between these nodes, it might be worthwhile to represent the two conditions by images. In a picture, composition of morphisms is represented by the idea that two paths with the same starting node and the same ending node can be considered to be the same. In category theory, one says that a diagram commutes to indicate that there is an equality between paths. Thus, for instance, one writes the composition of f and g in the following way:

$$\begin{tikzcd}
 & x \arrow[dl, "f"'] \arrow[dr, "f \circ g"] & \\
y \arrow[rr, "g"'] & & z
\end{tikzcd}$$

and says that the diagram commutes. Getting back to the two conditions in the definition, we can draw the first one as follows:

$$\begin{tikzcd}
 & w \arrow[ddl, "f"'] \arrow[d, "f \circ g"] \arrow[ddr, "(f \circ g) \circ h = f \circ (g \circ h)"] & \\
 & y \arrow[dr, "h"] & \\
x \arrow[ur, "g"] \arrow[rr, "g \circ h"'] & & z
\end{tikzcd}$$

We could have written two arrows on the right and indicate that they are equal. But it is not necessary.

And the second one can be pictured thus:

$$\begin{tikzcd}
 & x \arrow[dr, "f"] & \\
x \arrow[ur, "1_x"] \arrow[rr, "f"] \arrow[dr, "f"'] & & y \\
 & y \arrow[ur, "1_y"'] &
\end{tikzcd}$$

Thus, the equations in the definition correspond to diagrams. In practice, a lot of category theory amounts to setting up the right diagram and verifying that it commutes.

As we have indicated, categories pervades contemporary mathematics. Here is a (very short) list of examples of categories.

Example 1: The category **Set** whose objects are the usual sets and the morphisms are the usual functions $f : X \to Y$ between them.
Example 2: The category **Grp** of groups and group homomorphisms between them.
Example 3: The category **Top** of topological spaces and continuous functions between them.

Needless to say, we could extend this list considerably. The first important remark that has to be made is that (almost) all concepts in mathematics form a category with the suitable notion of morphism between them. Indeed, category theory came with the realization that mathematical notions come equipped with morphisms, or transformations, and that these have to be taken as part of the concept itself. For instance, groups come with group homomorphisms, that is a function that preserves the group structure. Similarly, topological spaces come with continuous functions between them and so on and so forth. This is what Eilenberg and Mac Lane had in mind when they wrote the foregoing passage.

And, indeed, categories themselves come with their own notion of morphism. These are called *functors*.

Definition 2 A *functor* $F : \mathcal{C} \to \mathcal{D}$ between categories $\mathcal{C}$ and $\mathcal{D}$ is given by the following data:

(1) For each $X \in Ob(\mathcal{C})$, an object $F(X) \in \mathcal{D}$;
(2) For each morphism $f : X \to Y$ in $\mathcal{C}$, a morphism $F(f) : F(X) \to F(Y)$ in $\mathcal{D}$;

These data have to satisfy the following conditions:

(1) For each pair of morphisms $f : X \to Y$, $g : Y \to Z$, $F(f \circ g) = F(f) \circ F(g)$, that is functors preserve composition of morphisms;
(2) For each $X \in Ob(\mathcal{C})$, $F(1_X) = 1_{F(X)}$, that is, functors preserve identities.

Functors are “translations” or “transformations” between categories. They play a key role in many branches of mathematics. For instance, algebraic topology study spaces by algebraic means. The basic strategy consists in translating spatial data into algebraic data. Thus, it is possible to associate to a topological space various groups by performing certain constructions on the given space. Each and every one of these translations is, in fact, a functor from the category of topological spaces into an appropriate category of algebraic objects, e.g. groups, abelian groups, modules, etc.

A moment’s thought and a routine verification suffices to realize that the collection of categories together with functors *is* a category! It is easy to define identity functors and the composition of functors is also immediate. It remains to verify that the conditions defining a category are satisfied. This is the category of categories. Thus, we have a universe of mathematics made up of categories and functors between them. It is a new network, a network of mathematical concepts. Within the standard set theoretical framework, the category of categories cannot be dealt with directly. It is "too big". There *are* ways of dealing with it, by tweaking the usual set theory appropriately, but they introduce extraneous considerations that seem somewhat artificial. We will get back to this issue in section "HD-Categories".

As we have said, Lawvere suggested that the standard set theory based on the relationship '$a \in A$' be replaced by the basic operation of category theory, namely the composition of morphisms. At first glance, this seems to be simple enough: sets and functions form a category after all. It is the simplest and probably the most natural example of a category. But, life is rarely easy and a considerable amount of work has to be carried out before we can say anything else. The challenge here is twofold. First, it is certainly not enough to declare that there is a category of sets. One has to be able to *do* set theory in this framework and one has to show that whatever it is that we do with set theory, it can be done in this new, different theory. This means that the language of category theory must make it possible to express the concepts that are considered essential to any set theory and that it allows us to do what we want a set theory to do in mathematics. Does the language of category theory allows that? This is a priori very difficult to assess. In fact, there is only one way to find out: it has to be carried out. We now know that it *is* possible. Second, it should be possible to use the language of category theory to characterize those categories that are to be taken as categories *of sets*. Lawvere originally thought that he could define a "unique" category of sets.[9] He then found out, some years later, that there was much more to categories of sets than he first thought. To understand that, we must say a few words about the language inherent to categories, namely the language of arrows and how that language can be used in mathematics.

The Language of Arrows

If a category can be thought of as a kind of network, it should be possible to use that network, to talk about that network, to know its objects and their properties. And it is indeed possible. We will illustrate how this works with a few simple examples. It will give the reader a glimpse of the language of arrows used in category theory.

We have to try to think about an arbitrary abstract network and try to see what kind of properties it could have and they could be characterized. A simple case is as follows. It is possible that in such a network, there is an object, call it 1, such that for all objects of the network, there is exactly one arrow into it. In other words, this

[9] There are important conceptual and technical issues involved here. One has to clarify what it means for a category to be unique and that question, far from being trivial, raises with it a host of interesting mathematical and philosophical considerations that we unfortunately have to ignore in such a short paper.

object is where are the processes end or terminate. We would therefore call such an object a *terminal* object. Once we have thought about a terminal object, it is easy to conceive of its mirror image: an *initial* object: an object of the network, call it 0, such that, for all objects of the network, there is exactly one arrow from it. In a sense, this is where everything originates in the network. Notice something crucial here: there may be more than one terminal object in a network. This might be surprising and it is possible only if a specific property of the network is satisfied: all such terminal objects have to be "the same", that is we have to have an internal criterion of identity that allows us to say that they are the same. The language of arrows provides us with such a criterion of identity. Here it is.

Definition 3 A morphism $f : x \to y$ in a category C is called an *isomorphism* between x and y if there is a morphism $g : y \to x$ such that $f \circ g = 1_y$ and $g \circ f = 1_x$. When such a morphism exists, x and y are said to be *isomorphic*.

This is an important notion in mathematics in general and it can be expressed directly in the language of categories. In our case, one has now to prove that all terminal objects in a category C are isomorphic and it can be proved. Moreover, there is a unique isomorphism between terminal objects. The same claims hold for initial objects (almost automatically).

The other nodes and arrows of the network that are worth noting require more familiarity with the possible kind of networks we are talking about. They are nonetheless easy to explain and understand. It requires a bit more time to understand how to *work* with them. What we can notice is that there are other nodes that are similar to terminal objects and initial objects with the difference that, in these cases, they are terminal (or initial) relative to certain "forms" in the network. Here is an example. Suppose we take the following fork in our network:

$$\begin{tikzcd} x & p_{x,y} \arrow[l] \arrow[r] & y \end{tikzcd}$$

The names x, y and $p_{x,y}$ play a role here and the notation $p_{x,y}$ is used to indicate that this object depends on x and y somehow. It may very well be that one of these forms, with x and y at each end of the fork and if it exists, acts like a terminal object with respect to the forms with x and y at each end of the fork. This means, in the language of arrows, that for all form $x \leftarrow q_{x,y} \to y$, there is a unique arrow $q_{x,y} \to p_{x,y}$ such that the following diagram commutes:

$$\begin{tikzcd} & q_{x,y} \arrow[dl] \arrow[d] \arrow[dr] & \\ x & p_{x,y} \arrow[l] \arrow[r] & y \end{tikzcd}$$

Again, any two such nodes satisfying this condition for x, y are isomorphic. Such an object $p_{x,y}$ *together with the arrows* $p_{x,y} \to x$ and $p_{x,y} \to y$ is called a *product* of x and y.

Two important remarks have to be made about these concepts. First, in all three cases, the expression "for all … such that…" appears in the description of these objects and arrows. Because of the "for all" in there, these objects and arrows are said to satisfy a universal property and they are characterized by that property. This way of defining objects in a category is central. Second, we have talked about hypothetical networks and identified some of their possible salient features. We haven't said anything about *existence*. It has to be verified, in specific instances, for given categories, whether these objects and arrows exist in these cases. For example, the category of sets does have terminal objects: any singleton set will do and notice that they are all isomorphic. It also has initial objects, but in this case it is unique in the set theoretical sense of being unique: it is the empty set. It also has products and they are known as cartesian products in the context of set theory. But they could be very different in different categories and they might not exist in certain categories.

This brings us to one last point about that language. It can be used and has been used to define certain kind of categories. This is something that Eilenberg and Mac Lane did not foresee when they came up with these notions. Then mathematicians realized that they could stipulate, in that language, certain properties that a category ought to have to be able to develop certain types of mathematics. For instance, one can say that a category C that has all finite products is *cartesian*. This method of working turns out to be extremely powerful. In this way, it is possible to define in an abstract manner fields of mathematics. Thus, homological algebra can be done in certain, abstractly given, categories. The same is true of homotopy theory. This gives us a different map of mathematics. Not only does category theory organizes types of structured sets, like in the foregoing examples, but it also cut mathematics at its methodological joints, so to speak. And, as it turns out, one of these joints *is* set theory.

Categories of Sets

If we were to go back to our extraterrestrial friends and try to explain to them this new way of thinking about sets, we would have to modify our language somewhat. It it important to note that we would start by talking informally about collections in the same way as before. It is only when we move to the more rigorous, formal theory that we change our discourse. We would not take the relation $a \in A$ as the primitive relation. We *declare* that an object a is of type A and we write $a : A$. This is not a statement, however. It cannot be true or false. It is a declaration. We start with a set A and we exhibit one of its members. We need one more relation. Each set A has a built-in identity relation that allows us to determine whether $a = b$ or $a \neq b$. There is no identity relation between *sets* themselves. We do *not* assume the axiom of extensionality for sets. What we have at our disposal is the language of arrows between sets and the latter allows us, as we have seen above, to say when two sets are *isomorphic*. This is now our identity relation or structure for sets. But we can be more specific and give axioms that characterize these abstract sets. Again, as with ZF set theory, we will not give all the axioms in the formal language. We will explain some of them informally. (We roughly follow Leinster (2014) here.)

Our first axiom is that sets and arrows, which are just the standard functions in this case, form a category. The second axiom stipulates that there is a set with no element and the third that there is a set with exactly one element. Our extraterrestrial friends might be surprised to hear us talk about elements. But there is an easy way to introduce the expression in the language of arrows and it is especially simple in the case of sets. It is easy to verify that a terminal object 1 in the category of sets is any singleton set $\{\star\}$. Indeed, take any set A, then there is exactly one function $A \to \{\star\}$. Clearly, any two singleton sets are isomorphic and there is a unique such isomorphism. Now, consider a function $\{\star\} \to A$. Such a function simply picks an element of A. We could in fact write $a : \{\star\} \to A$ to denote it.[10] There is thus a bijection between the (familiar) elements $a \in A$ and arrows $a : \{\star\} \to A$. Thus, the foregoing axioms say that the category of sets has both initial and terminal objects. Another axiom says that the product $A \times B$ of two sets A and B can be formed in the category, while the next one stipulates that the collection of all functions $f : A \to B$ between two sets A and B is a set of the category. (This last condition can be expressed in the language of arrows.) We also add that the natural numbers form a set. The remaining axioms are slightly more technical and we will refrain from trying to explain them in such a short paper. This, in a nutshell, is the set theory we take as being adequate for all ordinary mathematics, at least in the sense that ZF is.

There is a surprise that awaits us, however. Indeed, this is but one set theory! In fact, in our new framework, we have an infinite number of set theories. Indeed, as we have mentioned above, it is possible to use the language of arrows to characterize a type of mathematics. What we have just done is to give the axioms for abstract sets, which, in a precise sense, is the counterpart in the language of arrows of the standard ZF theory. We can do more. We can restrict the axioms above to what is now known as *elementary toposes*. This yields an extraordinary powerful theory with many different facets. It has, at the same time, depending on how one looks at it, rich geometric, algebraic and logical contents. The axioms of an elementary topos are extraordinarily simple. An elementary topos is a category, of course, that satisfies two simple conditions. The first condition stipulates that an elementary topos has a terminal object and that, for any two of its objects X and Y, their product $X \times Y$ exists. The second condition requires that for any object X of the topos, the so-called *power object*, denoted $\mathcal{P}(X)$, exists. That is it.[11]

So, what is mathematics about in this map? Topos can be said to be about sets. However, that answer is not entirely satisfactory, since toposes can also be said to be

[10] Notice that we can now do that in any category that has a terminal object.

[11] We are cheating here. For we do have to include the arrows that come with these constructions. They are an integral part of the definition. For more on topos theory, see Goldblatt (1984) or, for a more advanced treatment (Mac Lane and Moerdijk 1994).

about spaces. In fact, they combine in a unique way the discrete and the continuous. Moreover, these toposes, being categories, are related to one another by functors and it is relevant to mention that there are two important types of functors between toposes: there are geometric functors and logical functors. Thus, toposes extend the universe of mathematics and metamathematics. The territory has changed. Whereas our original answer was simple and clear, it seems to be somewhat more complicated now. And in a sense, it is. It is more complicated in the same way that 3-dimensional geometry is more complicated than plane geometry. We are adding depth to our picture. How is it? Well, we are in fact in a universe of categories and when categories are related to one another, a richer structure emerges.

HD-Categories

As we have seen above, categories form a network themselves, the links being provided by functors. But categories and functors were created by Eilenberg and Mac Lane in order to define a third concept, which was the focus of their work, namely the notion of natural transformation.

Definition 4 Let $F, G : \mathcal{C} \to \mathcal{D}$ be functors. A *natural transformation* $t : F \to G$ between F and G is a family of arrows $t_X : F(X) \to G(X)$, for all objects X of $\mathcal{C}$, such that for all arrows $f : X \to Y$ of $\mathcal{C}$, the following square commutes

$$\begin{tikzcd}
F(X) \arrow[r, "t_X"] \arrow[d, "F(f)"'] & G(X) \arrow[d, "G(f)"] \\
F(Y) \arrow[r, "t_Y"'] & G(Y)
\end{tikzcd}$$

Thus, there are arrows between functors! One can think of natural transformations in many different ways, but the following two are heuristically useful. First, if one think of a functor as transforming the category $\mathcal{C}$ into the category $\mathcal{D}$ in a systematic fashion, that is by preserving the structure of $\mathcal{C}$, then a natural transformation between such functors is a "translation" of one functor into the other in the category $\mathcal{D}$. Second, if functors are thought of as processes, then a natural transformation is a process that transforms one process into another process, that is, it is a process between processes.

The emerging structure results from the fact that natural transformations compose in two different ways: they compose vertically and horizontally and these compositions interact. Instead of giving the formal definitions, we will simply illustrate them. These illustrations also explain the choice of terminology.

A given natural transformation $t : F \to G$ can be depicted thus:

$$\begin{tikzcd}
\bullet \arrow[r, bend left=50, "F", ""{name=U, below}] \arrow[r, bend right=50, "G"', ""{name=D}] & \bullet
\arrow[Rightarrow, from=U, to=D]
\end{tikzcd}$$

This representation is suggestive: points, which represent the objects, are 0-dimensional, lines are 1-dimensional and thus, we should think of a double line as being 2-dimensional, as showing how to stretch a surface between two lines. Notice that we can draw more than one natural transformation between two functors. This representation is in fact occuring in 3-dimensional space. This become even more clear in the next picture:

$$\begin{tikzcd}
\bullet \arrow[r, bend left=70, "F", ""{name=U, below}] \arrow[r, "G" description, ""{name=M, above}, ""{name=M2, below}] \arrow[r, bend right=70, "H"', ""{name=D}] & \bullet
\arrow[Rightarrow, from=U, to=M]
\arrow[Rightarrow, from=M2, to=D]
\arrow[Rightarrow, from=U, to=D, bend left=40]
\end{tikzcd}$$

Here we have two natural transformations $t : F \Rightarrow G$ and $t' : G \Rightarrow H$ that compose vertically $t' \circ t : F \Rightarrow H$. Functors also compose. Thus, given $F : \mathcal{C} \to \mathcal{D}$ and $F' : \mathcal{D} \to \mathcal{E}$, we can compose them to get a functor $F' \circ F : \mathcal{C} \to \mathcal{E}$. Similarly, we can have parallel functors $G' \circ G : \mathcal{C} \to \mathcal{D}$ and natural transformations $t : F \Rightarrow G$ and $t' : F' \Rightarrow G$:

$$\begin{tikzcd}
\bullet \arrow[r, bend left=50, "F", ""{name=U, below}] \arrow[r, bend right=50, "G"', ""{name=D}] & \bullet \arrow[r, bend left=50, "F'", ""{name=U2, below}] \arrow[r, bend right=50, "G'"', ""{name=D2}] & \bullet
\arrow[Rightarrow, from=U, to=D]
\arrow[Rightarrow, from=U2, to=D2]
\end{tikzcd}$$

The natural transformations t and t' can now be composed horizontally to yield a natural transformation $t' \star t : F' \circ F \Rightarrow G' \circ G$:

$$\begin{tikzcd}
\bullet \arrow[r, bend left=50, "F", ""{name=U, below}] \arrow[r, bend right=50, "G"', ""{name=D}] \arrow[rr, bend left=70, "F' \circ F", ""{name=UU, below}] \arrow[rr, bend right=70, "G' \circ G"', ""{name=DD}] & \bullet \arrow[r, bend left=50, "F'", ""{name=U2, below}] \arrow[r, bend right=50, "G'"', ""{name=D2}] & \bullet
\arrow[Rightarrow, from=U, to=D]
\arrow[Rightarrow, from=U2, to=D2]
\arrow[Rightarrow, from=UU, to=DD, bend left=20]
\end{tikzcd}$$

Of crouse, we haven't defined these compositions properly, but it can be done. Furthermore, these two compositions satisfy certain equations, for instance they are associative, and interact in the form of another equation, called the interchange law.

The foregoing data can be collected and organized in an abstract presentation. Indeed, we start with 0-cells, represented by points, between them we have 1-cells, represented by directed lines, with an operation of composition satisfying certain conditions, between 1-cells, there are 2-cells, represented by double arrows denoting streched surfaces, with two operations of composition satisfying certain conditions. This is the system of categories and it is called a 2-category. In fact, there is another surprise here. There are two, at first sight, different definitions of 2-categories: the so-called strict 2-categories and the so-called weak 2-categories, the latter also known as bicategories. We will not give the definitions of these notions here. They are simply too long. Suffice to say that the difference between the strict and the weak 2-categories rest upon the equations the composition of natural transformations satisfy in the case of the strict 2-categories and do not satisfy in the weak ones. For instance, in a strict 2-category, the horizontal composition is associative, that is $t'' \star (t' \star t) = (t'' \star t') \star t$, whereas in a weak 2-category, one has a natural transformation, called the associator, $\alpha : t'' \star (t' \star t) \Rrightarrow (t'' \star t') \star t$ and it has to satisfy certain equations.

Weak 2-categories, like categories, are linked by certain 2-functors that preserve the structure of weak 2-categories: they are called pseudofunctors in the literature. And pseudofunctors have natural transformations between them, called lax-natural transformations. As it is to be expected, there is a new type of morphism emerging, they are the 3-cells, that is arrows between 2-cells. When these data are put together with the appropriate conditions, they form a weak 3-category, also known as a tricategory. At this stage, we can do an induction, although doing it rigorously is another matter.

The global picture of what one should end up with is clear. The abstract picture is as follows. At the bottom, one finds the 0-cells. Between them, the 1-cells with an operation of composition satisfying certain properties. Next, between the 1-cells, the 2-cells, now with two operations of composition satisfying certain conditions. We can continue like that up to the n-cells with operations satisfying certain conditions. Collecting all these, one gets an $(n + 1)$-category. We can stop at any n or, if one wants to, one can let it go all the way to obtain an ω-category, that is a system of n-categories, for all n. In the same way that in geometry, we talk about higher-dimensional spaces, these concepts refer to higher-dimensional categories, or HD-categories. This is the informal picture. It gives us a map of abstract mathematics and its organization. Thus, abstract mathematical concepts organize themselves in a complex system that reflects, and this is an important point, a geometric progression of sort. For this structure includes the intrinsic structure of what are called "homotopy types" which are, basically, the fundamental forms of space. They are to geometry what natural numbers are to arithmetic. Hence, this universe contains in its construction both the natural numbers, when we move form one level to another, together with the geometric basic blocks at each dimension. It is worth repeating it: this is the informal picture. Do we have the mathematics that goes with it?

Well, yes and no. There are, at present, many different definitions of n-categories. (See, for instance, Leinster 2002.) It should be mentioned that these definitions give directly, so to speak, the notion of an n-category, for an arbitrary n. It does not proceed by constructing step-by-step each n-category in the way we have suggested above. This is, in fact, quite a feat. But this richness generates its own problem. Which definition is the right one? How do they differ? Do they, in fact, differ? The latter question turns out to be mathematically quite challenging. Mathematicians have to devise a way to compare these definitions and it is far from obvious how one should proceed. A simple thought experiment should allow anyone to understand the nature of the challenge. Suppose we are given two apparently different definitions of n-categories, call them $nCat_1$ and $nCat_2$. The system of $nCat_1$ form an $(n + 1) - category$ and so does the system of $nCat_2$. But which definition of $(n + 1)$-category should we use? We need to pick one. For it is in the environment of $(n + 1)$-categories that we will find a notion of n-equivalence between n-categories that we need to determine when two n-categories are n-equivalent. Some solutions to solve that problem have been proposed and some equivalences have been proved. But we stil do not have *the* definition of n-category or ω-category at our disposal.

Categories, HD-Categories and the Foundations of Mathematics

It is time to wrap up, but not quite to conclude yet. Two important points have to be made. First, in the same way that the advent of category theory in mathematics has allowed for a vast generalization and abstraction of the discipline as a whole, its advent in the metamathematical domain has also allowed a vast generalization and abstraction as a whole.

Category theory *in mathematics* has lead to deep unifications of various fields, a clarification of various results and theories, the creation of new theories and domains. Its concepts have led to a better understanding of important conceptual aspects of various phenomena and theories. Although we haven't mentioned it here, the concept of adjoint functors sheds a new light on mathematical disciplines, theories and results. To use the terminology of this volume, it has lead to a vast expansion of the mathematical territory. At the same time, it also lead to new connections between some of its parts and a better understanding of the organisation of the whole territory.

Category theory *in metamathematics* has also lead to important generalizations and abstractions of the various parts of the field. We cannot give even a partial picture of these changes. We can perhaps capture the core of the modification by saying that categorical logic adds an algebraic stratum to the usual description of logic and mathematics. Thus, in proof theory, in model theory and in recursion theory, the categorical standpoint puts at the center of these fields algebraic constructions, namely categories, functors and natural transformations in various guises, that allow for a more abstract description and development of these pictures of mathematical knowledge.

Second, HD-categories does provide, at least in principle, a foundational framework for mathematics. The complete logical and mathematical details have still to be developed. In that respect, we have to point out that there is a new and beautiful foundational approach that has seen the day in the last ten years or so which has, at least, a very clear syntax and some models. That framework is called *homotopy type theory*. It uses categories in an essential manner for its semantics and develops an important portion of the higher-dimensional universe. (See Awodey 2014, 2015.) However, it does not cover it all, at least, not yet. In both cases, that is Homotopy

Type Theory and HD-categories, the new universe is tailored for geometrical needs whose foundations seem to be tangential to what the standard set theory offers.

Conclusion

Our extraterrestrial friends have been carefully listening our description of HD-categories and how it gives a map of the universe of mathematics. Needless to say, they come back with the same question: so, what is mathematics *now* about? We can answer again: mathematics is about *abstract structures* or *abstract forms*. Our new answer requires that we explain how two concepts can be the same *and* different at the same time, for this is what we mean by being abstract in this case. We also have to pay particular attention to the language we use in mathematics to talk about these concepts. That language has very specific properties, it has a special grammar that allows us to say certain things and prevent us to say other things, but in such a way that it is tailored specifically for abstract structures.[12] We can explain to our friends how these rest upon a theory of identity that is based on the notions of isomorphism and equivalence and how these notions define what it is to be abstract and a structure. In other words, we develop a map, a map composed of a language, a logic and a universe of interpretation for this language and this logic. The rest consists in directives to interpret the map properly. Of course, a map is different from what it maps. However, in this case, it might be sometimes difficult to say whether we are looking at the map or looking at what it represents.

References

S. Awodey, Structuralism, invariance, and univalence. Philos. Math. **22**(1), 1–11 (2014)

S. Awodey, *Homotopy type theory, in Logic and Its Applications*, vol. 8923, Lecture Notes in Computer Science (Springer, Heidelberg, 2015), pp. 1–10

S. Eilenberg, S. Mac, Lane, A general theory of natural equivalences. Trans. Am. Math. Soc. **58**, 231–294 (1945)

H.B. Enderton et al., Recursion theory, in *Handbook of Mathematical Logic, Part C*. Studies in Logic and the Foundations of Mathematics, vol. 90 (North-Holland, Amsterdam, 1977), pp. 525–815

J. Ferreirós, The road to modern logic-an interpretation. Bull. Symb. Log. **7**(4), 441–484 (2001)

J. Ferreirós, *A history of set theory and its role in modern mathematics, in Labyrinth of Thought*, 2nd edn. (Birkhäuser Verlag, Basel, 2007)

T. Franzén, *An incomplete guide to its use and abuse, in Gödel's Theorem* (A K Peters Ltd, Wellesley, MA, 2005)

R. Goldblatt, The categorial analysis of logic, in *Topoi*, 2nd edn., volume 98 of Studies in Logic and the Foundations of Mathematics (North-Holland Publishing Co., Amsterdam, 1984)

J. Hintikka, Logicism, in *Philosophy of Mathematics*, volume 4 of Handbook of the Philosophy of Science (Elsevier/North-Holland, Amsterdam, 2009), pp. 271–290

[12] We are being very sketchy. See, for more on this perspective (Makkai 1998, 2014; Marquis 2013).

T. Jech, *The third millennium edition, revised and expanded, in Set Theory* (Springer Monographs in Mathematics (Springer, Berlin, 2003)

S.C. Kleene, *Introduction to Metamathematics* (D. Van Nostrand Co., Inc, New York, N. Y., 1952)

R. Krömer, *Tool and Object. A History and Philosophy of Category Theory*, volume 32 of Science Networks. Historical Studies (Birkhäuser Verlag, Basel, 2007)

T. Leinster, A survey of definitions of *n*-category. Theory Appl. Categ. **10**, 1–70 (electronic) (2002)

T. Leinster, *Basic Category Theory*, vol. 143 (Cambridge Studies in Advanced Mathematics (Cambridge University Press, Cambridge, 2014)

T. Leinster, Rethinking set theory. Am. Math. Monthly **121**(5), 403–415 (2014)

S. Mac Lane, *Categories for the Working Mathematician*, 2nd edn., volume 5 of Graduate Texts in Mathematics (Springer, New York, 1998)

S. Mac Lane, I. Moerdijk, A first introduction to topos theory, in *Sheaves in Geometry and Logic*. Universitext, Corrected reprint of the 1992 edition (Springer, New York, 1994)

P. Maddy, How applied mathematics became pure. Rev. Symb. Log. **1**(1), 16–41 (2008)

P. Maddy, *Defending the Axioms: On the Philosophical Foundations of Set Theory* (Oxford University Press, Oxford, 2011a)

P. Maddy, Set theory as a foundation, in *Foundational Theories of Classical and Constructive Mathematics*, volume 76 of The Western Ontario Series in Philosophy of Science (Springer, Dordrecht, 2011b), pp. 85–96

M. Makkai, *Towards a Categorical Foundation of Mathematics, Logic Colloquium '95 (Haifa)*, vol. 11, Lecture Notes Logic (Springer, Berlin, 1998), pp. 153–190

M. Makkai, The theory of abstract sets based on first-order logic with dependent types, Oct 2014

D. Marker, An introduction, in *Model Theory*, volume 217 of Graduate Texts in Mathematics (Springer, New York, 2002)

J.-P. Marquis, *From a Geometric Point of View: A Study in the History and Philosophy of Category Theory*, volume 14 of Logic, Epistemology, and the Unity of Science (Springer, 2009)

J.-P. Marquis, Categorical foundations of mathematics or how to provide foundations for abstract mathematics. Rev. Symb. Log. **6**(1), 51–75 (2013)

J.-P. Marquis, Mathematical abstraction, conceptual variation and identity, in *Logic, Methodology and Philosophy of Science, Proceedings of the Fourteenth International Congress*, ed. by P. Schroeder-Heister, G. Heinzmann, W. Hodges, P. Edouard Bour (Forthcoming) (College Publications, London, 2015), pp. 299–322

J.-P. Marquis, Stairway to heaven: the abstract method and levels of abstraction in mathematics. Math. Intell. **38**(3), 41–51 (2016)

E. Riehl, *Category Theory in Context* (Dover Publications, New York, 2016)

P. Simons, Formalism, in *Philosophy of Mathematics*, volume 4 of Handbook of the Philosophy of Science (Elsevier/North-Holland, Amsterdam, 2009), pp. 291–310

H. Sinaceur, Différents aspects du formalisme, in *Le formalisme en question* (Saint-Malo, 1994), Probl. Controv. (Vrin, Paris, 1998), pp. 129–146

P. Smith, *An Introduction to Gödel's Theorems*, 2nd edn. (Cambridge Introductions to Philosophy (Cambridge University Press, Cambridge, 2013)

J. Stedall, A very short introduction, in *The History of Mathematics*, volume 305 of Very Short Introductions (Oxford University Press, Oxford, 2012)

G. Takeuti, *Proof Theory*, 2nd edn., volume 81 of Studies in Logic and the Foundations of Mathematics, With an appendix containing contributions by G. Kreisel, W. Pohlers, S.G. Simpson, S. Feferman (North-Holland Publishing Co., Amsterdam, 1987)

D. van Dalen, The development of Brouwer's intuitionism, in *Proof Theory* (Roskilde, 1997), volume 292 of Synthese Library (Kluwer Academic Publishers, Dordrecht, 2000), pp. 117–152

D. van Dalen, M. van Atten, Intuitionism, in *A Companion to Philosophical Logic*, volume 22 of Blackwell Companions Philosophy (Blackwell, Malden, MA, 2002), pp. 513–530

Chapter 20
Reconciling the Realist/Anti Realist Dichotomy in the Philosophy of Mathematics

Bharath Sriraman and Per Haavold

Introduction

Mathematical philosophy typically occurs in the background of mathematics. In the vast territory that characterizes modern mathematics, positions in the philosophy of mathematics can be viewed as a map or a guide through which one can understand some of its terrain. In classical mathematical philosophy there are four positions, namely Platonism, formalism,[1] logicism, and Intuitionism (or Constructionism). Each of these positions has been expounded on at length in the literature by philosophers like Reuben Hersh, Michele Friend, Penelope Maddy, among others. Platonism is also referred to as Realism and Intuitionism (or Constructionism) is referred to as Anti-Realism.[2] These two positions as their labels suggest are dichotomous with Realism conferring ontological status to mathematical objects whereas anti-Realism emphasizes epistemology in the sense that methods of construction are necessary to construct mathematical objects. More specifically there are different conceptions for the establishment of truth in these two positions.

[1] We deliberately rule out formalism for the primary reason that in keeping with Heyting's (1974) observation: "There is no conflict between intuitionism and formalism when each keeps to its own subject, intuitionism to mental constructions, formalism to the construction of a formal system, motivated by its internal beauty or by its utility for science and industry. They clash when formalists contend that their systems express mathematical thought. Intuitionists make two objections against this contention. In the first place, …[m]ental constructions cannot be rendered exactly by means of language; secondly the usual interpretation of the formal system is untenable as a mental construction." (p. 89).

[2] In this chapter we use the terms Realism and Constructionism for these two positions.

B. Sriraman (✉)
University of Montana, Missoula, MT 59812-0864, USA
e-mail: sriramanb@mso.umt.edu

P. Haavold
University of Tromso, Tromsø, Norway

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_20

For a realist, a proof by contradiction is sufficient to confer an irrational status to say $\sqrt{2}$, but for an anti-realist it is more important to know how to construct $\sqrt{2}$ or any other number for that matter! To paraphrase L.E.J. Brouwer, the founder and proponent of Constructionism, one does not ask a statement is true unless they know what it means (Bishop 1973). And further the methods used to construct an object or prove a theorem should not rely on "logical tricks" such as the law of the excluded middle. Richman (1999) illustrates this in the in direct proof of "There is a digit that appears infinitely often in the decimal expansion of π". The proof explained by Richman does not give any method for constructing these digits but merely confers an "existence status" to objects. Similarly there are other interesting and even absurd things that can proved using the Realist's criteria of an existence proof, without really knowing how to go about constructing these objects. This is the crux of the Realism-conferring status to objects without knowing what they are in the sense of being able to construct them without using the rule of the excluded middle. In other words, if a Realist proves "$\exists$O", the Constructionist would answer you have established "$\neg\forall \times \neg$O" or if the Realist proves "A $\vee$ B", the Constructionist would answer you have proved "$\neg[\neg$A $\wedge$ $\neg$B]"

The territory of mathematics particularly that found in textbooks relies on such proofs to establish results for undergraduate students. The question then is what (if any) are the benefits of using constructive methods. Further from a pluralist standpoint as expounded by Friend (2014), can one possibly hold both a realist and an anti-realist stance for particular objects or results? Better yet, in the exercise of "constructing the real numbers" (pun intended), an exercise which terminates in a real analysis course for some students, and an advanced geometry or abstract algebra course for others, can one highlight issues that arise in the philosophy of mathematics, particularly the realist and anti-realist stance to developing this mathematical object. In doing so, the territory of what constitutes a real number is illuminated by the map of developing particular constructions, especially notions of rationals and irrationals, and the subtleties of these objects. Can the seemingly dichotomous position of the realists and anti-realists find "points of convergence" (no pun intended), or can different ways to construct a particular number shed more insights for a student, and a pluralist view is thus possible? Another necessity to examine this approach is the fact that mathematical theories are constantly in a state of flux as evident in the development of non-Euclidean geometry, the paradoxes of set theory, and the development of special relativity with Minkowski's space-time metric as opposed to the older theory of Lorentz that used Newton's notions of space-time. Arguably bringing in examples from physics or examples from the physical world may be challenged by both realists and anti-realists as not being real mathematics. In the remainder of this chapter we will focus exclusively on mathematics.

There are different views of constructive mathematics (Bridges and Richman 1987; Raatikainen 2004) which suggest that old mathematical concepts need to be relearned and this is a non-trivial task, hence the recommendation to begin with younger students of mathematics. Schechter (2001) points to seemingly trivial notions that many take for granted such as inequality and apartness of real numbers also need to be carefully distinguished keeping with Brouwer's suggestion to constructionists that meaningful distinctions need to be maintained. One of the classical notions in analysis is that of an infimum of a set S of real numbers. Schechter writes:

> Suppose S is a set of real numbers, and r is a real number. To show constructively that $r = \inf(S)$, we must prove that $r \leq s$ for every $s \in S$, and we must also construct numbers $s_1, s_2, s_3, \ldots \in S$ satisfying $r > s_k - 1/k$. It is not enough merely to show the classical "existence" of some s_k's with that property.

The constructionist aspect suggests that merely having an algorithm is sufficient to meet the demands of constructionist mathematics. But Bishop (1967) never really explained what constitutes an algorithm for it to meet the burden of being constructionist. This leaves a very large grey area where algorithmic mathematics can be argued as being constructionist mathematics, a view which is corroborated by Richman (1999). However there is some clarification for what these grey areas might be. According to Mandelkern (1989), Errett Bishop said the following to explain what constructive mathematics is:

> How do you know whether a proof is constructive? Try to write a computer program. If you can program a computer to do it, it should be constructive. Notice I said write the program. Don't necessarily run it on the computer and wait around for the result.

In the 21st century, we have the advantage of retrospective on these words because of the huge program of experimental mathematics established by the Borweins, which not only involved writing a computer program but actually running it to ends never thought possible by Bishop.

Exploring the Grey Areas: Constructing the Real Numbers

The real numbers can be constructed in numerous ways. Typically one begins with the construction of Q, the set of rational numbers, which is an ordered field but not complete. For completeness considerations one has to venture into constructions that are too technical to discuss in this chapter. However the idea of infinity has to be developed since the types of sets one encounters now are infinite sets. Just like the natural numbers are countably infinite, the set of rationals are also countably infinite because it can be put in one-to-one correspondence with the natural numbers. For the realist there is no issue with lining up two infinite sets since the idea of an actual infinity is accepted, however for the constructionist there is a major issue here because the notion of actual infinity is rejected for "potential infinity". Actual infinity to the Constructionist suggests infinity is a closed realm that can be manipulated like an object as opposed to having different existential possibilities. Even though the arithmetic of infinity, called transfinite arithmetic is not viewed favorably by Constructionists (e.g., Kronecker who was an adherent of finitism), strangely enough the development of this theory by Cantor involved many constructionist proofs which are explored in the next section.

Constructing Objects in R

If one started with two numbers “a” and “b” and thought of them as lengths with b < a, then one can show the constructability of Q simply through Euclidean constructions, i.e., arithmetic with x and + gives it the properties of a field. In other words the four operations of arithmetic work and result in constructible lengths. In this process numbers such as $\sqrt{2}$, $\sqrt{3}$, … and well as nested radicals like $\sqrt{\sqrt{2}}$ etc. also arise which do not belong to Q.

There are three ways to deal with these new objects, either formally by extending the field of rational numbers to $Q\sqrt{a}$ for every new number $\sqrt{a}$ and showing arithmetic still works, leading to the construction of a tower of quadratic field extensions which in essence show that Euclidean numbers could be given the structure of a finite field. Another alternative for constructing Euclidean numbers like $\sqrt{2}$ is showing that an algorithm exists for constructing these numbers as multi-decked fractions called continuous fractions. The third alternative is viewing these numbers as being algebraic, i.e., as numbers that are solutions to polynomial equations in one variable with integer coefficients. $\sqrt{2}$ is the solution of $x^2 = 2$. Expressing these numbers as continued fractions allows for a constructive proof of establishing their irrationality. For example,

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{\dots}}}}$$

And this representation establishes irrationality because of another constuctive result that confers irrational status by producing an infinite continued fraction, as opposed to the traditional proof by contradiction that does not help us to construct the number.

By looking at the set of all the algebraic numbers, we produce not only all the rational numbers as solutions to these equations but all the numbers that are not rational like $\sqrt{2}$.

An interesting question now is that of countability—if Q is countable, are the Algebraic numbers also countable? At first glance this seems like a preposterous question because of the abstract nature of such a set. But Cantor’s proof for the countability of these numbers is a good example of a constructive proof because it relies on the tabulation of polynomials each given a particular index. Thus, for a general polynomial $a_0 + a_1x + a_2x^2 + \dots a_nx^n$, the index used is $n + |a_0| + |a_1| + |a_2| + \dots |a_{n-1}| + |a_{n,}|$ which neatly generates every polynomial and every algebraic number orders according to the index of the polynomial that generates it. This interesting object is called the height function and results in a systematic enumeration of the algebraic numbers! (Fig. 20.1).

The question now is why this approach is better. Before jumping to any conclusions about a preference for either approach, we critique each of these philosophies.

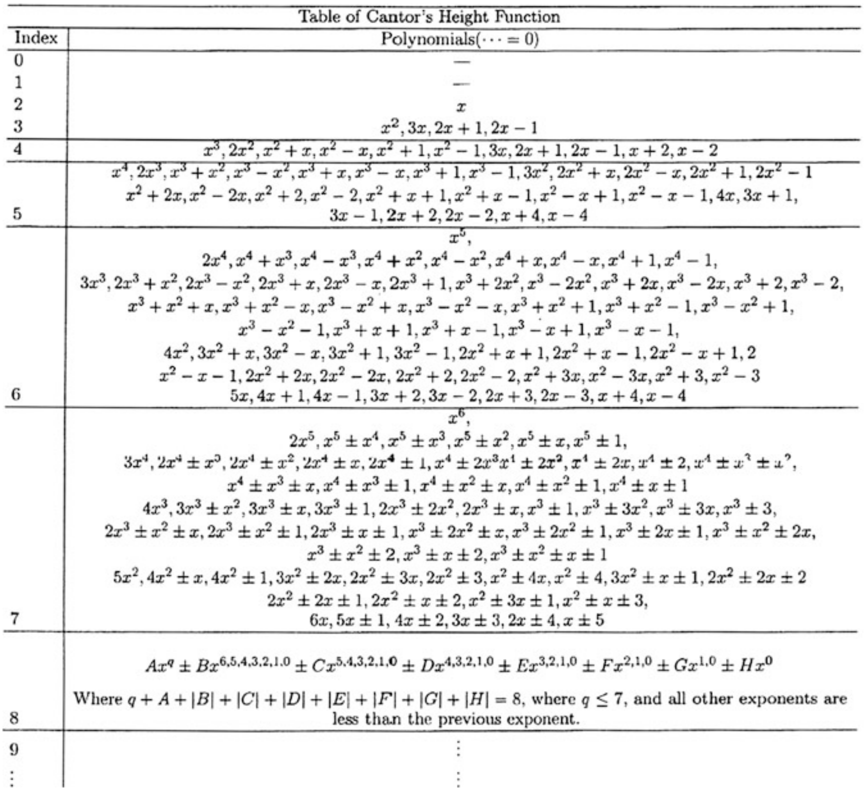


Table of Cantor's Height Function

Index	Polynomials($\cdots = 0$)
0	—
1	—
2	x
3	$x^2, 3x, 2x+1, 2x-1$
4	$x^3, 2x^2, x^2+x, x^2-x, x^2+1, x^2-1, 3x, 2x+1, 2x-1, x+2, x-2$
5	$x^4, 2x^3, x^3+x^2, x^3-x^2, x^3+x, x^3-x, x^3+1, x^3-1, 3x^2, 2x^2+x, 2x^2-x, 2x^2+1, 2x^2-1$ $x^2+2x, x^2-2x, x^2+2, x^2-2, x^2+x+1, x^2+x-1, x^2-x+1, x^2-x-1, 4x, 3x+1,$ $3x-1, 2x+2, 2x-2, x+4, x-4$
6	$x^5,$ $2x^4, x^4+x^3, x^4-x^3, x^4+x^2, x^4-x^2, x^4+x, x^4-x, x^4+1, x^4-1,$ $3x^3, 2x^3+x^2, 2x^3-x^2, 2x^3+x, 2x^3-x, 2x^3+1, x^3+2x^2, x^3-2x^2, x^3+2x, x^3-2x, x^3+2, x^3-2,$ $x^3+x^2+x, x^3+x^2-x, x^3-x^2+x, x^3-x^2-x, x^3+x^2+1, x^3+x^2-1, x^3-x^2+1,$ $x^3-x^2-1, x^3+x+1, x^3+x-1, x^3-x+1, x^3-x-1,$ $4x^2, 3x^2+x, 3x^2-x, 3x^2+1, 3x^2-1, 2x^2+x+1, 2x^2+x-1, 2x^2-x+1, 2$ $x^2-x-1, 2x^2+2x, 2x^2-2x, 2x^2+2, 2x^2-2, x^2+3x, x^2-3x, x^2+3, x^2-3$ $5x, 4x+1, 4x-1, 3x+2, 3x-2, 2x+3, 2x-3, x+4, x-4$
7	$x^6,$ $2x^5, x^5 \pm x^4, x^5 \pm x^3, x^5 \pm x^2, x^5 \pm x, x^5 \pm 1,$ $3x^4, 2x^4 \pm x^3, 2x^4 \pm x^2, 2x^4 \pm x, 2x^4 \pm 1, x^4 \pm 2x^3x^1 \pm 2x^2, x^4 \pm 2x, x^4 \pm 2, x^4 \pm x^3 \pm x^2,$ $x^4 \pm x^3 \pm x, x^4 \pm x^3 \pm 1, x^4 \pm x^2 \pm x, x^4 \pm x^2 \pm 1, x^4 \pm x \pm 1$ $4x^3, 3x^3 \pm x^2, 3x^3 \pm x, 3x^3 \pm 1, 2x^3 \pm 2x^2, 2x^3 \pm x, x^3 \pm 1, x^3 \pm 3x^2, x^3 \pm 3x, x^3 \pm 3,$ $2x^3 \pm x^2 \pm x, 2x^3 \pm x^2 \pm 1, 2x^3 \pm x \pm 1, x^3 \pm 2x^2 \pm x, x^3 \pm 2x^2 \pm 1, x^3 \pm 2x \pm 1, x^3 \pm x^2 \pm 2x,$ $x^3 \pm x^2 \pm 2, x^3 \pm x \pm 2, x^3 \pm x^2 \pm x \pm 1$ $5x^2, 4x^2 \pm x, 4x^2 \pm 1, 3x^2 \pm 2x, 2x^2 \pm 3x, 2x^2 \pm 3, x^2 \pm 4x, x^2 \pm 4, 3x^2 \pm x \pm 1, 2x^2 \pm 2x \pm 2$ $2x^2 \pm 2x \pm 1, 2x^2 \pm x \pm 2, x^2 \pm 3x \pm 1, x^2 \pm x \pm 3,$ $6x, 5x \pm 1, 4x \pm 2, 3x \pm 3, 2x \pm 4, x \pm 5$
8	$Ax^q \pm Bx^{6,5,4,3,2,1,0} \pm Cx^{5,4,3,2,1,0} \pm Dx^{4,3,2,1,0} \pm Ex^{3,2,1,0} \pm Fx^{2,1,0} \pm Gx^{1,0} \pm Hx^0$ Where $q + A + \lvert B\rvert + \lvert C\rvert + \lvert D\rvert + \lvert E\rvert + \lvert F\rvert + \lvert G\rvert + \lvert H\rvert = 8$, where $q \leq 7$, and all other exponents are less than the previous exponent.
9	⋮
⋮	⋮

Fig. 20.1 Enumeration of algebraic numbers

A Critique of Realism (Platonism)

According to Davis and Hersh (1981) your typical mathematician is a Platonist on weekdays and a formalist on Sundays. In other words, when the mathematician is actually doing mathematics he is convinced, at least implicitly and subconsciously, that he is dealing with an objective reality whose properties he is attempting to determine. However, when the mathematician is challenged to give a philosophical account of this reality, most of them would prefer to pretend that he does not believe in it after all. For instance, when the French mathematician, and Bourbakian, Jean Dieudonne was asked about his thoughts on the nature of mathematics, he answered that: "when philosophers attack us with their paradoxes we rush to hide behind formalism and say, mathematics is just a combination of meaningless symbols, and then we bring out Chapters 1 and 2 on set theory. Finally we are left in peace to go back to our mathematics and do it as we have always done, with the feeling each mathematician has that he is working with something real. This sensation is probably an illusion, but is very convenient." (1970, p. 145). So from this apparent contradiction between doing mathematics and thinking about mathematics, we can pose the following question: if the existence or non-existence has no impact on how we do mathematics, are mathematical objects even relevant?

Mathematical realism posits that mathematical objects exists independently of the human mind, language, and practices. However, these mathematical objects are not causally efficacious, or even observable. That means that mathematicians can work on mathematical problems, prove theorems and make computations, without ever encountering these abstract mathematical objects. In other words, human mathematical activity is possible regardless of the ontology of mathematics, unless there is some unknown link between human intuition and this abstract world of mathematical objects—which leads us to a second line of criticism raised against Platonism. Benacerraf (1973) formulated what is perhaps considered the most influential objection to Platonism and mathematical realism. The short version of the argument goes something like this: according to Platonism, mathematical objects are abstract objects that exist outside the spatio-temporal world of physical things like stars, cars and human beings. It is generally agreed upon that abstract entities cannot interact with concrete entities. So how can humans, who are very much concrete entities, acquire knowledge of abstract entities like mathematical objects? According to Davis and Hersh (1981), Platonists believe that human intuition must be the link between human awareness and mathematical reality. Take for instance the continuum hypothesis.[3] Its validity depends the version of set theory that is being used, and it is therefore undecidable (Gödel 1940; Cohen 1963). The Platonists, according to Davis and Hersh (1981), would say that this situation is just an example of human ignorance, and that human intuition must be developed until this situation can be resolved and truth established. The problem is of course that Platonists have yet to describe and explain human intuition, and how it could perceive an ideal and abstract reality, similarly to how our senses perceive a physical reality. Platonism in mathematics now has two problems that make it a difficult philosophy of mathematics for the rational and scientifically oriented person.

A third issue that has also been raised against Platonism, although not as influential as the previous two, is the identification problem first developed by Benacerraf (1965). The identification problem contends that since there are an infinite number of ways of identifying the natural numbers with sets, no particular set-theoretic method can be determined to be true. For instance, we could identify the natural numbers with sets in the following two ways: A: $0 = \emptyset$, $1 = \{\emptyset\}$, $2 = \{\{\emptyset\}\}$, $3 = \{\{\{\emptyset\}\}\}$ and so on, while set B: $0 = \emptyset$, $1 = \{\emptyset\}$, $2 = \{\emptyset, \{\emptyset\}\}$,

[3] The proposal originally made by Georg Cantor that there is no infinite set with a cardinal number between that of the infinite set of integers x_0 and the infinite set of real numbers (the "continuum").

$3 = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, … Benacerraf then simply asks which of these two consists of true identity statements? A or B? Both procedures could be used to define the natural numbers, and the two sets are isomorphic in their structure, but the definitions and arithmetical statements are not identical in the two sets. For instance, the two sets differ as to whether $0 \in 2$, insofar as $\emptyset$ is not an element of $\{\{\emptyset\}\}$ (Benacerraf 1965).

A Critique of Constructionism

Constructionism then seemingly offers the mathematicians a foundation for mathematics that avoids many of the paradoxes of Platonism. Yet only a few mathematicians have embraced constructionism, even though mathematicians often value constructive results with algorithmic meaning (Davis and Hersh 1981). Why is that? Perhaps the most basic and foundational consequence of constructionism, as opposed to Platonism, is the rejection of mathematical truth independent of the human mind. To the Platonists, mathematics can and must provide truth and certainty or "where else are we to find it?" (Davis and Hersh 1981); the purity of mathematics itself would be threatened. The constructionist denies mathematical truth as independent of human intuition and human mental constructions. To them, mathematics is a (inter-)subjective enterprise, in which understanding, intuition and human mental constructions are the foundations. This view of mathematics as a human, fallible and flawed enterprise becomes intolerable to the Platonists, who sees mathematics as infallible, perfect and eternally true, waiting to be discovered.

Now, the nature of truth is more of an esoteric critique, as most working mathematicians do not concern themselves with the philosophical mysteries of the foundations of mathematics—they just do mathematics. However, there are other, more mundane and practical reasons for why the mathematical community has rejected mathematical constructionism. One reason is that mathematicians do not want to give up many of the results that are valid within Platonism, or classical mathematics, but that would be rejected within mathematical constructionism, or as David Hilbert reportedly said in 1924: "the goal (of mathematics) is to obtain more, not less theorems." (Hesseling 2003, p. 74). To the constructionists, the many extra theorems of classical mathematics add no value, as they are not proved according to the principles of constructionism (as outlined earlier in this paper). One consequence of this, is that constructionism is probably less useful to the physical sciences than classic and Platonist mathematics, as the physical sciences are not directly dependent, or even concerned, with the ontological foundations of mathematics. Fewer valid mathematical results would produce a smaller toolbox for the physical sciences.

Other reasons, which are also less philosophical in nature, comes from how results are obtained in Platonist mathematics and constructionist mathematics respectively. Proofs that use classical techniques that are allowed in Platonist mathematics, but not constructionist mathematics, are often short, elegant and clever—ideas that are closely related to the concept of mathematical beauty—while the corresponding constructive proof is longer and far more convoluted.[4] The constructivist proof has lost all of its elegance (Snapper 1979). There are also theorems that are proved in constructionist mathematics, but that are considered meaningless and invalid in Platonist mathematics due to different definitions of concepts. One such example is the theorem that states that every real-valued function which is defined for all numbers is continuous. This sounds like a strange statement outside constructivist mathematics, but within constructionist mathematics a real-valued function is defined for all real numbers if and only if for each real number r, which has been constructed, the real number $f(r)$ can be constructed. Therefore, any discontinuous function that a Platonist mathematician might mention, would not satisfy this constructive criterion (Snapper 1979). Results like this seem so bizarre to many mathematicians, that they reject constructionist mathematics in its entirety.

Constructionism and Pedagogy

Brouwer's First Act of Intuitionism is the foundation for his intuitionist beliefs. In it, he separates mathematics from mathematical language and logic, and defines mathematics as a mental exercise. Mathematics is constructed by the mind by performing changes on its own thought in time, then abstracting away from the particulars of these constructions (Brouwer 1907). Brouwer's rejection of mathematics as pure logic was a reaction to the strong relationship between semantic and ontological realism in Platonism. The Platonist would argue that our mathematical theories should be taken at face value and that they are true, and that they could not be true in the absence of mathematical objects. Or, as Davis and Hersh puts it: "To show that all of mathematics is just an elaboration of the laws of logic would have been to justify Platonism, by passing on to the rest of mathematics the indubitability of logic itself." (1981, p. 332). Brouwer, on the other hand, meant that the truth of a mathematical proposition can only be determined by a mental construction that proves it to be true. He therefore, for instance, rejected the principle of the excluded middle, and contended that our usual logical principles were abstracted from our dealing with finite sets, and these principles could not be applied to infinite sets (Ferreiros 2008).

Take for instance the infinite series of the natural numbers: $1 + 2 + 3 + 4 + 5\ldots$ which is clearly a divergent series. However, if we treat and manipulate this series as if it was a finite series, we can see all kinds of strange effects. Srinivasa Ramanujan presented a simple heuristic example of this in chapter 8 of his first notebook:

He first assumes that the sum of the series can be expressed as $c = 1 + 2 + 3 + 4\ldots$ He then goes on to multiply this equation by 4, and subtract the second equation from the first equation:

[4] See for instance a classic and constructive proofs for the fundamental theorem of Algebra.

$$\begin{aligned} c &= 1+2+3+4+5+6\ldots \\ 4c &= 4+8+12\ldots \\ -3c &= 1-2+3-4+5-6\ldots \end{aligned}$$

Ramanujan then uses the fact that the alternating series of $1 - 2 + 3 - 4 + 5\ldots$ is the power series expansion of the function $\frac{1}{(1+x)^2}$, but with $x = 1$. He can then say that $-3c = 1 - 2 + 3 - 4 + 5\ldots = \frac{1}{1+1^2} = \frac{1}{4}$. Dividing both sides by −3, one gets: $c = -\frac{1}{12}$.

Which is clearly an absurd result, but illustrates how strange results can appear if you treat an infinite (divergent) series as a finite series. We chose to call this a *platonic leap of faith*, and it illustrates how logic and human intuition diverge (!) when we move from the finite to the infinite.[5]

Intuitionists, or constructionists, thus find non-constructive existence proofs unacceptable. Non-constructive existence proofs are proofs that claim to demonstrate the existence of a mathematical entity having a certain property without producing a method for generating such an entity. The difference between providing a method for creating a certain mathematical object and simply proving that such an object must exist, is in many ways related to the ideas of *need for certainty* and *need for causality*, which are two subcategories of what Harel (2013) calls *intellectual need*. Intellectual need is essentially defined as the knowledge an individual needs to learn, acquire or construct, to solve a particular problem. The *need for certainty* is, according to Harel (2013), based on a Piagetian theory of equilibration, a natural human desire to know whether a conjecture is true or false. Truth and certainty, however, may not be enough for an individual. The individual will often also want to know how and why something is true. The need for causality is a person's desire to explain and to determine a cause of phenomenon. Constructive proofs can be compared with a need for causality, while non-constructive proofs can be said to be more closely related to a need for certainty: "Mathematicians routinely distinguish proofs that merely demonstrate from proofs which explain." (Steiner 1978, p. 135). A typical example of noncausal, and non-constructive, proof would be the proof by contradiction to establish the irrational status of $\sqrt{2}$.

However, the analogy between constructionism in mathematical philosophy and the need for causality in teaching and learning (didactical situations) may not be perfect. Proofs by mathematical induction are for instance not rejected a priori, as they could be seen as a sort of iterated modus ponens, which is a logical principle generally accepted by the intuitionists. Within the mathematics education community, there are those who claim that proofs by induction establish certainty, but they do not provide an explanation for why a proposition is true: "a proof that explains must provide a rationale based upon the mathematical ideas involved, the

[5] A rigorous proof $\zeta(-1) = -1/12$ can be found in: Stopple, J. (2003). A primer of analytic number theory: from Pythagoras to Riemann. Cambridge University Press.

mathematical properties that cause the asserted theorem to be true." (Hanna 1990, p. 9). Harel proposes a possible resolution to this ostensible difference between constructive proofs and proofs that explain, by drawing on the ideas of Brouwer: "Hanna (1990), who argues that proofs by mathematical induction, for example, are proofs that prove but do not explain. Our position is different. We hold that it is the individual's scheme of doubts, truths, and convictions in a given context that determines whether an argument is a proof or an explanation." (2013, p. 128). Here, Harel presupposes mathematics as a human and mental activity, and proposes that whether or not a proof provides causality, depends on the individual learner's preexisting understanding and mental schemes.

Again, we go back to the series of sum of the natural numbers to illustrate Harel's point. For the first n numbers, we have that $0+1+2+3\ldots+n=\frac{n(n+1)}{2}$. Proof by induction would first start by showing that the statement holds for $n=1$, which is obviously true, as the two sides of the equation would be equal. The inductive step shows that if the statement is true for $n=k$, then it would also be true for $n=k+1$. We assume that the statement is true for some value of k and we must now demonstrate that the statement is true for $k+1$:

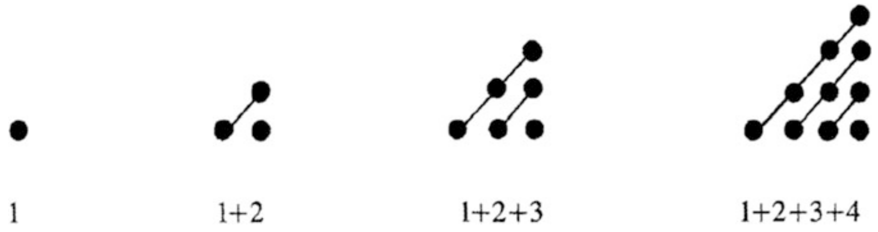
$$(0+1+2+3\ldots+k)+(k+1)=\frac{(k+1)((k+1)+1)}{2}$$

Using the induction hypothesis that the statement holds for $n=k$, the left hand side can be rewritten to:

$$\frac{k(k+1)}{2}+(k+1)=\frac{(k+1)((k+1)+1)}{2}$$

Thereby showing that indeed $n=k+1$ holds.

Now, Hanna (1990) claims that although this proof demonstrates that a certain mathematical statement is true, it does not show why the sum of the first n natural numbers is $\frac{n(n+1)}{2}$. However, if we look at proof by induction as a recursive process, we can illustrate this sequence in the following way:



Here we see that the dots form isosceles right triangles, and if we double them, we get rectangles with $n(n+1)$ dots. The rectangles are exactly twice the size of the corresponding sum, so the sum of the first n numbers is $\frac{n(n+1)}{2}$, and we can do this for $n = 1$, $n = 2$, $n = 3$, and so on. So, as Harel (2013) says, a proof by induction can very much be a proof that also explains—it depends on the individual's preexisting knowledge and how the individual perceives the proof. We now see how a constructionist proof, that is based on human mental activity and human intuition, is in many ways analogous to mathematics educators' call for proofs that explain—both begin with the human mind, and not the laws of logic, as a starting point!

Concluding Points

Mathematics is one single thing. The Platonist, formalist and constructionist views of it are believed because each corresponds to a certain view of it, a view from a certain angle, or an examination with a particular instrument of observation. This view is corroborated by Friend in her thesis on pluralistic views of mathematics being compatible with model building (Friend 2017). Grosholz (2016) gives other examples of this working philosophy through models (examples from celestial mechanics) which are developed simultaneously by different people using completely different methods from analysis that reflect different, even apposite views of the philosophy of mathematics. There are plenty of other examples that can be used to make the case that the realist/anti-realist dichotomy is false. One such classical result is: Gauss' result about the constructability of regular polygons and its relationship to Fermat primes. Most modern books use a realist approach using heavy tools from abstract algebra, whereas Gauss invented those tools very informally as he was tackling the problem from a number theoretic viewpoint. His approach is very anti-realist. More modestly put, the realist/anti-realist dichotomy is reconcilable.

References

P. Benacerraf, What numbers could not be. Philo. Rev. **74**, 47–73 (1965)

P. Benacerraf, Mathematical truth. J. Philos. **70**(19), 661–679 (1973)

E. Bishop, *Foundations of Constructive Analysis* (McGraw-Hill, New York, 1967)

E. Bishop, *Schizophrenia in Contemporary Mathematics*. American Mathematical Society Colloquium Lectures (University of Montana, Missoula, 1973); reprinted in *Errett Bishop: Reflections on Him and His Research*. American Mathematical Society Memoirs 39

L.E.J. Brouwer, Over de grondslagen der wiskunde. Ph.D. Thesis, University of Amsterdam, Department of Physics and Mathematics (1907)

D. Bridges, F. Richman, *Varieties of Constructive Mathematics*. London Mathematical Society Lecture Notes 97 (Cambridge University Press, Cambridge, 1987)

P.J. Cohen, The independence of the continuum hypothesis. Proc. Natl. Acad. Sci. **50**(6), 1143–1148 (1963)

P. Davis, R. Hersh, *The Mathematical Experience* (Springer Science & Business Media 1981)

J. Dieudonné, The work of Nicholas Bourbaki. Am. Math. Monthly **77**, 134 (1970)

J. Ferreirós, The crisis in the foundations of mathematics, in *The Princeton Companion to Mathematics* (2008), pp. 142–156

M. Friend, *Pluralism in Mathematics: A New Position in the Philosophy of Mathematics* (Springer, Netherlands, 2014)

M. Friend, Mathematical Theories as Models, ed. by B. Sriraman. in *Humanizing Mathematics and its Philosophy*. (Birkhäuser, Cham, 2017)

K. Gödel, *The Consistency of the Continuum-Hypothesis* (Princeton University Press, Princeton, NJ, 1940)

E. Grosholz, *Starry Reckoning: Reference and Analysis in Mathematics and Cosmology*. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 30 (Springer International Publishing, 2016)

G. Hanna, Some pedagogical aspects of proof. Interchange **21**(1), 6–13 (1990)

G. Harel, Intellectual need, *Vital Directions for Mathematics Education Research* (Springer, New York, 2013), pp. 119–151

D.E. Hesseling, *Gnomes in the Fog: The Reception of Brouwer's Intuitionism in the 1920s* (Birkäuser Verlag, Boston, 2003)

M. Mandelkern, Brouwerian counterexamples. Math. Mag. **62**(1), 3–27 (1989)

P. Raatikainen, Concepts of truth in intuitionism. Hist. Philos. Logic **25**, 131–145 (2004)

F. Richman, Existence proofs. Am. Math. Monthly **106**, 303–308 (1999)

E. Snapper, The three crises in mathematics: logicism, intuitionism and formalism. Math. Mag. **52** (4), 207–216 (1979)

E. Schechter, Constructivism is difficult. Am. Math. Monthly **108**, 50–54 (2001)

M. Steiner, Mathematical explanation. Philos. Stud. **34**, 135–151 (1978)

Chapter 21
To the Edge of the Map

Barry Mazur

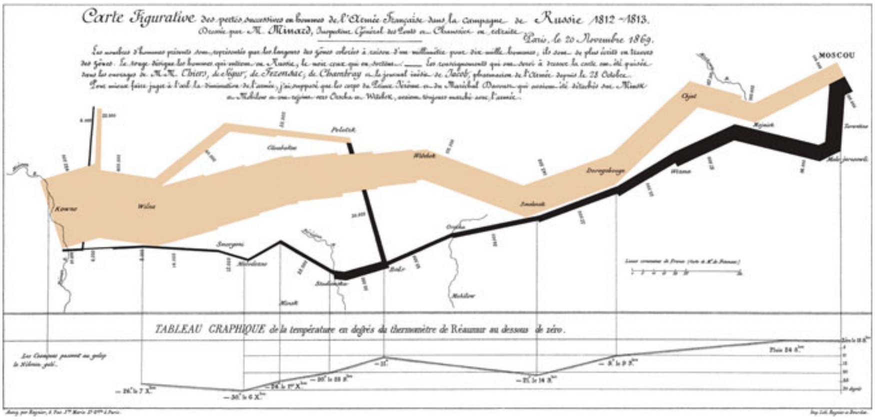
The title—*Map and Territory*—that Shyam Iyengar chose for this volume is, of course, rich in possible interpretations. The word *map* suggests any contrivance—perhaps of ephemeral utility—meant to model the geography of any territory. I'll take the title to be an invitation to write about the manner in which we fashion structures of thought—the '*maps*,'—in order to understand, and negotiate our way through, whatever realm it is that encompasses the objects of our thought—the '*territories*.'

It may be bad strategy to blurt out the point of this essay on the first page, but it is simply this: any faithful *map* of our thought processes has problems setting its boundaries: we don't think in closed systems and it is often at the edge of the map where things begin to get really interesting.

Maps offer (sometimes quite pointed) narratives—often quite political—for the territories they're meant to record. And consider the agonizing narrative palimpsested onto Minard's well known "figurative map" of Napoleon's losses of troops at the various stages of his 1812–1813 invasion of Russia (where the width of the colored band is meant to be proportional to the number of soldiers still active at that point in the campaign, and—below that—the graph of the temperature at that point).

B. Mazur (✉)
Department of Mathematics, Harvard University, One Oxford Street,
Cambridge, MA 02138-2901, UK
e-mail: mazur@math.harvard.edu

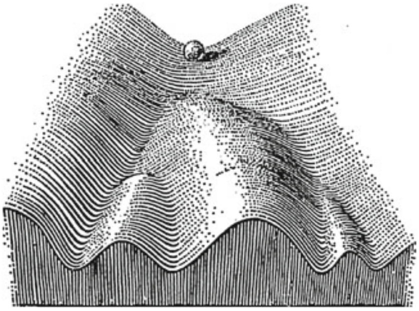
S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_21


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
MOSCOU
Kowno
Wilna
Polotzk
Gloubokoe
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Wiazma
Orscha
Bobr
Studienska
Smorgoni
Molodeczno
Minsk
Mohilow
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

But all maps orient, organize, delimit. What is fascinating is the manner in which a map frames its limits, and points to the dragons lurking at its ends, or distorts the perimeter of the onionskin of the globe by flattening it onto the page. The unproclaimed assumption of any map is that the boundary, the perimeter, makes sense: it encloses a meaningful-in-itself territory, a closed system. Precisely because of that it becomes particularly interesting to press beyond those borders.

In some contexts, it is natural to have a more relaxed view of the borders of one's map. The biologist C. H. Waddington devised a compelling geographical metaphor for the various paths an embryo might take in its development: it is as if the embryo were a stone poised on a hilltop, and morphogenesis consists in rolling down the hill, under the force of gravity. Well, there are the natural deep grooves that constitute the path for'normal' development, but for some embryos the ricochet of their downward ride might knock them into a neighboring (but abnormal) pathway. Therefore, to fully understand the repertoire of possible development routes, one must plot out all those deviant neighboring grooves—i.e., one must map out a significantly larger span of what he called the *epigenetic landscape*.

This cartographical spread beyond the probable to the possible is analogous to the use, in Physics, of Feynman's diagrams.[1]



[1] For example, in Feynman's book QED see the classical law: *angle of incidence equals angle of reflection* proved by dealing with *all possible reflecting paths*.

In this essay I will be considering the richness of going to the limit and beyond—in the 'maps and territories' that organize our thought.

Mathematical objects of study often don't come as singletons, to be examined in isolation. They tend to appear as particular individuals living in a family of like-structured objects. Often the members of such a family can be labeled by continuous parameters, these 'continuous parameters' constituting a geometric space of some sort, where two labels are very close to each other if the objects they label are 'very like each other.' That is, geometric features of the parameter space reflect the relative relationships between the various mathematical objects that are labelled by these parameters.

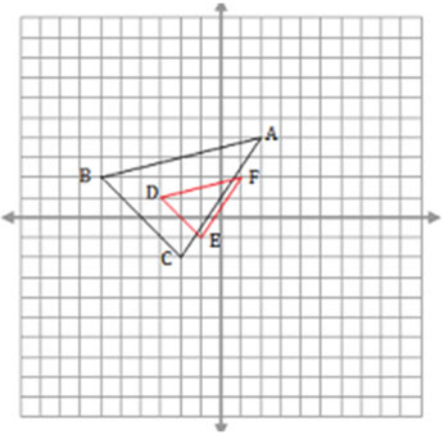
The technical term denoting a parameter space that labels the various members of a species of mathematical object is **Moduli Space**. ("Moduli" meaning "parameters"—and more specifically, parameters describing the possible ways of varying a mathematical object and staying within its species.) The idea of systematically studying mathematical objects in the context of their possible variation is ubiquitous in mathematics (and has achieved the status of a high art in Algebraic Geometry). A comprehension of the detailed structure of the moduli space of a given species often provides a powerful way to understand the deeper structures of the very objects of that species. And a surprise in store for anyone who thinks along these lines is that the moduli space classifying any given species of mathematical object is often extraordinarily rich in structure—*richer in structure than the species that it classifies*.

Continuing the metaphorical reach of *map and territory*, we can view a moduli space as *map* and the species it is meant to study as *territory*. Here, as I mentioned, the map may have more intricate structure than the territory it maps. In Algebraic Geometry, there is—at times—an interesting reversal of focus, where the moduli space, per se, becomes the primary object of investigation.[2]

Here are three examples: the first is something of a toy illustration; the second is an example where the exquisite complexity of the edge was quite a surprise when it was first appreciated; and the third is where the edge is nothing more than a *single point* and yet centering one's focus around that point makes a deep and rather amazing connection with a somewhat different field of mathematics.

[2] To allude to an example of this switch of focus, I might mention *Shimura varieties*, these being moduli spaces that classify a certain species of mathematical object interesting enough in its own right. But the *Shimura varieties* themselves play a key role in a significantly different project: establishing a bridge between two disparate mathematical fields: representation theory of reductive groups and algebraic number theory.

Consider Triangles in the Euclidean Plane, Taken Up to Similarity

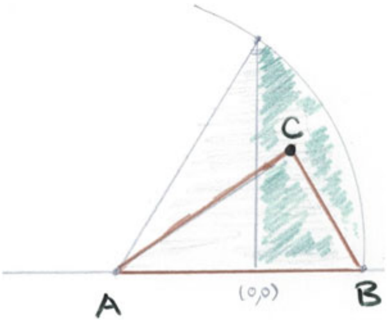


For any similarity-equivalence class of triangles, $\mathcal{S}$, choose a member-triangle of that class: $\Delta := ABC \in \mathcal{S}$. After possible renaming the vertices, suppose its longest side is AB, noting that Δ might be isosceles or equilateral, and therefore would have two or three 'longest sides'—if so, just choose one of them to be AB.

Rescale Δ—which doesn't change its similarity class—so that AB is of length 2. Now place Δ in the plane $\mathbf{R}^2$ with the vertex A at the point $(-1, 0)$, the vertex B at the point $(+1, 0)$. Flip Δ around the x-axis if necessary to arrange that its third vertex, C is in the upper-half-plane. Flip Δ about the y-axis if necessary to arrange that C is in the (closed) upper-right-quadrant of the plane. That is, $C = (a, b)$ with $a \geq 0$ and $b > 0$. Since AB is (one of) the longest side(s) of Δ, C lies in the shaded region

$$\mathcal{M} := \{(a, b) \in \mathbf{R}^2 \;\; |a \geq 0;\; b > 0; (a + 1)^2 + b^2 \leq 4\}$$

given in the diagram below.



Call this—so arranged—point C in the (closed) upper-right-quadrant of the plane the **modulus** of the similarity-equivalence class of triangles $\mathcal{S}$:

$$C = \mu(\mathcal{S}) \in \mathcal{M}.$$

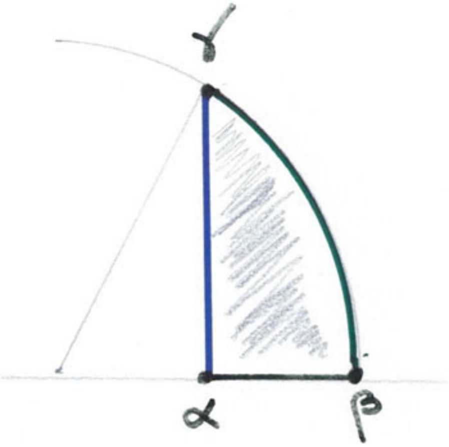
We have a one-one correspondence between similarity-equivalence class of triangles and their *moduli*; i.e., the points in $\mathcal{M}$:

$$\mathcal{S} \leftrightarrow \mu(\mathcal{S})$$

So $\mathcal{M}$ is the *moduli space* of the species: similarity equivalence classes of triangles in the Euclidean plane. It is our ‘map.’ The plane geometry of $\mathcal{M}$ relates nicely to the structure of the species that it ‘maps out’ in that points close to each other in $\mathcal{M}$ label similarity equivalence classes of triangles that have representatives that are close. Any real-valued continuous function, for example, $f : \mathcal{M} \to \mathbf{R}$ can be interpreted as providing a numerical invariant of any similarity-class of triangles, these invariants being sensitive to the closeness of different similarity-classes. And now, consider its boundary.

The boundary of $\mathcal{M}$ consists of three pieces:

- the vertical piece $\alpha\gamma$,
- the arc of the circle $\beta\gamma$, and
- *the horizontal piece* $\alpha\beta$.



The first two pieces are actually in $\mathcal{M}$ and (except for the common point γ) they label similarity equivalence classes of *isosceles* triangles. The distinction here being that points on $\alpha\gamma$ label such similarity equivalence classes where the two equal sides of the triangle are *shorter* than the third side, while points on $\beta\gamma$ label those where the two equal sides of the triangle are *longer* than the third side. The point γ at which they meet labels the (unique) equivalence class of equilateral triangles; i.e., triangles such that all three sides have equal length. In both of these sides, we encounter the issue of *nonrigidity*. Say that a mathematical object is *rigid* if it admits no nontrivial symmetries. ‘Most’ triangles are rigid but not the ones on these two sides of the boundary.

It is the third piece of the boundary, the horizontal interval $\alpha\beta\,:\,\{(x,0)\ \ 0 \leq x \leq 1\}$ to which I want to pay particular attention, even though it is not formally part of our moduli space $\mathcal{M}$ at all: points on the interval $\alpha\beta$ correspond—if to anything at all—to degenerate flattened triangles where the third vertex C lies in the line-segment AB. Studying these objects—a teratology of triangles—might seem bizarre. Yet it is precisely in the neighborhoods of such regions in many of the moduli spaces currently studied where profound things take place. It is often—to push the metaphor—the very edge of the map that captures one's attention. Not, perhaps, for this toy model, the moduli space of similarity-classes of triangles,[3] but for other moduli spaces in mathematics[4]—in particular, for the next example to be discussed: namely, the Mandelbrot set, viewed as moduli space.

The Mandelbrot Set

Let c a complex number. Consider the transformation of the complex plane

$$z \ \mapsto\ z^2 + c$$

If we ask questions about how this (seemingly comprehensible) geometric transformation—call it $\mathcal{D}(c)$—behaves–or 'performs'—when we iterate it:

[3] On the other hand, this moduli space has some interesting dynamics, tending toward that bottom edge. I'm thankful to Curt McMullen, Susan Holmes and Persi Diaconis for telling me about this, and about the relevant literature regarding the statistics of the operation of performing a **barycentric subdivision** of a triangle. That is, if $\Delta = ABC$ is our triangle, let D denote the barycenter of Δ and decompose Δ into a union of three triangles

$$\Delta_1 = ABD, \quad \Delta_2 = BCD, \quad \Delta_3 = CAD.$$

We can think of

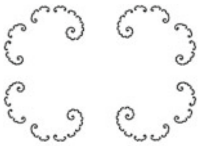
$$\Delta \mapsto \{\Delta_1, \Delta_2, \Delta_3\}$$

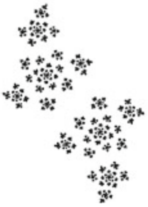
as a many-valued transformation on similarity classes of triangles, and therefore on points in $\mathcal{M}$. Statistically, this transformation produces thinner triangles, i.e., represented by lower points in $\mathcal{M}$. So, iteration of this operation can be thought of as producing gentle cascades, sending points on our moduli space—statistically speaking—towards the bottom horizontal interval. This is a consequence of I. Barany, A. Beardon, and T. Carne: *Barycentric subdivision of triangles and semigroups of Möbius maps*, Mathematika, **43** 165–171 (1996) and see Bob Hough's *Tesselation of a triangle by repeated barycentric subdivision*' in Elect. Comm. in Probab. **14** 270–277 (2009) for a refinement of the result and a discussion of the literature about it.

[4] Quite a few important results about the general structure of moduli spaces come from a close examination of the boundary—for example: a close analysis of degenerate—i.e. singular—curves and the manner in which a smooth curve might degenerate to the boundary of M_g, the moduli space of curves of genus g, leads to a proof of the irreducibility of that moduli space. To quote P. Deligne and D. Mumford in their paper *The irreducibility of the space of curves of given genus*: "The basis … is to construct families of curves X, some singular, … over non-singular parameter spaces, which in some sense contain *enough singular curves* to link together any two components that M_g might have.".

$$z \mapsto z^2 + c \mapsto \left(z^2 + c\right)^2 + c \mapsto \ldots,$$

i.e., when we ask questions about it as a *dynamical system*—for example what its orbits look like. The **forward orbit** of a point $z_0 \in \mathbf{C}$ is the set of points in $\mathbf{C}$ that occur as the image of z_0 under the sequence of iterations of the transformation $z \mapsto z^2 + c$. One feature of interest is called its **Filled Julia Set** $J(c)$ which by definition is the set of points $z \in \mathbf{C}$ whose forward orbits are bounded. The simplest (i.e., the most misleading) example of such a Filled Julia Set is when $c = 0$: $J(0)$ is the unit disc. It is a theorem that there are only two types of topology for any 'Filled Julia set' $J(c)$. These can be *connected*—or they can be *homeomorphic to a Cantor set*—in which case they are called **dust**:





Consider, then, our 'territory', which is this species of mathematical object:

$$\textit{Dynamical Systems } \mathcal{D}(\mathrm{c}) : \textit{ iteration of } z \mapsto z^2 + c$$

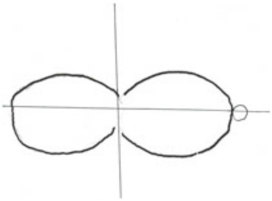
where c is a complex number for which the Filled Julia Set $J(c)$ is connected—i.e., is not 'dust'.

And consider its corresponding 'map'—i.e., the *moduli space* parametrizing this species:

> The region in the complex plane consisting of complex numbers c that correspond to such dynamical systems $\mathcal{D}(c)$; i.e., those with $J(c)$ connected.

This region is (now) called the **Mandelbrot set** (as are regions related to analogous problems).

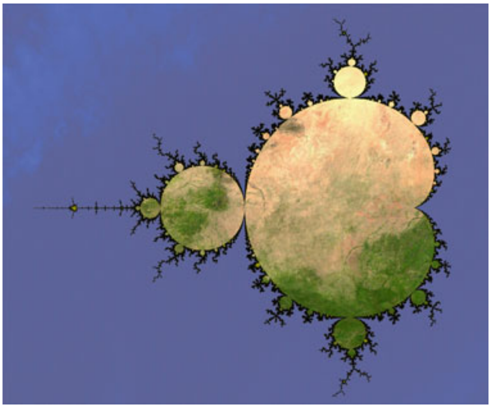
Up to the end of the First World War, the foundations of the theory of such structures was called Fatou-Julia theory. I'm guessing that Fatou or Julia would not have been able to make too exact a numerical plot of these regions. It would most likely look something like this:



(In fact, Julia made a drawing of the Mandelbrot set rather like that, a photo of which appears in *Fatou, Julia, Montel,: le grand prix des sciences mathématiques de 1918, et après* (Springer, 2009) by Michèle Audin.)

Partly due to the ravages of the First World War, and partly from the general consensus that the problems in this field were essentially *understood*, there was a lull, of half a century, in the study of *Julia* sets and the now-named *Mandelbrot* sets.

But in the early 1980s Mandelbrot made (as he described it) "a respectful examination of mounds of computer-generated graphics." His pictures of such sets were significantly more accurate, and tended to look like the figure below (which is an even more modern version of the ones Mandelbrot produced)[5]:



From such pictures alone it became evident[6] that there is an immense amount of structure to the regions drawn and to their perimeters. Specifically: the perimeters, whose self-similar infinitely laciness[7] was captured by Mandelbrot's then novel—but now ubiquitous—piece of vocabulary: *fractal*. This complexity at the edge of such maps almost immediately re-energized and broadened the field of research, making it clear that very little of the basic structure inherent in these Julia sets had been perceived, let alone understood. It also suggested new applications and

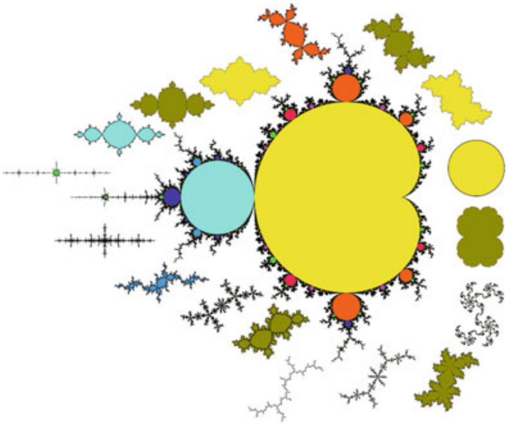
[5] I'm grateful to Sarah Koch and Xavier Buff for this elegant picture.

[6] Computers nowadays (as we all know) can accumulate and manipulate massive data sets. But they also play the role of *microscope* for pure mathematics, allowing for a type of extreme visual acuity that is, itself, a powerful kind of evidence.

[7] "as the small pool by the elm ices over," which is a line of Kevin Holden's poem *Julia Set* that appeared in his book *Solar* (Fence Books, 2016).

Mandelbrot proclaimed—with some justification—that “Fatou-Julia theory ‘officially’ came back to life”[8] on the day when, in a seminar in Paris, he displayed his illustrations.

Here is a picture[9] of the Mandelbot set with sample pictures of the corresponding Filled Julia Sets—color-coded so that a Filled Julia Set $J(c)$ will have the same color as the portion of the Mandelbrot set containing the point c.



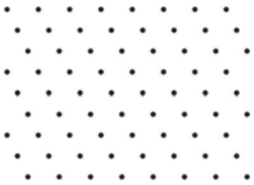
The Moduli Spaces Classifying Elliptic Curves

Elliptic curves play a key role in a surprising number of different mathematical subjects. Moreover, depending on the subject in which you want to consider them, *elliptic curves* will have surprisingly different appearances, definitions,[10] and uses. And mysteries. For the purposes of this essay, we will take these mathematical concepts as they make their first appearance in complex analysis: an *elliptic curve*—for us—can be thought of as given by an *equivalence class* of rectangular or parallelogram lattices in the complex plane **C**:

[8] This is from Benoit Mandelbrot’s book *Fractals and Chaos*.

[9] I’m grateful to Sarah Koch and Xavier Buff for the diagrams and comments. Xavier Buff mentioned that if you think of the Mandelbrot set as an island in an ocean; and each filled Julia set $J(c)$ corresponding to a point c of the Mandelbrot set, as an ‘inhabitant’ of this island, the main open question in the subject is whether there is a hidden landmass (component of the interior of M) where inhabitants are thin (have empty interior…but have positive measure). As far as we know, says Xavier, all inhabitants that live inland have some interior. Thanks, as well, to Curt McMullen for helpful comments.

[10] One often deals rather with elliptic curves endowed with a bit of extra structure.

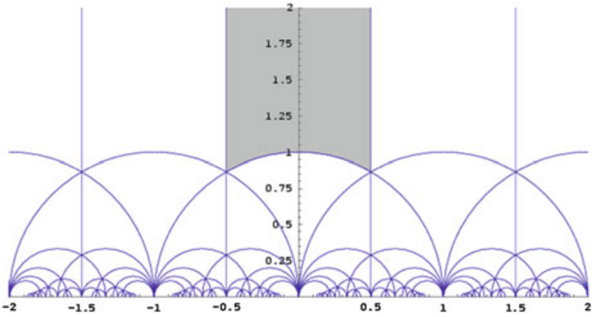


where, to be more precise, such a lattice, L, is a subgroup of the additive group $\mathbf{C}$ generated by two elements that are linearly independent over the field of real numbers $\mathbf{R}$ (so that L consists of the points of a configuration of the type drawn in the picture above). The equivalence relation that we will be considering might be called ‘complex-similarity.’ That is, two such lattices L, L' are in the same equivalence class if there is such a nonzero complex number a such that *scaling by* a brings L to L'; i.e., $a \cdot L = L'$.

How can we construct the ‘moduli space,’ that maps out the territory of this mathematical species: lattices in the complex plane taken up to complex similarity? Any lattice L is generated by two complex numbers, but we are allowed to scale our lattice so we can always arrange it so that one of those complex numbers is the real number 1. If we think, then, of L as generated by 1 and some other complex number τ, since 1 and τ are required to be linearly independent over the field of real numbers, τ is genuinely complex—i.e., not real—and by changing its sign, if necessary, we can arrange it so that τ is in the upper-half of the complex plane. Thus “τ” determines the complex similarity class of a lattice—i.e., the lattice L generated by 1 and τ.

But there are many points τ in the upper half plane that generate, when taken together with 1, this same complex similarity class of lattices. For instance, L is the same as the lattice generated by 1 and $\tau + 1$; and its similarity class is the same as the lattice L' generated by 1 and $-1/\tau$ (rescale L' by multiplying by τ; so $\tau L'$ is generated by $\tau = \tau \cdot 1$ and $-1 = \tau \cdot -1/\tau$).

So we can change τ by any of these two transformations $\tau \mapsto \tau + 1$ or $\tau \mapsto -1/\tau$ or by any combination of these transformations and their inverses and still have a “τ” that together with 1 generates a lattice in the same complex similarity class as L. The group generated by these two transformations ‘tiles’ the upper half plane in quite an intriguing way:



Here, for any similarity class of lattices, there is a unique τ in *any* of these tiles such that the lattice generated by 1 and τ is in that similarity class (if we are careful about our description of which of the boundary pieces belong to which tiles[11]). So, for example, the points of the shaded region (this comprises a single tile) is in one-one correspondence with the set of similarity class of lattices.

This abundance of different ways of generating the same complex similarity class turns out to be far more manageable than one might first think, thanks to the existence of something called the **elliptic modular function**—or, more familiarly, the *j-function*—"*j*."

The elliptic modular function $j(\tau)$ is a complex analytic function on the upper half-plane $\tau \mapsto j(\tau)$ that maps each of these tiles in a one-one manner onto the entire complex plane, thereby giving us a clean parametrization of this species: complex similarity classes of lattices.

That is, if L is generated by 1 and τ, the complex number $j(\tau)$ depends only on L, and—in fact—only on the similarity equivalence class $\{L\}$ of L. So rename $j(\tau) =: j(\{L\})$ and call the complex number $j(\{L\})$ **the *j*-invariant** of the similarity equivalence class $\{L\}$.

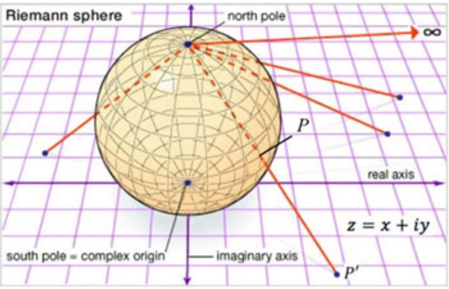
Having done this, we get a one-one correspondence between similarity equivalence classes and their j-invariants. Moreover, any complex number *is* the j-invariant of a (unique) similarity equivalence class of a lattice:

$$\{L\} \quad \leftrightarrow \quad j(\{L\}) \in \mathbf{C}.$$

So the complex plane **C** plays the role of the 'moduli space'—our metaphorical map—of the species: complex similarity classes of lattices.

What comprises the *edge* or the *end* of this map? The answer is that we must pass from the complex plane to the Riemann sphere, by adjoining the "point at infinity," ∞. Here is a picture of the stereographic projection of the Riemann sphere onto the complex plane, with the "North pole" acting as this point at infinity:

$$\mathbf{S} := \mathbf{C} \cup \{\infty\}.$$



[11] One should stipulate that the part of the boundary of the shaded region consisting of

- the right-hand vertical line—i.e., contained in the line with x-coordinate 0.5—and
- the part contained in the arc of the unit circle with positive x-coordinate

be *not* included.

The single point $\infty \in \mathbf{S}$ comprises the *end of our map*; i.e., of our moduli space. It corresponds, if to anything at all, to a curious degenerate similarity class of lattices: namely, the limit—as y tends to infinity—of the similarity classes of lattices L_y generated by 1 and $\tau = iy$.

Since $j(\tau) = j(\tau + 1)$, the value of j at τ depends only on $q := e^{2\pi i\tau}$ so we may rewrite the elliptic modular function as a function of this new variable q. Note that the limiting value of q for $\tau = iy$ with y tending to infinity is: $q = 0$.

Something quite curious happens when we view the elliptic modular function as centered about this missing point $q = 0$, i.e., as given by its Laurent series in $q = e^{2\pi i\tau}$,

$$j(\tau) = \frac{1}{q} + 744 + \sum_{n\geq 1} c_n q^n = \frac{1}{q} + 744 + 196884q + 21493760q^2$$
$$+\ 864299970q^3 + 20245856256q^4 + \dots$$

Or, if you wish, by this same formula presented as its Fourier expansion:

$$j(\tau) = e^{-2\pi i\tau} + 744 + \sum_{n\geq 1} c_n e^{2\pi i n\tau},$$

There are two surprising things about these (Fourier, or Laurent series) coefficients c_n. First, they are all non-negative integers. But also these numbers $c_1, c_2, c_3, c_4, c_5, \dots$ lead us to strikingly profound structure in a part of mathematics that one might imagine to be quite remote for our starting place, lattices in the plane.

Namely, the **Monster group** M (also known as the *Fischer-Greiss Monster group*). M is a finite simple group that is referred to as 'sporadic' because it isn't a member of any of the standard infinite families of finite simple groups connected with various types of geometries. It is quite large, having

$$2^{46}3^{20}5^{9}7^{6}11^{2}13^{3}171923293141475971$$

elements. As with all finite groups the group M can be 'represented' as a group of linear transformations of complex N-dimensional space for various dimensions N. The dimensions N for which M acts irreducibly on $\mathbf{C}^N$—including the 1-dimensional space on which M acts trivially—comprise a finite list of numbers:

$$1, 196883, 21296876, 842609326, 18538750076, \dots.$$

So, the very smallest dimension N for which this curious group M can be viewed as a 'group of linear transformations' on $\mathbf{C}^N$ is 196883. In 1978 John McKay made the following puzzling and somewhat amazing observation: the first few Fourier coefficients of the elliptic modular function j can be expressed as sums—with very few summands!—of the dimensions N for which M acts irreducibly on $\mathbf{C}^N$. For example:

$196884 = 1 + 196883$,
$21493760 = 1 + 196883 + 21296876$, and
$864299970 = 2 \times 1 + 2 \times 196883 + 21296876 + 842609326$,
$20245856256 = 3 \times 1 + 3 \times 196883 + 21296876 + 2 \times 842609326 + 18538750076$.

This extremely arresting purely numerical observation suggested a world of new structure: it led to the conjecture that there lurked an infinite sequence of 'natural in some sense' complex representation spaces of the Monster group,

$$V_1, V_2, V_3, \ldots, V_n, \ldots$$

where, for $n = 1, 2, 3, \ldots$ the Fourier coefficient c_n of the elliptic modular function is equal to the dimension of V_n. This seemed, perhaps, so startling at the time that the conjecture was labeled *monstrous moonshine*. When eventually proved[12] it has given birth to another profound field in mathematics, and intimate links with physics. And all this, inspired by the elliptic modular function j; i.e., by contemplating lattices in the plane.

There seems to be a tenacious unity to mathematics, where ideas trespass the borders of any field, any designated 'territory', and any map is merely provisional.

[12] This involved work of John Conway, Simon P. Norton, Igor Frenkel, James Lepowsky, Arne Meurman and Richard Borcherds.

Chapter 22
El Aleph, Or a Monster Lurks in the Belly of Computer Science

Francisco Antonio Doria

The shadow of that hideous strength
six miles and more it is of length.
Sir David Lindsay
apud C. S. Lewis

A Monster First Uncovered

Act 1.

In 1962 Tibor Radó uncovered a monster hiding in the midst of innocent-looking Turing machines.

Yes, this is what came to my mind when first I read Radó's 1962 paper, "On non-computable functions," published in the Bell System Technical Journal, vol. 41. Radó's paper begins with an innocent-looking presentation, a game played with Turing machines. Slowly the argument unfolds, and quite suddenly we reach a climax, which is as dissonant and unexpected as the big climax near the end of the *Adagio* in Bruckner's Ninth Symphony, or the crazy chord in the midst of Mahler's Tenth Symphony, first movement: for out of the nice, computable functions we were dealing with, we find out that the game we were playing is described by a non-computable function, an intractable function that moreover dominates all total recursive—computable—functions; a function which nevertheless can be very easily described. An aberration, easily describable but intractable.

F. A. Doria (✉)
Advanced Studies Research Group, HCTE, Fuzzy Sets Laboratory,
Mathematical Economics Group, Production Engineering Program, COPPE, UFRJ, 68507,
Rio de Janeiro 21945-972, Brazil
e-mail: fadoria63@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_22

An aberration, as it can be very precisely characterized, and yet is non-computable, as it tops all computable functions; it is non-computable because it grows too fast. Fast beyond the tools one uses to describe fast-growing computable functions.
Act 2.

Around 1992 Newton da Costa and I decided to try our hand at the *P vs. NP* problem. We were puzzled, like everybody: why is it so difficult? After all it begins in a simple, innocent-looking question, which I now quote as if telling a short tale:

> Mrs. H is a gentle and able lady who has long been the secretary of a large university department. Every semester Mrs. H is confronted with the following problem: there are courses to be taught, professors to be distributed among different classes of students, large and small classes, and a shortage of classrooms. She fixes a minimum acceptable level of overlap among classes and students and sets down in a tentative way to get the best possible schedule given that minimum desired overlap. It's a tiresome task, and in most cases, when there are many new professors or when the dean changes the classroom allocation system, Mrs. H has to redo everything again; again she has to check nearly all conceivable schedulings before she is able to reach a conclusion. In despair she asks a professor whom she knows has a degree in math: *"tell me, can't you find in your math a fast way of scheduling our classes with a minimum level of overlap among them?"*

Mrs. H unknowingly asks about the *P vs. NP* question. She is able to understand its basic contours as she aptly summarizes it: *there are questions for which it is easy to check whether a given arrangement of data fits in as a solution; however for the general case there are no known shortcuts in order to reach a solution.*

Where do we now begin our efforts? My colleague da Costa and I had a hunch from start: undecidability. The Gödel and Turing phenomenon. It assuredly plays a role—some role—in the *P vs. NP* question. More precisely, the *P vs. NP* question leads to formally undecidable sentences, we suppose. But how are we going to proceed in order to settle it?

Let us now take a closer look at the issue. Mrs. H noticed in her comments the chief characteristics (with some license) of problems in the class *NP*:

> *there are questions for which it is easy to check whether a given arrangement of data fits in as a solution; however for the general case there are no known shortcuts in order to reach a solution.*

Or:

> *These are problems where it is hard to find a solution, that is, no known shortcuts are efficient for all cases, while once we have a solution, it can be very easily tested.*

Easily tested, we add, means that we have a time-polynomial bounded algorithm ("poly algorithms," for short) that performs the testing. And hard to obtain in all known cases, means that only time-bounded exponential algorithms are known for the solution of problems in the *NP* class (we might say, "expo algorithms").

And Mrs. H's question becomes:

> *Are there poly algorithms that settle all problems in the NP class?*

We will translate that question as $P = NP$? (please wait a bit; I'll give the formal definition soon).

Joseph K Enters the Fray

We had a problem which could be phrased in a rigorous, precise, formal language. And we started by collecting folklore-like facts about it. We knew that the statement that $P = NP$ could be formulated as an arithmetic Σ_2 sentence. Then its negation $P < NP$ is a Π_2 arithmetic sentence (see the Appendix, section "Technicalities").

And folklore begins to pile up. Waving our hands a bit we have the easy but important results:

- If $P = NP$ is true, that is, if there is a fast algorithm to solve all problems in a NP set of problems, then $P = NP$ can be proved either by Peano Arithmetic (PA) or by a simple extension of PA, namely one with the same alphabet as PA, the same underlying language—classical first-order predicate calculus—and the same proof strength as PA, since the extended theory would have the same provably total recursive functions as PA.[1]
- If both $P = NP$ and $P < NP$ are independent of PA (supposed consistent), or even of a stronger theory S, then $P < NP$ holds true of the standard model for arithmetic, provided that PA or its extension S have one such model.
  That is to say, independence means that $P < NP$ is true.

We stalled. How do we go from here?

At this point something unexpected happens: Mr. Joseph K steps in.[2]

Mr. K is a sage with a prankster-like humor. He tells us, or better, warns us:

There is a monster lurking in the belly of the P vs. NP question.

The monster, he added, is a Busy Beaver like function. With the following essential characteristic: once you explicitly build that function, you settle the matter. The monster, he concluded, is the counterexample function to $P = NP$: it is a total (albeit noncomputable) function if and only if $P < NP$, and it is a partial function if and only if $P = NP$.

We consider a specific problem, say, SAT, the Boolean satisfiability problem. And with the help of SAT, here are the looks of the monster:

- List all poly Turing machines in the order induced by the listing by their usual Gödel numbers. (Their indices; this listing is of course nonrecursive.)
- For each poly machine P_n of Gödel number n get the Gödel number of the first instance of SAT which is input to it, and fails to output a solution for that instance; then add 1.

We just sketched the counterexample function, noted f, to a class of problems in NP. (Too vague? Again: please wait for the "Technicalities" section.) Such a function, Mr. K told us, has two remarkable properties, we insist:

[1] It isn't very elegant to repeat "PA" as I've done here, but I want to avoid ambiguities; it is inelegance for the sake of clarity.

[2] Joseph K is Georg Kreisel, with whom da Costa and Doria corresponded in 1993–1995.

- It is a total function if and only if $P < NP$ holds; it is partial if and only if $P = NP$.
- The function just described zigzags a lot. But if we isolate its peaks, we see that our monster overtakes all total recursive functions, infinitely many times. That is, its envelope function grows at least as the Busy Beaver.

The secret of the *P vs. NP* question is thus coded in a Busy Beaver like monster, concluded Mr. Joseph K with a subtle, ironic grin. However he gives us no proof of that crucial fact; and we started looking for it. Doing our homework, so to say.

And so we did: we didn't know how to describe the whole of f, but soon learned that it was possible to describe some of those "peaks" in f with the help of quite simple Turing machines.

Next slide, please!

An Invisible Monster?

Suppose that we want to build a function that tops all total recursive functions within some axiomatic frame. Let's do it formally. We will work within a formal theory S that "looks like" Peano Arithmetic (PA) or axiomatic set theory (ZFC; details below). We get:

Remark 1 For each n, $\mathsf{F}(n) = \max_{k \leq n}(\{e\}(k)) + 1$, that is the sup of those $\{e\}(k)$ such that:

1. $k \leq n$.
2. $\ulcorner \mathrm{Pr}_S(\ulcorner \forall x\, \exists z\, T(e, x, z) \urcorner) \urcorner \leq n$.

$\mathrm{Pr}_S(\ulcorner \xi \urcorner)$ means, there is a proof of ξ in S, where $\ulcorner \xi \urcorner$ means: the Gödel number of ξ. So $\ulcorner \mathrm{Pr}_S(\ulcorner \xi \urcorner) \urcorner$ means: "the Gödel number of sentence 'there is a proof of ξ in S.'" Condition 2 above translates as: there is a proof of [$\{e\}$ is total] in S whose Gödel number is $\leq n$. □

Proposition 2 *We can explicitly compute a Gödel number* e_{F} *so that* $\{e_{\mathsf{F}}\} = \mathsf{F}$. □

Proposition 3 *If S is consistent then* $S \nvdash \forall m \exists n\, [\{e_{\mathsf{F}}\}(m) = n]$. □

We do *not* get a Busy Beaver like function; actually we get a partial recursive function (and the Busy Beaver is noncomputable…) which can neither be proved nor disproved total in S—it is total in the standard model for arithmetic, provided that S has a model with standard arithmetic.

A related question: how do we mirror Busy Beaver like functions within a formal theory like S? Can we do it? Can we consider its relevant properties within our axiomatic system S?

El Aleph, or Properties of the Counterexample Function

…pues en un ángulo
del sótano había un Aleph.
Aclaró que un Aleph es uno de los puntos del espacio que contiene todos los puntos.

…because down in the cellar there was an Aleph.
He explained that an Aleph
is one of the points in space that contains all other points.[3]
J. L. Borges, *El Aleph*

f is our very Borgian El Aleph. But we soon learn that f has infinitely many avatars f^k in each theory S—moreover, and quite surprisingly, we see that the f^k are recursive, and defined with respect to sets noted BGS^k, which are recursive sets of poly machines.[4] How is that possible?

f and the infinitely many f^k are very peculiar objects.

They are fractal-like in several senses; say, in the following specific sense: the essential data about *NP*-complete questions is reproduced mirror-like in each of the f^k (or over each BGS^k). The different BGS^k are distributed over the set of all Turing machines by the primitive recursive function $\mathsf{c}(m, k, a)$.

Let us take a closer look at that quite unexpected property. Let g be a partial recursive function. Its graph can be written as $\langle n, \mathsf{g}(n)\rangle$, $n \in \omega$. Then there is a primitive recursive c that depends on the Gödel number of g so that we map:

$$\langle n, \mathsf{g}(n)\rangle \text{ onto } \langle \mathsf{c}(n), f(\mathsf{c}(n))\rangle.$$

That is to say, any partial recursive function is reproduced in a segment of f. Now consider the usual axiomatic systems, like PA or ZFC. They can be given (under an adequate coding for its formal sentences) as a Turing machine that outputs all theorems of the theory. As a consequence:

> *all axiomatic theories that have a recursively enumerable set of theorems and which include arithmetic (at least rules for* $+$ *and* $\times$ *plus the trichotomy axiom) can be mapped within* f*.*

Another fractal-like behavior: we may start from f and embed into it the theory coded by g, and then into that theory we can embed the theory coded in $\mathsf{g}_{(2)}$ and so on, indefinitely. And of course, given f (the BGS partial recursive function that reproduces f) we can code in various ways f into itself and into f. And so on.

In a nutshell: f copies itself in its own belly. Infinitely many times.

Yet we cannot argue within S that for all k, f_k dominates …, as that would imply the totality of the recursive function F_S.

[3] For the translation, http://www.phinnweb.org/links/literature/borges/aleph.html.

[4] See the Appendix for definitions and technical details.

Still More Intuitions

It is interesting to always keep in mind a picture of these objects. First notice that the BGS and BGS^k machines are interspersed among the Turing machines. The quasi-trivial Turing machines have their Gödel numbers given by the primitive recursive function $\mathsf{c}(k, n)$—we forget about he other parameters—where:

- k refers to f^k and to BGS^k as already explained;
- n is the argument in $\mathsf{f}^k(n)$.

So, fast-growing function f^k is sort of cloned among the values of the BGS^k counterexample function while slightly slowed down by c. (Recall that c is primitive recursive, and cannot compete in growth power with the f^k.)

Function f^k compresses what might be a very large number into a small code given by the Gödel number of g^k and by n (recall that the length of the numeral $\mathsf{f}^k(n)$ is the order of $\log \mathsf{f}^k(n)$). The effect is that all functions f^j, $j < k$ embedded into the k-counterexample function via our quasi-trivial machines keep their fast-growing properties and allow us to prove that the counterexample function is fast-growing in its peaks for BGS^k.

For $j > k$ the growth power of f^k doesn't compensate the length of the parameters in the bounding polynomial that regulates the coupled clock in the BGS^k machines.

Finally while $j < k$, the compressed Gödel numbers of the quasi-trivial machines—they depend on the exponent and constant of the polynomial $x^{\mathsf{f}^k(n)} + \mathsf{f}^k(n)$ which regulates the clock—grow much slower that the growth rate of the counterexample function over these quasi-trivial machines (depending on f^j) and so their fast growing properties come out clearly (For details see the appendix).

Two questions here: can we prove that the Busy Beaver is total, within ZFC? Similarly, suppose that the counterexample function f is (naïvely) total. Then can we prove it within ZFC? Within $S \supset$ ZFC? Notice that g can be mapped onto a segment of f. But we must thread very carefully here.

Appendix

Technicalities Galore

We now definitively change the language to best operate the weapons we have to try to tame, or at least to control, our monster functions. And we will now dive into the obscurities of a formal language. Portions of these technical details have already been presented in Costa and Doria (2016).

We deal here some possible formalizations for $P = NP$ and $P < NP$; we have called the unusual formalizations (there are infinitely many) the "exotic formalizations." They are intuitively equivalent, but if we formalize things there are difficulties to be dealt with when trying to establish their equivalence.

Remark 1 The reason we actually have such a plethora of (not always equivalent) definitions has to do with the fact that when we say we have a polynomial bound we are talking about *some* polynomial bound, and not, say, a minimal bound. That apparently minor fact is the source of many complications, as we will soon see. □

Let $t_m(x)$ be the primitive recursive function that gives the operation time of $\{m\}$ over an input x of length $|x|$.

Recall that the operation time of a Turing machine is given as follows: if $\{m\}$ stops over an input x, then:

$$t_m(x) = |x| + [\text{number of cycles of the machine until it stops}].$$

t_m is primitive recursive and can in fact be defined out of Kleene's T predicate.

Definition 2 (STANDARD FORMALIZATION FOR $P = NP$.)
$[P = NP] \leftrightarrow_{\text{Def}} \exists m, a \in \omega \, \forall x \in \omega \, [(t_m(x) \leq |x|^a + a) \wedge R(x, m)]$. □

$R(x, y)$ is a polynomial predicate; as its interpretation we can say that it formalizes a kind of "verifying machine" that checks whether or not x is satisfied by the output of $\{m\}$. (There is an equivalent formalization for $[P = NP]$ where again one uses Kleene's T predicate to get the time measure t_m.)

Definition 3 $[P < NP] \leftrightarrow_{\text{Def}} \neg[P = NP]$. □

Now suppose that $\{e_{\mathsf{f}}\} = \mathsf{f}$ is total recursive and strictly increasing:

Remark 4 The naïve version for the exotic formalization is:

$$[P = NP]^{\mathsf{f}} \leftrightarrow \exists m \in \omega, a \, \forall x \in \omega \, [(t_m(x) \leq |x|^{\mathsf{f}(a)} + \mathsf{f}(a)) \wedge R(x, m)].$$

However as we will soon see, there is no reason why we should ask that f be total; on the contrary, there will be interesting situations where such a function may be partial and yet it may provide a reasonable exotic formalization for $P < NP$ (Guillaume 2003). □

So, for the next definitions and results let f be in general a (possibly partial) recursive function which is strictly increasing over its domain, and let e_{f} be the Gödel number of an algorithm that computes f. Let $p(\langle e_{\mathsf{f}}, b, c\rangle, x_1, x_2, \ldots, x_k)$ be an universal Diophantine polynomial with parameters e_{f}, b, c; that polynomial has integer roots if and only if $\{e_f\}(b) = c$. We may if needed suppose that polynomial to be ≥ 0. We omit the "$\in \omega$" in the quantifiers, since they all refer to natural numbers.

Definition 5 $M_{\mathsf{f}}(x, y) \leftrightarrow_{\text{Def}} \exists \, x_1, \ldots, x_k \, [p(\langle e_{\mathsf{f}}, x, y\rangle, x_1, \ldots, x_k) = 0]$. □

Actually $M_{\mathsf{f}}(x, y)$ stands for $M_{e_{\mathsf{f}}}(x, y)$, or better, $M(e_{\mathsf{f}}, x, y)$, as dependence is on the Gödel number e_{f}.

Definition 6 $\neg Q(m, a, x) \leftrightarrow_{\text{Def}} [(t_m(x) \leq |x|^a + a) \to \neg R(x, m)]$. □

Proposition 7 (STANDARD FORMALIZATION, AGAIN.)

$$[P < NP] \leftrightarrow \forall m, a\, \exists x\, \neg Q(m, a, x). \qquad \square$$

Definition 8 $\neg Q_{\mathsf{f}}(m, a, x) \leftrightarrow_{\text{Def}} \exists a'\, [M_{\mathsf{f}}(a, a') \wedge \neg Q(m, a', x)].$ □

Remark 9 We will sometimes write $\neg Q(m, \mathsf{f}(a), x)$ for $\neg Q_{\mathsf{f}}(m, a, x)$, whenever f is—safely, in some sense—total. □

Definition 10 (EXOTIC FORMALIZATION.)

$$[P < NP]^{\mathsf{f}} \leftrightarrow_{\text{Def}} \forall m, a\, \exists x\, \neg Q_{\mathsf{f}}(m, a, x). \qquad \square$$

Notice that again this is a Π_2 arithmetic sentence:

$$\forall m, a\, \exists x, a', x_1, \ldots, x_k\, \{[p(\langle e_{\mathsf{f}}, a, a' \rangle, \ldots, x_1, \ldots, x_k) = 0] \wedge \neg Q(m, a', x)\}.$$

(Recall that Q is primitive recursive.)

Definition 11 $[P = NP]^{\mathsf{f}} \leftrightarrow_{\text{Def}} \neg[P < NP]^{\mathsf{f}}.$ □

The Monster Shows Glimpses of Its Face

For the definition of SAT see Machtey and Young (1979); for the BGS recursive set of poly Turing machines see Baker et al. (1975). In a nutshell, SAT is the set of all Boolean expressions in conjunctive normal form (cnf) that are satisfiable, and BGS is a recursive set of poly Turing machines that contains emulations of every conceivable poly Turing machines.

The full counterexample function f is defined as follows; let ω be also a set of codes for an enumeration of the Turing machines (see on what we mean by "standard coding," Mendelson 1997, p. 320ff). Similarly we code by an analogous standard code SAT onto ω:

- If $n \in \omega$ isn't a poly machine, $f(n) = 0$.
- If $n \in \omega$ codes a poly machine:
  - $f(n)$ = first instance x of SAT so that the machine fails to output a satisfying line for x, plus 1, that is, $f(n) = x + 1$.
  - Otherwise $f(n)$ is undefined, that is, if $P = NP$ holds for $n, f(n)$ = undefined.

As defined, f is non computable. It will also turn out to be at least as fast growing as the Busy Beaver function, since in its peaks it tops all intuitively total recursive functions.

The idea in the proof of that fact goes as follows:

- Use the *s–m–n* theorem to obtain Gödel numbers for an infinite family of "quasi-trivial machines"—soon to be defined. The table for those Turing machines involves very large numbers, and the goal is to get a compact code for that value in each quasi-trivial machine so that their Gödel numbers are in a sequence $\mathsf{g}(0), \mathsf{g}(1), \mathsf{g}(2), \ldots$, where g is primitive recursive.
- Then add the required clocks as in the BGS sequence of poly machines, and get the Gödel numbers for the pairs machine + clock. We can embed the sequence we obtain into the sequence of all Turing machines.
- Notice that the subsets of poly machines we are dealing with are (intuitive) recursive subsets of the set of all Turing machines. More precisely: if we formalize everything in some theory S, then the formalized version of the sentence "the set of Gödel numbers for these quasi-trivial Turing machines is a recursive subset of the set of Gödel numbers for Turing machines" holds of the standard model for arithmetic in S, and vice versa.
  However S may not be able to prove or disprove that assertion, that is to say, it will be formally independent of S.
- We can thus define the counterexample functions over the desired set(s) of poly machines, and compare them to fast-growing total recursive functions over similar restrictions.

Recall:

Definition 12 For $f, g : \omega \to \omega$,

$$f \textbf{ dominates } \mathrm{g} \leftrightarrow_{\mathrm{Def}} \exists \mathrm{y}\, \forall \mathrm{x}\, (\mathrm{x} > \mathrm{y} \to \mathrm{f}(\mathrm{x}) \geq \mathrm{g}(\mathrm{x})).$$

We write $f \succ g$ for f dominates g. □

Quasi-trivial Machines

Recall that the operation time of a Turing machine is given as follows: if M stops over an input x, then the operation time over x,

$$t_{\mathsf{M}} = |x| + \text{number of cycles of the machine until it stops.}$$

Example 13

- **First trivial machine**. Note it O. O inputs x and stops.

$$t_{\mathsf{O}} = |x| + \text{moves to halting state } + \text{stops.}$$

So, operation time of O has a linear bound.

- **Second trivial machine**. Call it O'. It inputs x, always outputs 0 (zero) and stops. Again operation time of O' has a linear bound.
- **Quasi-trivial machines**. A *quasi-trivial machine* Q operates as follows: for $x \leq x_0$, x_0 a constant, fixed value, $\mathsf{Q} = \mathsf{R}$, R an arbitrary total machine. For $x > x_0$, $\mathsf{Q} = \mathsf{O}$ or O'.
  This machine has also a linear bound. □

Remark 14 Now let H be any fast-growing, superexponential total machine. Also let H' be a total Turing machine. Form the following family $\mathsf{Q}_{\dots}$ of quasi-trivial Turing machines with subroutines H and H':

1. If $x \leq \mathsf{H}(n)$, $\mathsf{Q}^{\mathsf{H},\mathsf{H}',n}(x) = \mathsf{H}'(x)$;
2. If $x > \mathsf{H}(n)$, $\mathsf{Q}^{\mathsf{H},\mathsf{H}',n}(x) = 0$. □

Proposition 15 *There is a family* $\mathsf{R}_{\mathsf{g}(n,|\mathsf{H}|,|\mathsf{H}'|)}(x) = \mathsf{Q}^{\mathsf{H},\mathsf{H}',n}(x)$, *where* g *is primitive recursive, and* $|\mathsf{H}|, |\mathsf{H}'|$ *denotes the Gödel number of* H *and of* H'.

Proof By the composition theorem and the *s–m–n* theorem. □

Remark 16 Very important! We are interested in quasi-trivial machines where $\mathsf{H}' = \mathsf{T}$, the standard truth-table exponential algorithm for SAT. □

Notice that, for the counterexample function when defined over all Turing machines (with the extra condition that the counterexample function = 0 if M_m isn't a poly machine), we have:

Proposition 17 *If* $\mathsf{g}(n)$ *is the Gödel number of a quasi-trivial machine as in Remark 14, then* $f(\mathsf{g}(n)) = \mathsf{H}'(n) + 1$.

Proof Use the machines in Proposition 15 and Remark 16. □

That Hideous Strength …

Our goal here is to prove the following result: *no total recursive function dominates f*.

Remark 18 We sketch below the idea of the proof. Suppose that there is a total recursive function $\mathsf{h}(n)$ that dominates f. Get a total recursive $\mathsf{k}(n)$ that dominates h and so that the relative growth speed of k with respect to h is faster that any primitive recursive function.

Why do we need such a condition? We use the quasi-trivial machines to reproduce k within f, that is, we (sort of) replicate function $\langle n, \mathsf{k}(n)\rangle$ within f by a sequence of machines with Gödel numbers $N(n), n = 0, 1, 2, \dots$ (see above Proposition 17), where N is primitive recursive, so that we have that k becomes the sequence of machines $N(n), n = 0, 1, 2, \ldots$, and we get the value of f at k as $\langle N(n), \mathsf{k}(n) + 1\rangle$ with $f(N(n)) = \mathsf{k}(n) + 1$.

As N can be taken to be primitive recursive, monotonic increasing on n, it slows down the growth of k by a primitive recursive function. Given our construction—which is trivially fulfilled—we have that f still overtakes h infinitely many times, as k grows faster than h, and we are done. □

One side comment: f is extremely complex, as it contains infinitely many copies of itself (if we suppose that f is total). Just for starters… Mr. K pointed to us the fast growing property of f but gave no proof of the fact. When we (da Costa and I) first managed to sketch a proof of f's fast growing behavior, I was shocked, and sent an email to K, with a comment, "I would never expect that out of a pedestrian, seemingly naïve question one would have to confront such a crazy growth."

He answered me, like Darth Vader explaining the rituals surrounding the Force to Luke Skywalker, "this kind of problem has a sacred status, and should only be approached by very few people, by the high priests of the cult."

That is, I was a heretic on the verge of facing the stake.

Proposition 19 *For no total recursive function* h *does* $\mathsf{h} \succ f$.

Proof Suppose that there is a total recursive function h such that $\mathsf{h} \succ f$.
Notice:

- Given such a function h, we can obtain another total recursive function h' which satisfies:

  1. h' is strictly increasing.
  2. For $n > n_0$, $\mathsf{h}'(n) > \mathsf{h}(\mathsf{g}(n))$, with g as in Proposition 17. □

- Given a total recursive h, there is a total recursive h' that satisfies the previous conditions.

For given h, we obtain out of that total recursive function by the usual constructions a strictly increasing total recursive h^*. Then if, for instance, F_ω is Ackermann's function, $\mathsf{h}' = \mathsf{h}^* \circ \mathsf{F}_\omega$ will do. (The idea is that F_ω dominates all primitive recursive functions, and therefore h^* composed with it dominates $\mathsf{g}(n)$.)

We have that the Gödel numbers of the quasi-trivial machines Q are given by $\mathsf{g}(n)$. Choose adequate quasi-trivial machines, so that $f(\mathsf{g}(n)) = \mathsf{h}'(n) + 1$, from Proposition 17. We now conclude our argument. If we make explicit the computations, for $\mathsf{g}(n)$ (as the argument holds for any strictly increasing primitive recursive g):

$$f(\mathsf{g}(n)) = \mathsf{h}'(n) + 1 = \mathsf{h}^*(\mathsf{F}_\omega(n)) + 1,$$

and

$$\mathsf{h}^*(\mathsf{F}_\omega(n)) > \mathsf{h}^*(\mathsf{g}(n)).$$

For $N = \mathsf{g}(n)$,

$$f(N) > \mathsf{h}^*(N) \geq \mathsf{h}(N), \text{ all } N.$$

Therefore no such h can dominate f. □

Corollary 20 *No total recursive function dominates f.* □

BGS-Like Sets

We use here the BGS (Baker et al. 1975) set of poly machines:

$$\langle \mathsf{M}_m, |x|^a + a \rangle,$$

where we couple a Turing machine M_m to a clock regulated by the polynomial $|x|^a + a$, that is, it stops M_m after $|x|^a + a$ steps in the operation over x, where x is the machine's binary input and $|x|$ its bit-length.

A more general machine-clock couple will also be dealt with here:

$$\langle \mathsf{M}_m, |x|^{(a)} + \mathsf{f}(a) \rangle \mapsto \mathsf{M}_{\mathsf{c}(m, |\mathsf{f}|, a)},$$

Its Gödel number is given by $\mathsf{c}(m, |\mathsf{f}|, a)$, with c primitive recursive by the *s–m–n* theorem.

Remark 21 Notice that we can have c such that, for parameters a, b, if $a < b$, then $\mathsf{c}(\ldots a \ldots) < \mathsf{c}(\ldots b \ldots)$. □

$P < NP$ is given by a Π_2 arithmetic sentence, that is, a sentence of the form "for every x there is an y so that $P(x, y)$," where $P(x, y)$ is a very simple kind of relation.[5] Now given a theory S with enough arithmetic in it, S proves a Π_2 sentence ξ if and only if the associated Skolem function f_ξ is proved to be total recursive by S. For $P < NP$, the Skolem function is what we have been calling the counterexample function.

However there are infinitely many counterexample functions we may consider, an *embarras de choix*, as they say in French. Why is it so? For many adequate, reasonable theories S, we can build a recursive (computable) *scale of functions*[6] $\mathsf{f}_0, \mathsf{f}_1, \ldots, \mathsf{f}_k, \ldots$ with an infinite set of S-rovably total recursive functions so that f_0 is dominated by f_1 which is then dominated by $\mathsf{f}_2, \ldots,$ and so on, up to the corresponding function F_S.

Given each function f_k, we can form a BGS-like set BGS^k, where clocks in the time-polynomial Turing machines are bounded by a polynomial:

$$|x|^{\mathsf{f}_k(n)} + \mathsf{f}_k(n),$$

[5] It is a primitive recursive predicate.

[6] Such a "scale of functions" exists and can be explicitly constructed.

where $|x|$ denotes the length of the binary input x to the machine. We can then consider the recursive set:

$$\bigcup_k \mathrm{BGS}^k$$

of all such sets.

Each BGS^k contains representatives of all poly machines (time polynomial Turing machines). Now, what happens if:

- There is a function g which is total provably recursive in S and which dominates all segments f_k of counterexample functions over each BGS^k.
- There is no such an g, but there are functions g_k which dominate each particular f_k, while the sequence $\mathsf{g}_0, \mathsf{g}_1, \ldots$ is unbounded in S, that is, grows as the sequence $\mathsf{F}_0, \mathsf{F}_1, \ldots$ in S?

We will take a look into these queries.

Exotic BGS$^{\mathsf{F}}$ Machines

Now let F be a fast growing, intuitively total, algorithmic function. We consider exotic $\mathrm{BGS}^{\mathsf{F}}$ machines, that is, poly machines coded by the pairs $\langle m, a\rangle$, which code Turing machines M_m with bounds $|x|^{\mathsf{F}(a)} + \mathsf{F}(a)$. Since the bounding clock is also a Turing machine, now coupled to M_m, there is a primitive recursive map c so that:

$$\langle \mathsf{M}_m, |x|^{\mathsf{F}(a)} + \mathsf{F}(a)\rangle \mapsto \mathsf{M}_{\mathsf{c}(m,|\mathsf{F}|,a)},$$

where $\mathsf{M}_{\mathsf{c}(m,|\mathsf{F}|,a)}$ is a poly machine within the sequence of all Turing machines. We similarly obtain a g as above, and follows:

Proposition 22 *Given the counterexample function* f_k *defined over the* BGS^k*-machines, for no* ZFC*-provable total recursive* h *does* $\mathsf{h} > \mathsf{f}_k$.

Proof As in Proposition 19; use Gödel number coding primitive recursive function c to give the Gödel numbers of the quasi-trivial machines we use in the proof. □

Remark 23 Notice that we have a perfectly reasonable formalization for our big question:

$$[P < NP]^k \leftrightarrow [P < NP]^{\mathsf{f}^k}.$$

Also, $S \vdash [P < NP]^k \leftrightarrow [\mathsf{f}_c^k$ is total]. So our analysis will give estimates for the growth rate of each counterexample function f_c^k. □

Remark 24 The previous statements have interesting consequences, which we will briefly pursue below. For the proof of the proposition choose a BGS^k so that f_k dominates all strictly increasing fast growing provably total recursive functions that eventually appear in the proof. □

We can state, for total f_c^k:

Proposition 25 *For each j there is a k, $k > j + 1$, so that S proves the sentence "f_k doesn't dominate the* BGS^k *counterexample function f_c^k."* □

However we cannot conclude that "for all *j*, we have that..." since that would imply that *S* proves "for all *j*, f_j is total" as a scholium, which cannot be done (as that is equivalent to "F_S is total," which again cannot be proved in *S*).

What can be concluded: let S' be the theory $S + \mathsf{F}_S$ is total. Then:

Proposition 26 *If S is consistent and if f_c^k is total in a model with standard arithmetic for each k, then S' proves: there is no proof of the totality of f_c^k, any k, in S.*

Proof See below the discussion. □

Remark 27 Notice that:

- $S' \vdash \forall k\,([P < NP]^k \leftrightarrow [\mathsf{f}_c^k$ is total]), while *S* cannot prove it.
- $S' \vdash \forall k\,([P < NP]^k \leftrightarrow [P < NP])$ while again *S* cannot prove it.
- S' is $S + [S$ is Σ_1- sound]. □

Remark 28 That means that we can conclude:

> *S' proves that, for every k, S cannot prove $[P < NP]^k$.*

Now : does the $[P < NP]^k$ adequately translate our main question? □

Remark 29 Notice that theory $S + \mathsf{F}_S$ is total is $S + S$ is Σ_1-sound. This will have further consequences. □

References

T. Baker, J. Gill, R. Solovay, Relativizations of the *P =?NP* question. SIAM J. Comp. **4**, 431–442 (1975)

S. Ben–David, S. Halevi, On the independence of *P vs. NP*, Technical Report # 699, Technion (1991)

N.C.A. da Costa, F.A. Doria, Undecidability and incompleteness in classical mechanics. Int. J. Theor. Phys **30**, 1041–1073 (1991)

N.C.A. da Costa, F.A. Doria, Suppes predicates and the construction of unsolvable problems in the axiomatized sciences, in *Patrick Suppes, Scientific Philosopher*, ed. by P. Humphreys, vol. II, pp. 151–191 (Kluwer, 1994)

N.C.A. da Costa, F.A. Doria, Consequences of an exotic formulation for $P = NP$. Appl. Math. Comput. **145**, 655–665 (2003); also Addendum. Appl. Math. Comput. **172**, 1364–1367 (2006)

N.C.A. da Costa, F.A. Doria, Computing the future, in *Computability, Complexity and Constructivity in Economic Analysis*, ed. by K. Vela Velupillai (Blackwell, 2005)

N.C.A. da Costa, F.A. Doria, On the O'Donnell algorithm for *NP*-complete problems. Rev. Behav. Econ. **3**, 221–242 (2016)

N.C.A. da Costa, F.A. Doria, E. Bir, On the metamathematics of the *P vs. NP* question. Appl. Math. Comput. **189**, 1223–1240 (2007)

M. Davis, Hilbert's Tenth Problem is unsolvable, in *Computability and Unsolvability* (Dover, 1982)

R.A. DeMillo, R.J. Lipton, Some connections between computational complexity theory and mathematical logic, in *Proceedings of 12th Annual ACM Symposium on the Theory of Computing* (1979), pp. 153–159

R.A. DeMillo, R.J. Lipton, The consistency of $P = NP$ and related problems with fragments of number theory, in *Proceedings of 12th Annual ACM Symposium on the Theory of Computing* (1980), pp. 45–57

F.A. Doria, Is there a simple, pedestrian, arithmetic sentence which is independent of ZFC? *Synthèse* **125** (1/2), 69 (2000)

F.A. Doria, Informal vs. formal mathematics. Synthèse **154**, 401–415 (2007)

S. Fortune, D. Leivant, M. O'Donnell, The expressiveness of simple and second-order type structures. J. ACM **38**, 151–185 (1983)

T. Franzen, Transfinite progressions: a second look at completeness. Bull. Symb. Log. **10**, 367–389 (2004)

H.M. Friedman, Finite functions and the necessary use of large cardinals. Ann. Math. **148**, 803 (1998)

M. Guillaume, What counts in an exotic formulation..., unpublished draft (2003)

J. Hartmanis, J. Hopcroft, Independence results in computer science. SIGACT News, 13 (1976)

H. Hermes, *Enumerability, Decidability, Computability* (Springer, 1965)

J.E. Hopcroft, J.D. Ullman, *Formal Languages and their Relation to Automata* (Addison–Wesley, 1969)

D. Joseph, P. Young, Independence results in computer science?, in *Proceedings of 12th Annual ACM Symposium on the Theory of Computing* (1980), pp. 58–69

D. Joseph, P. Young, Fast programs for initial segments and polynomial time computation in weak models of arithmetic. STOC Milwaukee **1981**, 55–61 (1981)

R. Kaye, *Models of Peano Arithmetic* (Clarendon Press, 1991)

S.C. Kleene, General recursive functions of natural numbers. Math. Ann. **112**, 727 (1936)

S.C. Kleene, *Mathematical Logic* (Wiley, 1967)

W. Kowalczyk, A sufficient condition for the consistency of $P = NP$ with Peano Arithmetic. Fund. Inform. **5**, 233–245 (1982)

G. Kreisel, On the interpretation of non-finitist proofs. I, J. Symb. Log. **16**, 241 (1951); II, **17**, 43 (1952)

G. Kreisel, On the concepts of completeness and interpretation of formal systems. Fund. Math. **39**, 103–127 (1952)

M. Machtey, P. Young, *An Introduction to the General Theory of Algorithms* (North–Holland, 1979)

A. Maté, Nondeterministic polynomial-time computations and models of arithmetic. J. ACM **37**, 175–193 (1990)

E. Mendelson, *Introduction to Mathematical Logic* (Chapman & Hall, 1997)

M. O'Donnell, A programming language theorem which is independent of Peano Arithmetic, in *Proceedings of 11th Annual ACM Symposium on the Theory of Computation* (1979), pp. 176–188

J. Paris and L. Harrington, "A mathematical incompleteness in Peano arithmetic, in *Handbook of Mathematical Logic*, ed. by J. Barwise (North–Holland, 1977)

T. Rado, On non-computable functions. Bell Syst. Tech. J. **91**, 877 (1962)

H. Rogers Jr., *Theory of Recursive Functions and of Effective Computability*, reprint (MIT Press, 1992)

S. Shapiro, Incompleteness, mechanism and optimism. Bull. Symb. Log. **4**, 273–302 (1998)

C. Smorýnski, *Logical Number Theory*, I (Springer, 1991)

J. Spencer, Large numbers and unprovable theorems. Am. Math. Monthly **90**, 669 (1983)

Chapter 23
Two Algorithms for *NP*-Complete Problems and Their Relevance to Economics

C. A. Cosenza and Francisco Antonio Doria

Introduction

Maps and territory suggest problems which have to do with the opening of pathways in some poorly explored domain. We can perhaps recall Heidegger's *Holzwege*—pathways that sometimes lead nowhere, within some lost backwoods—or we can remember something more specific, such as K. Menger's *Das Botenproblem*, which is now best known as "the traveling salesman problem," where one looks for some specific path within a miriad, sort of inadequate, ones.

Well, let's be more specific: can we find semantics for a given family of mathematical problems out of its solution procedures, like the collection of algorithms that settle them? Are they very diverse? Can we get unity out of diversity?

We present and describe here two algorithms for *NP*-complete problems, the O'Donnell algorithm, which is an exact one, and the so-called COPPE–COSENZA approximate procedure for allocation problems. Our sources are (Cosenza 1981; da Costa and Doria 2016), which we liberally quote in our characterization of those algorithms. They are wildly different: the COPPE–COSENZA technique arises out of concrete examples, and follows an intuitive path up to a step where we make an approximation that works quite well for the great majority of cases. In contrast, O'Donnell's algorithm is highly abstract from the beginning, and requires examination of the very innards of the Big Question it is related to.

But our query here is a kind of collateral damage: these computational recipes are—so to say—very different in their construction. Can we develop a semantics for the *NP*-class of problems that "locally" (in some sense of locality, say some

C. A. Cosenza (✉) · F. A. Doria
Advanced Studies Research Group and Fuzzy Sets Laboratory, PEP-COPPE, UFRJ,
68507, Rio RJ 21945-972, Brazil
e-mail: cosenza@pep.ufrj.br

F. A. Doria
e-mail: fadoria63@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_23

restriction on our set of problems) reduces to either O'Donnell's or the COPPE–COSENZA procedure?

Let's take a look at our examples.

A Brief Review of *NP*-Completeness in Economics

While we can trace the *traveling salesman problem* back to Sir William Rowan Hamilton in the 19th century, the first formulation of TSP is usually credited to (Karl Menger in 1932). Menger's formulation of the TSP is quite straightforward[1]:

> We call the Messenger Problem [. . .] the task of finding, for a finite number of points whose pairwise distances are known, the shortest path connecting the points. The problem is naturally solvable by making a finite number of trials. No rules are known that would reduce the number of trials below the number of permutations of the given point.

The name "traveling salesman problem" makes one of its first appearances nearly two decades later, in a 1949 report prepared by Julia Robinson (1949) for the RAND Corporation. Then follows Gödel's much-quoted letter to von Neumann in March 1956 (Velupillai 2009) where the problem is again formulated, now in the context of a Boolean satisfiability problem.

Cook and Karp (see Machtey and Young 1979 for references and details) characterized *NP*-complete problems in the early 1970s; they are seen to pop up everywhere in both concrete and abstract situations. And Menge's question is one of them.

One may now state the $P = NP$ question:

> Is there a polynomial algorithm that settles all instances of some *NP*-complete problem?

NP-complete and *NP*-hard problems have the following features:

- The obvious search for a solution involves a search over some combination or permutation of the elements involved.
- For the *NP*-hard case, we look for a combination of permutation of elements that maximize or minimize some condition K, which is easily calculated.
- For the *NP*-complete case, there is also a condition K' which must be satisfied by the solution; again it is easy to check it.

This is the general picture. Let's now consider the situation in economics where several theoretical situations may also depend on the solution of a *NP*-complete problem. Consider the following case of a game with coalitions:

- We have a n-person game.
- We have some criterion K that has to be satisfied by coalitions in the game (say, returns for the coalitions).

[1] We quote a translation of Menger's text.

In the general case the obvious solution procedure is to examine all possible coalitions in order to check for K, and that demands an exponential effort.

Yet if we test for K we see that such a test is usually a polynomial task. More specific examples are now given. They show the widespread presence of *NP*-completeness in economics:

- *Computation of Nash equilibria.* Nash games are undecidable in the general case, a fact which is known since Lewis' results obtained in the late 1980s (see the references in Bartholo et al. 2009; see also da Costa and Doria 2005; Tsuji et al. 1998). However even for the simple case of explicitly given outcomes of games we get *NP*-completeness (Baron et al. 2004).
  We can offer an intuition about it: in order to compute Nash equilibria one must consider all possible combinations of players and strategies, so that out of the combinatorial explosion of alternatives we have the exponential growth typical of *NP* hard and *NP*-complete problems. If we ask for a minimum return in the game, then to check whether some choice of strategies satisfies it is in general a polynomial task. Therefore the actual computation of Nash equilibria (whenever possible) may then turn out to be a hard problem.
- *Risk management.* The choice of a portfolio given simultaneous risk and return constraints is also *NP*-complete, for we must consider all possible securities combinations and see whether they satisfy the desired constraints or not.
  This fact has led Huberman and coworkers to the suggestion that we should look at *NP*-completeness from the viewpoint of economics (Huberman et al. 1997). The idea is a very interesting one, for in most tasks that lead to *NP*-complete problems we can evaluate each alternative in terms of gains or in terms of risks incurred. For instance, given the current status of DNA research, and given the parents' genetic makeup, if we define a level of risk of defects in the possible offspring, we have a *NP*-complete problem which can be evaluated in economic terms, that is, in terms of gains (the children) and risks (namely inherited defects and diseases).
- *Shapley values. NP*-complete questions appear in the theory of cooperative games in general. A recent example is given in Conitzer and Sandholm (2004), which also lists related references. The chief result in the Conitzer and Sandholm paper is the proof that core membership in a cooperative game is *NP*-complete, a result that mirrors an earlier result by Deng and Papadimitriou (1994). However the calculation of Shapley values is easy, again as shown by Deng and Papadimitriou.

So, we believe that our case for the relevance of *NP*-hardness and *NP*-completeness for economists has been successfully pleaded. Not only the class of *NP* problems do matter for economics: we can even follow the converse path and look at some *NP* problems from the viewpoint of typical economic concepts such as gain, risk or value.

The O'Donnell Algorithm

Required concepts from mathematical logic and computer science can be found in Bartholo et al. (2009). We will essentially need a few intuitions about Turing machines, plus some results on logic which can be found in the reference.

Polynomial Turing Machines

Here Turing machines are supposed to input sequences of 0s and 1s such as, say, 001010. They are called *bit strings*. A *polynomial Turing machine* (or *poly machine*) is a Turing machine whose operation time is bounded by some polynomial on the length of the input string. (The length of a string is the number of 0s and 1s in it; given a bit string x, its length is noted $|x|$.)

The NP Class of Problems and the P = NP Conjecture

It is known that the so-called nondeterministic poly (that is, *NP*-) machines can settle any *NP* problem in polynomial time, whereas the use of the abbreviation *NP* to denote that class of problems. So, *NP* stands for "nondeterministic polynomial." The *NP* class can be described as:

> If you know the solution for a problem in *NP* then you can check it very fast. However if you don't know the solution then it usually takes a to get one solution in the general case.

More precisely: they are easy to check because they can be checked by a poly machine. They are hard to find because in the general case nobody knows an algorithm for it which is polynomial, i.e., can be implemented by a poly machine. Follows the $P = NP$ question (this equation means, all *NP* problems are solvable by poly machines).

The O'Donnell Algorithm

We consider the Boolean satisfiability problem, noted SAT (Machtey and Young 1979). The O'Donnell algorithm requires the set of all time-polynomial algorithms (actually the so-called BGS set of Turing machines (Baker et al. 1975)) and scans it in a prescribed way until it stops—and stop it will.

We frame our discussion in a theory S:

- S has classical first-order predicate logic as its underlying language.
- S has a recursively enumerable set of theorems.

- S includes Peano Arithmetic, and has a model with standard arithmetic (we then say that S is *arithmetically sound*).

Poly Machines

We deal here with Turing machines bounded by a polynomial clock on their binarily coded input, that is, if x is a binary input to machine $\{m\}$, the operation time of $\{m\}(x)$ is bounded by a polynomial function on $|x|$, say, $|x|^p + p$, $p \in \omega$. (We will indifferently write M_k, $\{k\}, e_k$, for Turing machines of Gödel number k.)

The (always satisfiable) instances of SAT are coded by all elements of ω in a monotonic way w.r.t. the length of the expressions in cnf.

The Algorithm

We will describe the O'Donnell quasi-polynomial algorithm for SAT. Here f_c is the (recursive) counterexample function to $[P = NP]$. The present section is based on the characterization of O'Donnell's algorithm in (da Costa and Doria 2016); it goes as follows. First the requirements:

- We use the enumeration of finite binary sequences

  $$0, 1, 00, 01, 10, 11, 000, 001, 010, 011, \ldots.$$

  If FB denotes the set of all such finite binary sequences, form the standard coding $FB \mapsto \omega$ which is monotonic on the length of the binary sequences.
- We use a binary coding for the Turing machines which is also monotonic on the length of their tables, linearly arranged, that is, a 3-line table s_1, s_2, s_3, becomes the line $s_1 - s_2 - s_3$.

  We call such monotonic codings *standard codings*.
- We consider the set of all Boolean expressions in cnf, including those that are unsatisfiable, or totally false. We give it the usual coding which is 1–1 and onto ω.
- Consider the poly (that is, time-polynomial on the binary length of the input) Turing machine $\mathsf{V}(x, s)$, where $\mathsf{V}(x, s) = 1$ if and only if the binary line of truth values s satisfies the Boolean cnf expression x, and $\mathsf{V}(x, s) = 0$ if and only if s doesn't satisfy x.

- Consider the enumeration of the BGS (Baker et al. 1975) machines, $\mathsf{P}_0, \mathsf{P}_1, \mathsf{P}_2, \ldots$.

We start from x, a Boolean expression in cnf binarily coded:

- Consider x, the binary code for a Boolean expression in cnf form.

- Input x to $\mathsf{P}_0, \mathsf{P}_1, \mathsf{P}_2, \ldots$ up to the first P_j so that $\mathsf{P}_j(x) = s_j$ and s_j satisfies x (that is, for the verifying machine $\mathsf{V}(x, s_j) = 1$).
- Notice that there is a bound $\leq j = \mathsf{f}_c^{-1}(x)$.

  Eventually a poly machine (in the BGS sequence) will produce a satisfying line for x as its output given x as input. The upper bound for the machine with that ability is given by the first BGS index so that the code for x is smaller than the value at that index of the counterexample function.

  That means: we arrive at a machine M_m which outputs a correct satisfying line up to x as an input, and then begins to output wrong solutions.
- Alternatively check for $\mathsf{V}(x, 0)$, $\mathsf{V}(x, 1)$, $\ldots$ up to—if it ever happens—some s so that $\mathsf{V}(x, s) = 1$; or,
- Now, if f_c is fast-growing, then as the operation time of P_j is bounded by $|x|^k + k$, we have that $k \leq j$, and therefore it grows as $O(\mathsf{f}_c^{-1}(x))$. This will turn out to be a very slowly growing function.

  The BGS machines are coded by a pair $\langle m, k \rangle$, where m is a Turing machine Gödel index, and k is as above. So we will have that the index j by which we code the BGS machine among all Turing machines is greater than k, provided we use a monotonic coding.

  More precisely, it will have to be tested up to j, that is the operation time will be bounded by $\mathsf{f}_c^{-1}(x)(|x|^{\mathsf{f}_c^{-1}(x)} + \mathsf{f}_c^{-1}(x))$.

  Again notice that the BGS index $j \geq k$, where k is the degree of the polynomial clock that bounds the poly machine.

(We give elsewhere a more precise depiction of f_c^{-1}.) Notice that the alternate testing procedure $\mathsf{V}(x, 0), \mathsf{V}(x, 1), \ldots$ amounts to the use of a poly machine that will eventually be found in the sequence of BGS machines. Also, a specific V ties the procedure to a specific problem, so this construction is quite general.

Then either x is unsatisfiable—and therefore one will have to test all possible s up to the envelope function of the counter example function (da Costa and Doria 2016)—or, if satisfiable, operation time has the nearly polynomial time given above. This means that if independence holds, then we will have something that might be informally written as $P \approx NP$—there are nearly polynomial algorithms, as independence means that f is total in the standard model for arithmetic for our theory S.

As an example: suppose that Peano Arithmetic proves $P < NP$, while Primitive Recursive Arithmetic doesn't prove it. Then O'Donnell's algorithm performs as an exponential that grows at most as $O(2^{\mathsf{F}_\omega^{-1}})$, where F_ω is Ackermann's Function, already a fast-growing function. (We may measure provability strength of a theory S by the set of provably total recursive functions in the theory.)

The COPPE–COSENZA Procedure

The COPPE–COSENZA technique for approximate solutions of allocation problems is now briefly described Cosenza (1981). It expands an earlier Italian model, the

MASTERLI model (MODELO DI ASSENTO TERRITORIALE E LOCALIZZAZIONE INDUSTRIALE) which was conceived and applied in the early 1970s. The COPPE–COSENZA procedure is partly heuristic and uses a fuzzy-logic approach in its handling of data. It is widely used in locational studies in Brazil (Bartholo et al. 2010) and in dealing with allocation problems from economics and engineering to medicine (Cosenza et al. 2006, 2007); empirical data support the contention that it in general gives better, more efficient, solutions to the issues considered.

The main idea is disarmingly simple: we define two matrices, A and B. The first one, matrix A, tells us the required factors. The second matrix, B, exhibits the possible alternatives we have in the real world in order to implement our wishes.

Then there is a third matrix, C, which is a function of A and B, which allows us to compute optimal allocations out of our desiderata and out of the real-world alternatives we have at our disposal. That computation is both simple and fast, and may have heuristic components. The main ideas can be formulated for crisp sets, but a more sensitive algorithm may be obtained with the help of fuzzy objects.

Construction of Matrix A

Suppose that we have several industries to distribute over a given geographic space, and suppose that we have different potential placements for those industries. The first matrix describes the industries we are interested in, and relates them to requirements for these industries (say, a shoe factory requires a continuous leather supply, water, energy, some chemical inputs and pollution control).

The first matrix, A, with k lines and m columns, has the following structure:

- *Lines* list the industries, $p_1, p_2, \ldots, p_k$.
- *Columns* list the requirements (factors) for these industries, $f_1, f_2, \ldots, f_m$.
- Given matrix A, its entries are linguistic variables A_{ij}, say:
  - Critical factor.
  - Decisive factor.
  - Indecisive factor.
  - Irrelevant factor.

Construction of Matrix B

Matrix B has n lines and k columns and tells us what we have to offer to the demands in matrix A.

Matrix B has the same structure as A but in a transposed way:

- *Lines* list the same requirements $f_1, f_2, \ldots, f_n$ that appear in A. Matrix B tells us what is available in our prospective placements.

- *Columns* list possible placements for our industries, $z_1, z_2, \dots, z_m$, where in general $n \neq m$.
- Again the entries B_{jk} of B are linguistic variables:
  - Optimal availability.
  - Good availability.
  - Regular availability.
  - Poor availability.

Construction of Matrix C

Matrix C will be the tool we require to do actual computations. Its entries are given by a heuristic procedure:

1. Suppose that there is demand for factor f_i (1 value of demand), and that region z_j doesn't have that factor (0 value of offer). We put $C_{ik} = 0$.
2. Suppose that there is no demand for f_i (0 value of demand), and yet that region z_j has that factor (value equal to 1). We put $C_{ik} = 1/n$.
3. Suppose that there is no demand for f_i (0 value of demand), and that region z_j doesn't have that factor (value also equal to 0). We put $C_{ik} = 1/n!$.
4. Finally suppose that there is a demand for f_i (1 value of demand), and that region z_j has that factor (value equal to 1). We put $C_{ik} = 1$.

These are simply "marks" or "grades" we give for the possible alternatives. Cases 1 and 4 are obvious: they correspond to 0% and 100%, respectively. Case 2 gives an intermediate nonzero value because the fact that (momentarily) one doesn't require a factor that is available and which may be required in the future must be taken into account. Finally Case 3—no demand and absence of a prescribed factor in the region—is given a nonzero value not to penalize the possibility, as 0 should only be given to a factor that is required and isn't available.

Ranking Techniques

There are several alternative, empirically tests, ranking techniques, which depend on the optimal goals to be attained. We describe here a simple ranking scheme that has given very reliable results in actual optimizing situations:

- Consider the demand for industry j. Form the demand value $D_j = \sum_k D_{kj}$. That is to say:
  - Fix industry j. For site k, sum over all "grades" of the factors required for j. This gives the demand D_j of industry j.

- Examine the offer for site m: $O_m = \sum_i O_{ik}$. The offer is calculated as we do for the demand.
- The rank grade r_{jm} of site m with respect to industry j is given by:

$$r_{jm} = \frac{O_m}{D_j}.$$

Given matrices A, B, C, there are of course many other possibilities to rank locations for industries with respect to the required factors. This is just the first possibility.

We can also have global rankings:

- Compute the global demand, for all industries, $D = \sum_j D_j$.
- Compute the global offer, $O = \sum_j O_j$.
- The rank grade r_m of site m with respect to all industries is: $r_m = O_m/D$.

Of course $r_m < 1$ means that not all requirements are fulfilled, while $r_m \leq 1$ means the possible fulfilment of the requirements. One usually ranks the solutions with respect to the global demand D (and not the D_j), which allows for a fast algorithmic treatment of the procedure (or it would be exponention, if we were to consider all D_j.

Comparison with An Analytic Solution

Given a fixed budget X, the distribution of F activities among Y locations, and the calculation of an arrangement, if any, that satisfies the prescribed budget is clearly *NP*-complete. For the obvious algorithm for the computation of an adequate arrangement of activities and locations is exponential, while testing whether some particular arrangement fits the budget can be done in polynomial time on the length of the input data.

Let's elaborate on that. In order to obtain one exact solution (if it exists) for one such *NP*-complete allocation problem as described, one would need:

- Crisp values for resources and demand at each location.
- We would have to consider all locations and the corresponding data; a solution for the problem might involve e.g. raw material obtained at places A_1 and A_2, plant location at A_4, administrative tasks at A_6, and so on. So, we would have to consider all possibilities at all places, which together with budget constraints makes the problem thus formulated into a *NP*-complete one.

The COPPE–COSENZA procedure is a kind of cutting the Gordian knot solution. It abandons crisp values for fuzzy data such as "critical" or "good." The ranking matrix sort of aggregates all data instead of considering each and one individual possibility, and the final ranking procedure derives from that aggregate consideration of all possible locations in a single demand index.

This semi-heuristical procedure however is known to provide efficient responses to most actual situations where it has been applied specifically because it improves over already existing unplanned situations. An example is its application to the Brazilian Biodiesel Program (Cosenza 2005; Cosenza et al. 2005). To apply it to another class of *NP* problems one must at first obtain a poly map between our allocation problem and the new kind of problems.

Discussion

A brief summary of what we have out of these situations is:

- If we deal with mathematical models for economic situations, we may have to cope with (among other things) *NP*-completeness.
- If we confront a practical, everyday, particular situation which can be usefully treated with a semi-heuristic procedure, then we will probably be able to forget about undecidability, incompleteness and *NP*-completeness.

Of course there are crisp problems, say, cryptographic decoding procedures, that will always require exact solution for a *NP*-complete problem. But if we only need approximate solutions then an algorithm like the COPPE–COSENZA procedure is enough.

Yet—think about the situation in physics. When Einstein first published his general theory of relativity in 1916, nobody would even consider the possibility that it would affect us in ordinary, everyday situations. Well, today our homely GPS guiding devices use general relativity corrections to help is in establishing our location.

This is a sobering remark. Follows that undecidability, incompleteness and again *NP*-completeness may eventually matter, after all.

The present section and the previous one were based on Bartholo et al. (2011).

Acknowledgements The ongoing research program that led to this text has been sponsored by the Advanced Studies Group, Production Engineering Program, COPPE-UFRJ, Rio, Brazil. FAD acknowledges support from the Brazilian Academy of Philosophy and its chairman Professor J. R. Moderno. The same author acknowledges partial support from CNPq, Philosophy Section.

References

P. Arestis, C.A. Cosenza, Hierarchy models for the organization of economic spaces. Preprint Land Department Economy, Cambridge University, UK (2009)

T. Baker, J. Gill, R. Solovay, Relativizations of the $P = ?NP$ question. SIAM J. Comp. **4**, 431–442 (1975)

R. Baron, J. Durieu, H. Haller, P. Solal, Finding a Nash equilibrium in spatial games is an *NP*-complete problem. Econ. Theory **23**, 445–454 (2004)

R. Bartholo, C.A. Cosenza, F.A. Doria, C. Lessa, Can economic systems be seen as computing devices? J. Econ. Behav. Organ. **70**, 72–80 (2009)

R. Bartholo, C.A. Cosenza, F.A. Doria, M. Doria, A heuristic algorithmic procedure to solve allocation problems with fuzzy evaluations. Preprint (2010)

R. Bartholo, C.A. Cosenza, F.A. Doria, M. Doria, A. Teixeira, On exact and approximate solutions for hard problems in economics: an alternative look. Preprint Fuzzylab–UFRJ (2011)

S. Ben–David, S. Halevi, On the independence of *P vs. NP*. Technical Report no. 699, Technion (1991)

V. Conitzer, T. Sandholm, Computing Shapley values, manipulating value division schemes, and checking core membership in multi–issue domains, in *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04).* (San Jose, California, USA, 2004), pp. 219–225

C.A.N. Cosenza, *Industrial Location Model: A Proposal.* preprint (Martin Centre, Cambridge U., Cambridge, 1981)

C.A.N. Cosenza, *Brazil's Biodiesel Project* (Oxford University, Centre for Brazilian Studies, 2005)

C.A.N. Cosenza, C. Neves, F. Rodrigues Lima, A decision–making method for the selection of biodiesel sites in Brazilian regional planning. Internal report, PIT–Coppe (2005)

C.A.N. Cosenza, M.E. Cosenza-Andraus, C.F. Andraus, S. Leon, R.G. Nunes, Monitoración prolongada por vídeo eeg de pacientes con diagnostico ambulatorial de epilepsía del lobo frontal de difícil control. Rev. Neurología de Barcelona **43**, 1–20 (2006)

C.A.N. Cosenza, R.C.M. Nobre, O.C. Rotunno, W.J. Mansur, M.M.M. Nobre, Ground water vulnerability and risk mapping using gis modeling and a fuzzy logic tool. J. Contam. Hydrol. **1**, 1–36 (2007)

N.C.A. da Costa, F.A. Doria, Computing the future, in *Computability, Complexity and Constructivity in Economic Analysis*, ed. by K.V. Velupillai. (Blackwell, 2005), pp. 15–50

N.C.A. da Costa, F.A. Doria, On the O'Donnell algorithm for *NP*-complete problems. Rev. Behav. Econ. **3**, 221–242 (2016)

X. Deng, Ch. Papadimitriou, On the complexity of cooperative solution concepts. Math. Oper. Res. **19**, 257–266 (1994)

B. Huberman, R.M. Lukose, T. Hogg, An economics approach to hard computational problems. Science **275**, 51–54 (1997)

M. Machtey, P. Young, *An Introduction to the General Theory of Algorithms* (North–Holland, 1979)

K. Menger, Das Botenproblem, in *Ergebnisse eines Mathematischen Kolloquiums*, ed. by K. Menger, Vol. 2 (1932), pp. 11–12

M. O'Donnell, A programming language theorem which is independent of Peano Arithmetic, in *Proceedings of the 11th Annual ACM Symposium on the Theory of Computation* (1979), pp. 176–188

J. R. Robinson, "On the Hamiltonian game (a traveling salesman problem," RAND Research Memorandum RM–303 (1949)

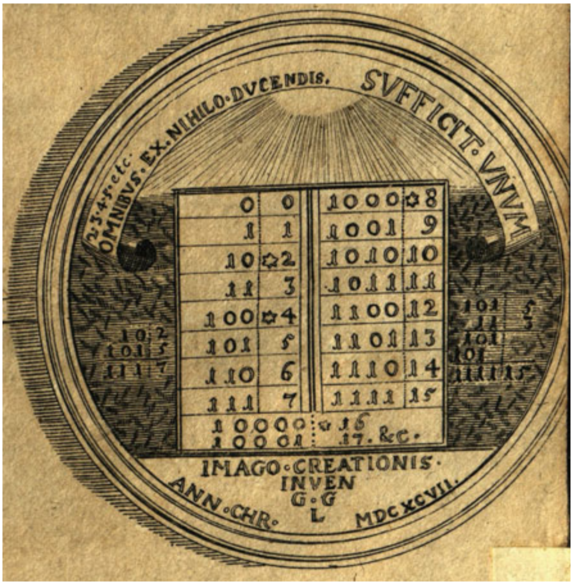
M. Tsuji, N.C.A. da Costa, F.A. Doria, The incompleteness of theories of games. J. Phil. Logic **27**, 553–563 (1998)

K.V. Velupillai, A computable economist's perspective on computational complexity, in *The Handbook of Complexity Research*, ed. by J. Barkley Rosser Jr., Ch. 4, 36–83. (Edward Elgar Publishing Ltd, 2009)

Chapter 24
Building the World Out of Information and Computation: Is God a Programmer, Not a Mathematician?

Gregory J. Chaitin

Leibniz Medallion



- Medallion designed by Leibniz in 1697 to celebrate his discovery of **base-two arithmetic** and of the combinatorial potential of 0 and 1 to generate the entire universe (**digital philosophy**).
- Reproduced from Rudolf August Nolte, *Mathematischer Beweis der Erschaffung und Ordnung der Welt* (Leipzig: Langenheim, 1734).

Talk presented at the Leibnizfest at the University of Turin in Italy on November 14, 2016, the three-hundredth anniversary of the death of the polymath.

G. J. Chaitin (✉)
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: gjchaitin@gmail.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_24

- *On the left,* an example of binary arithmetic: 2 plus 5 = 7.
  On the right, another example: 5 times 3 = 15.
- *In the middle,* the integers from 0 to 17 in binary and decimal.
- *Latin inscriptions:*
  Above: OMNIBUS EX NIHILO DUCENDIS, SUFFICIT UNUM / God has created everything from nothing / The 1 has created everything from 0.
  Below: IMAGO CREATIONIS / Image of the creation. Invented by Godefroi Guillaume Leibnitz. Christian year 1697.

We are gathered here today at the University of Turin to celebrate Leibniz, a truly remarkable thinker.

Most of his best work was described in letters to other European intellectuals or was written for his drawer, but due to the fact that he was also a diplomat for successive Dukes of Hanover, his papers—all his papers, not just those that dealt with affairs of state—were sealed and preserved at the Royal Archives in Hanover.

Among his many achievements, is the fact that he anticipated our digital age. He invented base-two notation and realized that it was particularly well-suited for hardware implementations. He also invented a calculating machine that could multiply. Pascal's original calculating machine could only add. And he realized how useful it would be to be able to perform calculations—arithmetical calculations or logical chains of reasoning—completely mechanically.

All of this more than three centuries before we routinely do this and this revolutionary technology has completely transformed our society and our lives!

He also had a vision, celebrated in the famous **Leibniz Medallion**, described above, that the entire world could be built out of information and computation, out of 0's and 1's and mechanical algorithms. So he is the patron saint of a new, contemporary school of philosophy, *digital philosophy*, which I shall be describing today, and he inspires people like Edward Fredkin and Stephen Wolfram to conceive of completely discrete, cellular automata worlds, which is called *digital physics.*

In particular, I want to tell you about Leibniz's ideas on *complexity*, very deep ideas, and how they are being developed mathematically three centuries later.

Leibniz on Complexity

- Leibniz, *Discours de métaphysique, VI* (1686):
  Dieu a choisi celui qui est le plus parfait, c'est à dire celui qui est en même temps le plus simple en hypotheses, et le plus riche en phenomenes [*God has chosen that which is most perfect, in other words, that which is simultaneously simplest in hypotheses and richest in phenomena*].
  Mais quand une regle est fort composée, ce qui lui est conforme passe pour irregulier [*But when a rule is very complex, that which obeys it passes for random*].
- *If arbitrarily complex theories are permitted then the concept of theory becomes vacuous because there is always a theory!*—Hermann Weyl (1932)

Leibniz's ideas on complexity are contained in a document, the *Discours de métaphysique* [Discourse on Metaphysics], a small document in French with an interesting history.

A copy of the *Discours* was found in the Leibniz *Nachlass* more than a century after Leibniz's death by a scholar, Georg Heinrich Pertz, who gave the document that name; the original had no title. Among other things, the document contains an analysis of elastic collisions and introduces the idea of *vis viva* which would eventually became kinetic energy.

The importance of Leibniz's remarks on complexity in sections V and VI of the *Discours* was only realized a century later, by the mathematician and mathematical physicist Hermann Weyl, David Hilbert's most distinguished student.

Leibniz jotted down these notes during a visit to the silver mines in the Harz mountains, where he was trying to improve the technology for pumping water out of the mines, when his engineering work was temporarily suspended by a snow storm.

He later sent a summary to the theologian Antoine Arnauld, who was horrified by the potential heretical implications. So Leibniz never sent Arnauld nor anyone else the *Discours.* Arnauld himself was, like Pascal, a Jansenist, an early form of Protestantism, and was in hiding at the time because the king of France, Louis XIV, had decided to suppress the Jansenists.

I should add that the *Discours* was written the year before the publication of Newton's *Principia,* a very different world from today, in which the mechanical philosophy—which was to became modern science—still coexisted with medieval theology.

That's the story of the *Discours.* But what does the *Discours* have to say about complexity?

There are two very fundamental ideas: Firstly, that science is possible because all the richness and diversity that we see in Nature is actually the product of a small number of laws of physics. God is parsimonious, and the apparent complexity we see everywhere hides an inner simplicity. This is what it **means** for science to be possible, for the world to be comprehensible. This is in section V.

Secondly, how can we distinguish a world governed by law from one which is lawless or governed by capricious Gods? You might think that if empirical, experimental data about a physical system is graphed as a finite set of data points on graph paper, for example, temperature as a function of time, and there is a mathematical equation passing through those points, then the physical system is lawful, not random.

But, as Leibniz observes in section VI there is always a mathematical equation passing through **any** finite set of points, so this cannot enable us to distinguish between a lawful and a lawless system. What then can?

Well, says Leibniz, if the equation going through those points is very complicated, the system is lawless, but if the equation is simple, then the system actually follows a law.

Of course, how true, why didn't I think of that myself!? Well I actually did, in a somewhat different context, as I will soon explain.

Hermann Weyl not only identified the treasure in sections V and VI of the *Discours,* he made two dramatic observations. First, Weyl observes that if arbitrarily complex laws are permitted, then the concept of law becomes **vacuous**, because there is **always** a law. So the concept of law requires an accompanying concept of complexity in order to be meaningful.

Furthermore, Weyl observes that identifying the complexity of a law-equation with its size is a reasonable first step, but has the somewhat unsatisfying feature that it makes this important concept dependent on mathematical notation, which is somewhat arbitrary and varies as a function of time.

What to do? Algorithmic information theory to the rescue!

We must change the context, from continuous mathematics to discrete mathematics, and from laws as equations to laws as computer programs, whose complexity will naturally be measured in bits of software.

Metaphysics

- **Leibniz 1686**:
  Empirical data: points in a plane (pairs of real numbers)
  Theory: a mathematical equation for a curve passing through those points
  The equation must be simple!
- **Algorithmic Information Theory 1960s** (Solomonoff, Chaitin, [Kolmogorov]):
  Empirical data: a finite string of bits
  Theory: software, a program that calculates exactly the empirical data
  The number of bits of theory must be much smaller than the number of bits of data.
- **Analogy**:
  program → **COMPUTER** → *output*
  physical theory → **COMPUTER** → *experimental data*

Algorithmic information theory (AIT) was independently proposed by three people in the 1960s: Ray Solomonoff, Andrey Kolmogorov, and myself. Solomonoff and I were interested in epistemological toy models, in analyzing mathematically how empirical science works. We were interested in being able to evaluate how good a theory is. And Kolmogorov and I were interested in providing a mathematical definition of randomness as lack of pattern or structure, which might be referred to as *logical randomness* as opposed to *physical randomness*.

The fundamental idea of AIT, inspired by considering the above toy model of how science works, is to define the complexity or *algorithmic information content* of something to be the size in bits of the smallest program for calculating it. And AIT solves the problem of the dependence of this concept on the programming language being employed, by identifying those languages that have the most concise programs.

In a nutshell, we consider only binary programs, bit strings p, and use universal Turing machines $U(p)$ with the property that for any other computer C there is a bit string π_C such that $U(\pi_C p) = C(p)$, i.e., the machine U can simulate the machine

C by adding a fixed prefix π_C to each binary program p for C that indicates which computer to simulate.

And a decade later, in the 1970s, I and, independently, Leonid Levin, realized that U should also be *self-delimiting,* which means that as U reads the program p it realizes by itself where p ends without there having to be a blank or other end-marker symbol, i.e., without having to read beyond the end of p. This has the important consequence that the set of all possible computer programs acquires a very natural probability measure: the probability of a program p is merely $2^{-|p|}$, one over two raised to the size in bits of p.

So what can we do with all this machinery which was inspired by considering epistemological models of how empirical science works? Well, in my opinion, the most interesting application is the new light that AIT sheds on metamathematics and, in particular, on the celebrated—or infamous—incompleteness phenomenon.

And AIT can also help us to analyze biological evolution by natural selection, at least at a meta level, i.e., to extract some fundamental mathematical concepts from the seemingly impenetrable jungle of contemporary biological thought. That's the way biology initially appears to a bewildered pure mathematician: quite ill-adapted to mathematization. Not a context where you can prove anything. We will see!

Metamathematics

- A *formal axiomatic theory* is a computer program for systematically generating all the theorems from the axioms. Software!
- Looking at the number of bits of software in a mathematical theory (at its *algorithmic information content*) gives us a new path to Gödel incompleteness.
- The platonic world of math has **infinite** complexity, but our theories can only have **finite** complexity, and thus capture only an infinitesimal part of the truth.
- The halting probability Ω has infinite complexity: each bit in its binary representation is an independent, **irreducible** mathematical truth!
- Mathematics is *quasi-empirical,* i.e., different from physics, but not that different. New principles can be justified **pragmatically**, because of their usefulness, without proof. Experimental mathematics!
- **Analogy**:
  program → **COMPUTER** → *output*
  mathematical theory → **COMPUTER** → *theorems*

The conventional view of mathematics is that, at least in principle, it is the only field of human knowledge capable of providing *absolute certainty*. And mathematical truth is, it is thought, totally objective, not subjective, **black** or **white**, not at all gray.

This was Leibniz's view, when he anticipates the central idea of modern logical and formal axiomatic theories, that deductions can be performed **mechanically**, and this was also the basis for David Hilbert's metamathematical program, the so-called

Hilbert program, whose goal was to identify a single formal axiomatic theory that, at least in principle, would enable one to mechanically generate all possible theorems, all mathematical truths, an endless computation of course.

And now a contrarian view: According to AIT, mathematics may be different from physics—as Vladimir Arnold joked, the experiments are much cheaper—but it is not that different. What counts is unification, compression, identifying unifying principles, discovering that seemingly unrelated questions have a common element. This is how mathematical or physical theories organize and compress our mathematical and physical experience—in one case experiments performed in physics laboratories, in the other case, experiments performed *in silico*.

The loss of certainty is actually a boon not a curse. If Hilbert had been right, mathematics would be a dead subject. Students would only have to study the great works of the past. Instead, as Emil Post emphasized to an uncomprehending world in the 1940s, AIT shows that mathematics is an open system, one in which creativity will always be required, one embracing unifying principles whose justification is pragmatic, not deductive. In other words, mathematics should also allow inductive or experimental proofs, not just deductive proofs.

What really counts in mathematics as well as in physics, is to identify fertile principles that help us to comprehend and to organize new fields of mathematical or physical experience.

So AIT teaches us that mathematics is endlessly creative, much like biology. A wild thought: Can we apply some of these ideas to theoretical biology? Does Gödel's incompleteness theorem have anything to do with biological creativity? Can we do fertile concept migration from metamathematics to theoretical biology?

Yes, I think so! So let's now try to apply this software view of science in one more area, in biology.

Metabiology

- It is commonly said that DNA is software.
- **Proposal**: Study the effect of natural selection on randomly mutating computer programs rather than randomly mutating DNA.
- Organisms are pure DNA, i.e., computer programs. Software organisms calculate an integer, which determines their fitness, i.e., the bigger the better. Mutations are arbitrary algorithms for transforming organisms. The population consists of a single organism at a time.
- The probability of an $|M|$-bit mutation M is $2^{-|M|}$. (The program M must be self-delimiting so that the total probability is ≤ 1.)
- Evolution is a hill-climbing random walk in software space; a random mutation M is accepted only if it increases the fittness of our one and only organism.
- **Goal**: A toy model that mathematically expresses fundamental biological concepts and for which it is possible to prove that Darwinian evolution works, that is, is open-ended.

- **Analogy**:
  program → **COMPUTER** → *output*
  organism → **COMPUTER** → *fittness of organism*
  organism → **MUTATION** → *mutated organism*

The basic idea of this metabiological toy model of evolution is that organisms are mathematicians working on the Busy Beaver Problem, the problem of trying to concisely name extremely large integers, an open-ended problem which can absorb—which can take advantage of—an unlimited amount of mathematical creativity. Not surprisingly, because the BB problem is equivalent to Turing's famous halting problem, which has no algorithmic solution, and which no one formal axiomatic theory can enable us to settle in all cases.

That is how we transform a meta-theorem about open-ended mathematical creativity into a meta-biological theorem about open-ended biological creativity.

Does this have any relevance to real biology?! Maybe not, or maybe algorithmic mutations suggest trying to explain discontinuities in evolution as the products of more powerful mutational mechanisms than are currently envisioned. At any rate, metabiology is an attempt to take the idea of DNA as software seriously and run with it mathematically. This was already done by John von Neumann in a visionary lecture in 1948 that he published in 1951 identifying DNA as digital software—visionary because this was well before Watson and Crick and the rest of molecular biology.

But what metabiology does to enrich von Neumann's vision, is to provide a mathematically compelling setting for organisms and for mutations. It proposes treating the space of all possible organisms and the space of all possible mutations as software spaces. I believe that here, finally, are mathematical spaces that are rich enough to take as the basis for a theoretical biology, for a biology seen through mathematical spectacles, through a mathematical magnifying glass.

Crazy, you may say! Or perhaps not crazy enough? But now for our final topic, an even wilder vision.

The Perfect Language

What is the perfect language?

It is the adamic language of creation, the language used by God to create the universe, the language that directly expresses the innermost structure of the physical world.

The medieval search for this language is described in Umberto Eco's delightful book, which even has a chapter on Leibniz.

Contemporary physicists suppose that this language of creation is pure mathematics, in other words, that God is a mathematician, in fact, a mathematician who uses continuous mathematics.

But what if, following the message in the Leibniz Medallion, we assume that the world is discrete, that it is built out of information, out of 0's and 1's? Then God would have to be a programmer, not a mathematician!

And in this new neo-pythagorean ontology, **matter and energy** are respectively replaced by **information and computation**. In other words, instead of the Pythagorean **All is Number!** now **All is Algorithm!**

Yes of course, I know this is a bit far-fetched. You don't have to rub it in! But three centuries from now, who knows?!

After all, some of Leibniz's ideas that we most admire also seemed far-fetched three centuries ago.

Further Reading

1. Gottfried Leibniz, *Discours de métaphysique (Nouvelle édition, collationnée pour la première fois avec le texte autographe de l'auteur),* (Paris: Félix Alcan, 1907), page 33, available from Gallica (Bibliothèque nationale de France) at http://gallica.bnf.fr. Published in Italian as *Discorso di metafisica.*
2. Hermann Weyl, *The Open World: Three Lectures on the Metaphysical Implications of Science* (New Haven, CT: Yale University Press, 1932). Published in Italian as *Il mondo aperto.*
3. Ugo Pagallo, *Introduzione alla filosofia digitale. Da Leibniz a Chaitin* [Introduction to Digital Philosophy: From Leibniz to Chaitin] (Torino: Giappichelli, 2005).
4. Ugo Pagallo, *Leibniz. Una breve biografia intellettuale* [Leibniz: A Brief Intellectual Biography] (Padova: CEDAM, 2016).
5. Giuseppe O. Longo and Andrea Vaccaro, *Bit Bang. La nascita della filosofia digitale* [Bit Bang: The Birth of Digital Philosophy] (Milano: Apogeo Education, 2013).
6. Umberto Eco, *The Search for the Perfect Language* (Oxford: Blackwell, 1995), published in Italian as *La ricerca della lingua perfetta.*
7. Gregory Chaitin, *Meta Math! The Quest for Omega* (New York: Pantheon, 2005), published in Italian as *Alla ricerca di Omega.*
8. Gregory Chaitin, *Proving Darwin: Making Biology Mathematical* (New York: Pantheon, 2012), published in Italian as *Darwin alla prova. L'evoluzione vista da un matematico*.
9. Robert J. Marks II, William A. Dembski and Winston Ewert, *Introduction to Evolutionary Informatics* (Singapore: World Scientific, 2017).
10. Gregory Chaitin, "Consciousness and information," draft of a reaction to Chap. 8 of David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, 1996.
11. Many papers by Chaitin are available from his website at https://ufrj.academia.edu/GregoryChaitin.

Part IV
Biology/Cognitive Science

Chapter 25
The Invention of Consciousness

Nicholas Humphrey

The literary critic William Empson said of his own profession: “Critics are of two sorts: those who merely relieve themselves against the flower of beauty, and those, less continent, who afterwards scratch it up. I myself, I must confess, aspire to the second of these classes; unexplained beauty arouses an irritation in me” (Empson 1930). We could say that students of consciousness are of two sorts also. On the one hand, those who want to see the mystery left intact, well watered but otherwise untouched. On the other, those who see it as a scientific challenge, a natural phenomenon that we need to dig up and explain.

Yet we all start from the same place. We relish the heat and redness of a fire, the sour tang of a lemon, the caress of a lover’s hand. Mystic or sceptic, we all agree that consciousness is wonderful. Conscious sensations lie at the core of our being. Without them we’d be poorer creatures living in a duller world. What’s more we all agree that consciousness is inexplicable—or at any rate that it is at present *unexplained.* The problem is not that we do not understand consciousness at all. Some aspects of it are relatively easy to account for in scientific terms. The problem is that *one* aspect continues to baffle everyone, and that’s the “qualitative feel of consciousness”: the redness of red, the painfulness of pain. The *qualia*—or, as Tom Nagel has put it, simply “what it’s like.”

The biologist H. Allen Orr probably speaks for the majority of scientists when, in a review of Nagel’s book “Mind and Cosmos,” he writes: “I share Nagel’s sense of mystery here. Brains and neurons obviously have everything to do with consciousness but how such mere objects can give rise to the eerily different phenomenon of subjective experience seems utterly incomprehensible” (Orr 2013). Or, as Colin McGinn has colourfully put it: “The brain is just the wrong kind of thing to

Published in Topoi, https://doi.org/10.1007/s11245-017-9498-0. Based on the “Mind and Brain Prize” lecture, Turin, 2015.

N. Humphrey (✉)
Darwin College, Cambridge, UK
e-mail: humphrey@me.com

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_25

give birth to consciousness. You might as well assert that numbers emerge from biscuits or ethics from rhubarb" (McGinn 1993).

Well, let's see. I've called this paper "The Invention of Consciousness" because I want to play on two different meanings of the word "invention" in the English language.

An invention can be:

1. *A device or process, developed by experiment, designed to fulfill a practical goal.*

For example, a light-bulb or a telescope.

But alternatively, an invention can be:

2. *A mental fabrication, especially a falsehood, designed to please or persuade.*

For example, a fairy tale or a piano sonata.

I am going to argue that human consciousness is an "invention" in both these senses.

That's to say, consciousness is:

1. *A cognitive faculty, evolved by natural selection, designed to help us make sense of ourselves and our surroundings.*

But, on another level, consciousness is:

2. *A fantasy, conjured up by the brain, designed to change how we value our existence.*

I'll argue that qualia make little if any contribution to the cognitive faculty. However they lie at the very heart of the fantasy.

I must start, of course, by defining the scope of the term "consciousness". People sometimes make a big meal of this. But I don't think this first step need be controversial—at any rate, not if we can ground it in the case we each know best subjectively, our own. If I may speak objectively on your behalf, consciousness is surely just *what you are conscious of*: that's to say the various states of mind of which at any one time you are the subject, and which are accessible to you by introspection.

It's true that consciousness, defined this way, may be difficult to access in nonverbal animals. But fortunately grown-up human beings can indeed *tell* us about it (at least up to a point). And what all agree is that you can be conscious of a range of rather different kinds of mental state: perceptions, memories, wishes, thoughts, feelings, and so on. When you introspect, you observe these various states, as it were with an inner eye. So, it comes naturally to you—and people everywhere do this—to think of consciousness as some kind of window on the mind, a private view of the stage where your mental life is being played out.

A view from whose standpoint? Well, from the standpoint of whom else but “you”, your *self*. And this brings us immediately to one of the most striking features of consciousness: its *unity*. There’s only one “you” at the window. Only one self. When you find yourself feeling pain, or wanting breakfast, or remembering your mother’s face, it’s the same you in each case.

We might think it obvious that it has to be so. But actually this unity is by no means a logical necessity. I’d say it’s quite conceivable—and indeed psychologically plausible—that your brain could house several independent you’s, each representing a different segment of the mind. Indeed this fragmented state may have been the way you and every other human being started out at birth. Back then, and for the first few months of life, the different you’s might hardly have known each other. Thankfully, however, it was never going to stay that way. As your life got going and your body—your *one* body—began interacting with the outside world, these separate selves were destined to come into register—orchestrated, as it were, by the single line of music that, as it happened, made up your one life (Humphrey 2000).

Was this “binding of selves” genetically pre-programmed? Not necessarily. I think it could have been the automatic outcome of the dynamics of mind and body. In fact something like it can be seen occurring in quite simple physical systems. In the 17th century Christian Huygens, the inventor of the pendulum clock, made a surprising observation. When two or more of his clocks were hung from the same beam, he noticed that their pendulums spontaneously began to beat in synchrony, showing as he put it an “odd kind of sympathy”. In a more recent demonstration, a set of five metronomes are placed on a floating table, and they too soon begin beating as one (Harvard 2016). It happens because each individual metronome, interacting via the table, feels the pull of the others. In the case of consciousness, presumably the story must be more complex. Yet perhaps not very much more complex. Perhaps the separate parts of a newborn mind, interacting with a single body, also somehow feel the pull of the others.

Whatever the truth of this, let’s turn to the big question. Once your mental states all have the same subject, what does this unity achieve? The answer is a big one too. The unity of consciousness underwrites the most obvious cognitive function of consciousness, which is to create what Marvin Minsky has called the “society of mind” (Minsky 1986). Just as—in fact just because—there is only one “you” at the window, there comes to be only one mind on the other side. Information from different agencies is being brought to the same table, as it were, and it’s here that your sub-selves can meet up, shake hands and engage in fertile cross-talk. This means you now have a mind-wide forum for planning and decision making. And the way is then open for a *central processing unit* to take control: an intelligent agent that can recognise patterns, marry past and future, assign priorities and so on across the mind as a whole. A computer engineer might recognise this as an “expert system”. You of course recognise it as “I”.

But, alongside this, another opportunity emerges. Once you can observe the parts of the mind interacting on a single stage, you are in a position to *make sense* of the interaction. And this can support a second important function of consciousness: namely, to allow you to appreciate just *how* your mind works. Observing, for example, how "beliefs" and "desires" generate "wishes" that lead to "actions", you find your mind revealed as having a clear psychological structure. Thus you begin to gain insight into *why* you think and act the way you do. This means you can explain yourself to yourself, and explain yourself to other people too. But, equally important, it means you have a model for explaining other people to yourself. When you meet another person, you can assume his mind works much as yours does. So you can work out what he is likely to be thinking and how he will behave. Consciousness has laid the ground for what psychologists call "Theory of Mind."

So far, so good. We have a workable definition of consciousness in terms of introspection. And we've identified two ways in which introspection can be put to practical use. So that's two reasons why this kind of consciousness would have been likely to be selected in the course of evolution. What's more we have a plausible metaphor for how it works: consciousness provides a window onto—and at the same time creates—the society of mind.

Yet, what about the imagery I'm using here? Doesn't it smack of the "Cartesian theatre" on which Dennett (1991) has poured such scorn? No, I think that's a false worry. What Dennett has objected to is the idea that the brain contains a projection space where a replica of the outside world is on show to an inner observer. But I hope it's clear this is not what's being proposed. What the window of consciousness opens onto is a picture not of what's outside but of what's inside—the mental states whose turns and twists and conflicts underlie the way you think and act. If this is theatre, it is indeed more like a proper human theatre, where a *play* is running.

Imagine yourself at a performance of a Shakespeare play. Shakespeare was not concerned with copying reality. His plays are stories, dramatic mock-ups, designed to analyse, expose and explain. And indeed as he himself made plain, the stories rely on codes and shorthand. In a famous prologue to Henry V, the Chorus apologises on behalf of the actors—mere *ciphers or symbols*—for daring to recreate the pageant of history on stage. "Pardon," the Chorus says, "The flat unraised spirits that have dared on this unworthy scaffold to bring forth so great an object: Can this cockpit hold the vasty fields of France?" The secret, he continues, lies in the encryption. Just as a string of zeros can represent a huge number—"Since a crooked figure may attest in little place a million"—so the players and props on stage can a reality of quite a different order. "So let us, ciphers to this great account, on your imaginary forces work."

It's a startlingly prescient passage—almost as if Shakespeare has anticipated modern ideas about how mental states are represented in the brain. But, now the words are in front of us, I want to take up another remarkable allusion: "Can this *cockpit* hold the vasty fields of France?"

The term "cockpit" originated of course as the name of an arena for staging cock-fights. Already by Shakespeare's time it had morphed into the name for any confined space where important things get done. He could not have known that the word cockpit would later come to mean the wheel-room of a ship and later still the control room of an aeroplane. Yet, now, when we're discussing consciousness, I want to suggest the cockpit of a plane provides an even better analogy for consciousness than the theatrical stage does.

So picture, if you will, the cockpit of a plane. And place yourself where the pilot sits. You'll see before you an array of instrument panels, that display the output of a variety of modules that are monitoring the plane's external and internal states: speed, altitude, fuel reserves, global position, intended course, and so on. Let's say then that, from your privileged seat, you have a window on the plane's beliefs, desires, and intentions—presented in coded form, of course, as numbers, icons, graphs. Your job as pilot is to integrate all this information, so as to decide what to do to achieve certain goals. You must observe, then think, then act. You have a joystick with which you can control the plane's wing flaps and tail fin, so as to steer the plane in the intended direction. Oh, and by the way, you also have a cockpit radio, so you can report verbally to ground control. You have become in effect the *plane's self*,

You'll appreciate the analogy. And yet, you may be wondering what the point is. A *conscious human pilot* as an analogy for a *conscious agent* in the brain? If there's consciousness on both sides of the equation, where does that get us? But that's just it. It doesn't have to be on both sides. I want to use the analogy as a further way of demystifying consciousness.

We already know for a fact that there's no need to have a conscious agent in the pilot's seat. An electronic autopilot, made of nothing but circuit boards, can—and in many planes does—fulfill exactly the same function as the pilot, collating information, referencing a knowledge base, choosing the best path, and so on. The autopilot can even be designed to report on what it's doing and why, to a base on the ground, in simulated speech if required. And it can keep a historical record of its own activity (tucked away in a Black Box so that it can be accessed posthumously if necessary).

True, no one has yet engineered a plane's autopilot to be capable of reading the minds of other planes. But as it happens just such meta-cognitive abilities are already being incorporated into the computers of driverless cars. To navigate traffic safely, the computer must be able to anticipate how other cars are likely to behave. The computer has to have, in effect, a "Theory of Drivers". How does it learn this theory? I don't know the facts here, but I wouldn't be surprised if engineers are already working on having one computer learn how to model other computers by reflecting on its own example.

So, back to the problem of consciousness. My point of course is that if an electronic autopilot can be engineered to do all this, then it's not so surprising that a brain can. We're talking normal science and engineering, here. In fact the science is well under way. To mention a few areas of good progress: Stanislas Dehaene (2014) has been mapping what he calls "the global neuronal workspace"; Giulio Tononi (2012) has proposed a statistical model of "integrated information"; Crick and Koch (2005) have identified a brain structure, the claustrum, as a likely candidate for the master of ceremonies.

I suggested at the start that consciousness is an invention in the first sense of the term: "A cognitive faculty, evolved by natural selection, designed to help us make sense of ourselves and our surroundings". Exactly. So far it seems this is just what consciousness is. And, as I suggested would be the case, we haven't yet had to say anything about the mysterious feel. We get this cognitive faculty—the workspace, the integration, the theory of mind—without having so much as to mention the *eeriness* of consciousness.

This is good news, in its way. But bad news too. The good news is that we're getting an account of consciousness that looks like being scientifically respectable. The bad news is we're getting an account of consciousness that leaves out the very thing that many of us think of as its most baffling and intriguing feature. What about the eery phenomenal feel of consciousness? Where's the "what it's like" that everyone beefs about?

We defined consciousness at the outset as comprising all those mental states that are available to introspection. But now, if we want to make the eeriness of consciousness the issue, we'll have to focus in. Does the quality in question pervade all mental states? No, that's the thing: it does not seem to be a feature of higher-level cognitive states. At any rate it's not a necessary feature. There is no special *feel* associated with your having the thought, say, that today is Thursday. It's not *like anything* for you to believe it's going to rain, or to remember where you put your hat.

Rather, it seem the phenomenal quality kicks in only at a more animal level. It's there especially, perhaps exclusively, in the way you represent what's happening at your bodily sense organs—skin, eyes, nose, ears, tongue. It's there—and it's only there—with your experience of *sensations*: the pain of a bee sting, the salt taste of an anchovy, the blue look of the sky. Among conscious mental states, sensations have the very special property of being *intrinsically eery*, they simply couldn't be the states they are without having this mysterious dimension to them.

As I said at the opening, sensations lie at the heart of our being. No one would or could wish *qualia* out of existence. Indeed there will have been times for all of us when conscious experience is *about* little else. A science of consciousness that leaves qualia out is not just ignoring the elephant in the room, it is ignoring the elephant that *is* the room. Yet so far it seems that this is all the science we're getting. How can that be?

There may be several explanations for why qualia are not been given the priority we might expect. No doubt it's partly because, as we have just seen, cognitive science can indeed go a long way towards explaining consciousness without any reference to them. But it's also because of the fear, expressed by a good many scientists—and philosophers too—that it will never be possible to explain qualia in conventional scientific terms. H. Allen Orr, as we saw, said that qualia are "utterly incomprehensible". Christof Koch wrote to me not long ago: "it is bizarre that brain matter should exude these phenomenal feelings. Consciousness is so vivid, and its properties appear so otherworldly, that it seems to call for God.". Koch may have been half-joking. But who's laughing? Short of invoking some supernatural agency, where are we to go?

There are indeed a good many theorists who simply don't want to go anywhere with it. It's not so much a case of *qualia denial*—though that exists too—as *qualia avoidance*. Isaac Newton set the tone five hundred years ago: "But, to determine more absolutely, what Light is, after what manner refracted, and by what modes or actions it produceth in our minds the Phantasms of Colours, is not so easie. And I shall not mingle conjectures with certainties" (Newton 1671). Jerry Fodor has echoed Newton's pessimism: "We don't know, even to a first glimmer, how a brain (or anything else that is physical) could manage to be a locus of phenomenal experience. This is, surely, among the ultimate metaphysical mysteries; don't bet on anybody ever solving it." (Fodor 1998).

Of course not everyone has been so ready to surrender. In the coffee room, if not yet the lab, there has been ongoing debate about what just what kind of thing qualia are and what to do about them. The answers that have been proposed have not always been helpful. Yet it does seem a consensus is emerging, at least about the boundaries of the problem. Most theorist now accept that there are only two options that can be taken seriously. We can be *Realists* about qualia, or else we have to be *Illusionists* (Frankish 2016).

The names make the meaning of these alternatives clear. Realists take qualia at face value. In their view, if your sensations appear to have qualities that lie beyond the scope of physical explanation, then it must be they really do have such qualities. And this is possible because the brain activity that underlies sensations already has consciousness latent in it as an additional property of matter—a property as yet unrecognised by physics, but one that you the conscious subject are somehow able to tap into. Tom Nagel, for example, writes: "The existence of consciousness seems to imply that the physical description of the universe, in spite of its richness and explanatory power, is only part of the truth, and that the natural order is far less austere than it would be if physics and chemistry accounted for everything." (Nagel 2012). So, according to the Realists, when you experience pain, say, you are in effect breaking through the veil of mundane physics to access a higher-order realm.

Illusionists, by contrast, will have none of this. They argue that if your sensations appear to have these marvellous non-physical properties, then this can only be because your physical brain is playing tricks on you. And this is possible because the brain is a computational engine that deals in symbols, and physically based symbols can perfectly well represent states of affairs that do not and even could not exist (thank you, Shakespeare!). Dan Dennett, for example, has it that: "Consciousness is an illusion of the brain, for the brain, by the brain." Qualia are like "a beautiful discussion of purple, just about a colour, without itself being coloured" (Dennett 1991, p. 371). So, according to Illusionists, when you have a sensation—of purple, or sweetness, or pain—you are accessing your own brain's magic show and being *tricked into believing* you have reached through to another level of reality, when in fact it's all coming from your side.

Realism and Illusionism. The trouble is that both these theoretical positions come at a considerable price. On the one hand, the price of Realism is that it implies that the standard physical description of the world is radically incomplete. Some people actually welcome this. Nagel thinks it would make the natural order less austere! But others—including me—find it a lazy and inelegant solution. Do we really need to dream of there being unknown dimensions to the physical world? If we can send a probe on a journey lasting ten years, crossing 4 billion miles of empty space, so as to land it on a comet speeding at 34,000 mph, and actually get it there within 2 min of its planned arrival, doesn't that suggest that our existing physics is pretty much complete? If those other dimensions are really out there, I have to say they are exceedingly coy.

But then, on the other hand, there's a price to illusionism too. Illusionism undermines—and in many people's eyes devalues—the mystery of human experience. Some people welcome that too. Dennett clearly takes wicked delight, in discomforting what he calls the Mysterians. He's happy to be, as he puts it, "the cop at Woodstock" (the policeman at a pop festival). But many others find illusionism deeply depressing, complaining that it "unweaves the rainbow" and so on.

Still, which is right? No one yet knows sure. But I'm not hiding which I hope is right. Although I myself have recently questioned the language of illusionism (Humphrey 2016b), I hope to see a resolution of the "hard problem" within the bounds of our standard world model.

Here's an appealing analogy. I expect you are familiar with the "real impossible triangle", or "Gregundrum", a wooden object invented by Richard Gregory which, when looked at from one particular viewpoint, looks exactly like a solid Penrose triangle—a structure that simply couldn't exist in the physical world. My suggestion—my hope—is that the apparent "unreality" of consciousness comes down to a similar trick of perspective.

Can we do better than merely hope for this? Does anyone have any idea about what kind of physical processes in the brain might possibly underlie it? Actually yes, as I'll explain in a moment, I think—contrary to Fodor—we do have at least "a first glimmer". But before going there I want to consider a much simpler example. When sceptics are questioning whether any scientific theory can deliver the semi-magical effects, it will be good if we can point to a model mechanism that can emulate some of these effects. Then, at least we'll have a proof of principle.

So let's go back to my cockpit analogy. And let's suppose now that the plane you are flying has specialised sensors in its body, analogous to human sense organs, whose job is to represent what's happening at its body surface—heat, pressure, tissue damage and so on. Let's suppose, too, that there is a special set of "sensory instruments" in the cockpit, which display this information. But here's what's special: while all the other instruments on the panel use simple flat graphical or numerical displays, the sensory instruments—and only the sensory instruments—dress them up in a very special way… as *holograms.*

We've all seen holograms. The picture appears to rise above the flat surface. Of course we know it's not real. It only looks as if there's a third dimension. However, you, in the magical cockpit *don't know this*. To *you* it seems that the numbers really are jumping out of the screen. No wonder, then, that you find these sensory displays specially attention-grabbing and impressive. You do your best to explain to others, over the radio, just what it's like. But sadly, words often fail you. Still, it is your own first-person experience that matters to you above all. From now on you will go flying just to immerse yourself in these extraordinary displays. As Lord Byron said: The great object of life becomes sensation—"to feel that we exist, even though in pain" (Byron 1813).

But I must not get carried away, just because you the pilot have been. I'm running ahead of my own argument.

OK. An analogy is an analogy. A hologram is a hologram. What can this actually have to do with the brain and qualia? Well, dare I say it, maybe it's not just an analogy. I want to draw your attention to the so-called "holographic principle" which has come out of cosmology and the physics of black holes. The principle states that, not only can a 3-dimensional world always be represented without loss of information by a 2-dimensional surface (as in a conventional hologram), but *an n-dimensional world can always be represented by a (n-1) dimensional surface.*

Thus, to start with, when 3-dimensional objects disappear into black holes, the information they contain need not have been finally lost—which would be problematical for physics — but instead could be preserved on the hole's 2-dimensional surface*, from which an illusion of the original objects could be regenerated.* In fact, in light of this, cosmologists have suggested that the 3-dimensional world we ourselves believe we inhabit could actually be just such an illusion arising from a flat 2-dimensional surface. But more to the point, we can now suggest that the 4-dimensional world of conscious qualia could quite well be an illusion generated by a 3-dimensional brain. As someone said about the black hole case: "This idea is so odd, it's comparable to finding that the instruction manual for a dishwasher holds the recipe to making a good chocolate soufflé" (Maynard 2015). Ah ha! As someone else said about consciousness: "You might as well assert that numbers emerge from biscuits or ethics from rhubarb" (McGinn 1993). Looks as though we might be on to something!

Yes, but how precisely could it work? As it happens, Karl Pribram, back in the 1970s, did indeed raise the possibility that information in the brain is stored in holograms. But no one today takes Pribram's detailed model seriously. So how else might the brain be generating a higher-dimensional sensory display? I've been working on an answer to this question for many years (Humphrey 1992, 2006, 2011). I've wanted an answer that takes account of evolutionary history. This isn't the place to give you the full story, but I'll try to give a brief overview.

It begins, as I see it, with the creatures that were our far distant ancestors, floating in the seas, making evaluative responses to stimuli at the body surface: "wriggles of acceptance or rejection." These responses, to which I've given the general name "sentition", have been honed by natural selection, so as to be well adapted to the creature's needs—taking account of what kind of stimulus is reaching the body surface, what part of the body is affected, and what import this has for biological well-being. From the start then, the responses can be said to be *meaningful*—which is to say they potentially carry a lot of information about what the stimulation means for the creature. However, to begin with, there is no one at home in the brain to realise this potential, no one to *take an interest* in the meaning.

But, evolution is inventive. Before long there arises in the brain a special module —a proto self, if you like—whose job is exactly that: to discover "what the stimulation means for me". And, as luck would have it, it turns out it can do this by the simple trick of *reading—extracting the meaning from—the motor command signals* being sent out to produce the reflex response.

So now, we have an agent who is reading the brain's own responses and making a sensory interpretation of them. In truth this is the first *subject of sensation.* But let's note there is nothing fancy or magical about the interpretation at this stage. The subjective experience does not have had any special phenomenal feel. What happened?

I've argued that the key lay in how sentition went on evolving. Back at the start, the reflex responses are overt bodily actions occurring at the site of stimulation at the body surface. However things are never going to stay like this. As the descendants of the original creatures evolve to be more sophisticated, these overt responses soon enough become inappropriate, even inconvenient—you don't always want to grimace when you're touched by red light, say. So now the creature faces a problem. How to lose the bodily behaviour but keep the information about the meaning of the stimulus?

The solution natural selection hits on is ingenious. It is for the responses to become internalised, or "privatised," such that the motor signals no longer reach the actual body surface, but rather begin to target the body-map where the sense organs first project to the brain. Thus sentition evolves from being an actual form of bodily expression to being a *virtual* one—yet still a response that the subject can milk for information.

Now, this privatisation has a remarkable—if fortuitous—result. It means that a feedback loop is created between motor and sensory regions of the brain—a loop that has the capacity to sustain recursive activity, going round and round, catching its own tail. And this, as I see it, has been game-changing. Crucially, it means that the activity can be drawn out in time, so as to create the "thick moment" of sensory experience. But, more than this, the activity can be channelled and stabilised, so as to create a mathematically complex "attractor" state. And such an attractor can have remarkable hyper-dimensional properties (Krisztin 2008). Real, unreal, surreal? The answer will be in the eye of the beholder—the subject whose reading of this brain activity is giving rise to the sensory experience.

At any rate, from now on, whenever the opportunity arises to "improve" the quality of sensations—to make further adaptive changes—natural selection has a whole new design space to explore. Small adjustments to the circuitry can have dramatic effects. And this provides the evolutionary context, I believe, for the invention of a special kind of attractor that will be read by the subject as a sensation with an unaccountable *phenomenal feel.* On the analogy of the Gregundrum, I've called this attractor the "ipsundrum", to signify a real "impossible brain state" that is actually self made. The ipsundrum is still a species of sentition, that originates as a response to sensory stimulation, and still carries information about the objective properties of the stimulation. But this information now comes in a remarkable new guise. It comes, if you like, as part of "a riddle written on the brain" (Humphrey 2016a).

As I mentioned, I put forward this account of sensations more than twenty five years ago. My arguments were largely theoretical, rather than empirical. But I'm happy to say it looks as if the key features has been getting experimental backing: namely that visual sensations depend on brain activity in a loop running between primary visual cortex and areas further forward. In a masterly review of recent neuroscientific evidence, Stan Dehaene (who, oddly enough, is something of a "qualia denier") sums up the picture he sees emerging: "Consciousness lives in the loops: reverberating neuronal activity, circulating in the web of our cortical connections, causes our conscious experiences" (Dehaene 2014, p. 156).

So there we have it: my glimmer of a theory of what gives consciousness its astonishing quality. With so much of the detail missing, I acknowledge it's not much more than a glimmer. But it must be better than no theory at all. Colin McGinn has written: "It is not that we know what would explain consciousness but are having trouble finding the evidence to select one explanation over the others; rather, we have no idea what an explanation of consciousness would even look like" (McGinn 1999, p. 61). I humbly suggest that's no longer true.

This is all I have to say for now about how a physical system could deliver conscious experience. However, for an evolutionist, of course it's too soon to wrap up the discussion. We may have found a possible answer to the question of *what* evolved, but we haven't yet begun to address the question of *why* it evolved. Even if we did know all the detail—if we could explain how conscious experience is created neuron by neuron, from red light touching your retina through to your making all the claims you do about the red qualia—we still would not know *what this is good for*. What can possibly have been the biological advantage, the contribution to fitness, of dressing up sensations in this provocatively mysterious way?

It's a real problem. Let's return to the idea of consciousness as an invention. Under the first meaning of invention we saw that consciousness could indeed be considered to be "a cognitive faculty, evolved by natural selection, designed to help us make sense of ourselves and our surroundings." But now, when we consider the role of qualia, this meaning of invention looks much less of a good fit. At first sight at least, *qualia* are neither cognitive, nor helpful!

Jerry Fodor has stated the difficulty in his typically blunt way: "Consciousness"—and it's clear he's referring to qualia in particular—"seems to be among the chronically unemployed. As far as anybody knows, anything that our conscious minds can do they could do just as well if they weren't conscious. Why then did God bother to make consciousness?" (Fodor 2004). John Searle has made much the same claim, about qualia having no impact at the level of behaviour: "As far as the ontology of consciousness is concerned, behaviour is simply irrelevant. We could have identical behaviour in two different systems, one of which is conscious and the other totally unconscious." (Searle 1992).

If these philosophers are right, it would mean that consciousness—at least its phenomenal side—could not have had any impact on our ancestors' survival. In which case the genes specifying the underlying brain circuits could not have been selected by natural selection.

Then, *are* these philosophers right? I think the plain answer is, No. They are guilty of a massive failure of imagination.

Fodor says qualia are "unemployed". He seems to take it for granted that, if consciousness does have a job to do, this can only be to provide us with some special kind of skill—helping us to act more intelligently or more efficiently in the service of some practical goal. But what if this notion of employment is simply not appropriate when discussing the phenomenal aspect of consciousness? What if phenomenal consciousness, rather than making us more intelligent or more productive *on the outside,* makes us somehow *bigger on the inside*—emotionally and spiritually bigger? What if consciousness is actually an invention in the second sense I mentioned at the start: "a fantasy, conjured up by the brain, designed to change designed to change how we value what becomes of us"?

Think about it. Think again about the real impossible triangle, the Gregundrum. Why, for what purpose, did Richard Gregory, invent this brilliant illusion? It surely wasn't to serve any practical purpose. There's a photo showing him with his face framed by the real impossible object (Gregory 2011). Look at his broad smile. He did it simply to *amaze us.* Then, could it be that Nature, when she invented qualia, did it so that we conscious creatures should *amaze ourselves*??

Don't get me wrong. I am a card-carrying Darwinian reductionist. I've no wish to get off the explanatory hook by substituting fuzzy answers for clear ones. But still, I do think there are times when, in the interests of science, we need to loosen up a bit. Before we pronounce on the employability of phenomenal consciousness, we need to undertake a proper natural history. We should be studying how conscious experience actually changes the way people live in the world. How does exposure to qualia change people's psychology? What beliefs and attitudes are generated? How does it affect people's ideas about who and what they are, and what kind of world they live in?

These are—or ought to be—empirical questions to be asked of ordinary people. And we should be ready to consider all sorts of possible answers, not just those we'd find discussed in the science or philosophy section of the library but perhaps those that belong in the self-help section, or even the New Age. But, most important, we should begin the inquiry close to home, by taking seriously our own intuitions about just how and why phenomenal consciousness matters to ourselves.

Think about it. Suppose the magic for *you* were *not* there. Suppose your sensations were in fact just brown bag numbers. What would be missing from your life?

It's clear to me that in such a semi-zombie state I—you—would lose out, on several levels. First, you'd lose your psychological essence, your core self. Next, you'd lose your sense of intimacy with things in the outside world. And then, finally, you'd lose your soul, and other humans would lose their souls as well.

Self

We saw, early on, how the binding of sub-selves leads to the creation of the core self as the singular subject of a range of mental states. But, now let me say it, even when all the sub-selves are gathered together, the larger self is by no means secure.

A self, stripped of sensations, would remain a pretty anaemic kind of self. But add in the qualia, and everything changes. By lifting sensory experience onto that mysterious, non-physical plane, qualia deepen and enrich your sense of your own presence. You find yourself living in thick time. So you become the owner of a self that you want to expand and preserve for its own sake—in short, a *self worth having*. Take away this primary sense of your own presence, and your existence would simply be less well-founded, less convincing—to you and everyone else

World

Next, though this isn't so obvious, you'd lose the external world—at least the world as you've come to know and love it. Even though it's your own brain that creates the qualia, you can't but project the special qualities of sensations out onto the objects of perception in the outside world. In doing so, you spread a kind of fairy-dust around you. You enchant the world. Take away this magic paintbrush, and the world would lose much of its significance. You'd find it a less awesome place, less fun, less promising.

Soul

You did it. It's all yours. The things out there, experienced through bodily sensation, are singing your song. It's bound to dawn on you that when you pay homage to the beauties of nature you are really paying homage to yourself. So, by a strange inversion, the magical world you've made returns the compliment and further enhances your sense of your own significance. Then add in the poetry of human culture, and by one path or another, your core self becomes elaborated into that marvellous cultural construct: the human soul. A soul that, with your generous theory of mind, you recognise in other people too.

Now, I *will* draw this to a close. Earlier, when I quoted Shakespeare's prologue, I omitted the first lines. They read.

O for a Muse of fire, that would ascend
The brightest heaven of invention.

The Chorus means "invention" in the second sense: he's seeking permission for the actors to create an extraordinary work of fiction on stage. I like to think that Nature did it first. Qualia are just such an invention, arguably the brightest heaven—the most remarkable story that anyone has ever dared to tell. Thanks to natural selection, we all contain within ourselves that muse of fire.

References

G. Byron, Letter to Annabella Milbanke, in *The Bride of Science: Romance, Reason and Byron's Daughter*, Quoted by B. Woolley (1999). (MacMillan, London 1813), p. 28

F. Crick, C. Koch, What is the function of the claustrum? Phil. Trans. Roy. Soc. B **360**, 1271–1279 (2005)

S. Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Viking Penguin, New York, 2014)

D. Dennett, *Consciousness Explained* (Little Brown, New York, 1991)

W. Empson, *Seven Types of Ambiguity* (Chatto and Windus, London, 1930), p. 9

J. Fodor, *In: Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind* (MIT Press, Cambridge, MA, 1998), p. 83

J. Fodor, You can't argue with a novel. London Rev. Books, March 3 issue, 31 (2004)

K. Frankish, Illusionism as a theory of consciousness. J. Conscious. Stud. (2016) in press

R. Gregory, Gregorian reflections (2011). Image available at http://www.richardgregory.org/gregorian-reflections.htm

Harvard Natural Sciences Demonstrations, Synchronization of metronomes (2016), https://youtu.be/Aaxw4zbULMs

N. Humphrey, *A History of the Mind* (Chatto and Windus, London, 1992)

N. Humphrey, One-self: a meditation on the unity of consciousness. Soc. Res. **67**, 32–39 (2000)

N. Humphrey, *Seeing Red: A Study in Consciousness* (Harvard University Press, Cambridge, MA, 2006)

N. Humphrey, Soul Dust: The Magic of Consciousness. (Princeton University Press, 2011)

N. Humphrey, A riddle written on the brain. J. Conscious. Stud. **23**, 278–287 (2016a)

N. Humphrey, Redder than red: illusionism or phenomenal surrealism. J. Conscious. Stud. (2016b) in press

T. Krisztin, Global dynamics of delay differential equations. Periodica Math. Hung. **56**, 83–95 (2008)

J. Maynard, Is The Universe A Hologram? Holographic Principle Suggests 'Yes'. Tech Times, 27th April (2015)

C. McGinn, *The Mysterious Flame: Conscious Minds in a Material World* (Basic Books, New York, 1999)

C. McGinn, Consciousness and cosmology: hyperdualism ventilated, in *Consciousness* ed. by M. Davies, G.W. Humphreys (Oxford, Blackwell, 1993), pp. 155–177, p. 160

M. Minsky, *The Society of Mind* (Simon & Schuster, New York, 1986)

Nagel, *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False* (Oxford University Press, New York, 2012), p. 35

I. Newton, A letter from Mr. Isaac Newton containing his new theory about light and colours. Phil. Trans. Roy. Soc. **6**, 3075–87 (1671)

H.A. Orr, Awaiting a New Darwin. (New York Rev Books, 2013) Feb 7 issue

J. Searle, *The Rediscovery of the Mind* (MIT Press, Cambridge, MA, 1992), p. 71

G. Tononi, The integrated information theory of consciousness: an updated account. Arch. Ital. Biol. **150**, 56–90 (2012)

Chapter 26
The Fantasy of First-Person Science

Daniel C. Dennett

A week ago, I heard James Conant give a talk at Tufts, entitled "Two Varieties of Skepticism" in which he distinguished two oft-confounded questions:

> Descartes: How is it possible for me to tell whether a thought of mine is true or false, perception or dream?
>
> Kant: How is it possible for something even to *be* a thought (of mine)? What are the conditions for the possibility of experience (veridical or illusory) at all?

Conant's excellent point was that in the history of philosophy, up to this very day, we often find philosophers talking past each other because they don't see the difference between the Cartesian question (or family of questions) and the Kantian question (or family of questions), or because they try to merge the questions. I want to add a third version of the question:

Turing: How could we make a robot that had thoughts, that learned from "experience" (interacting with the world) and used what it learned the way we can do?

There are two main reactions to Turing's proposal to trade in Kant's question for his.

A. Cool! Turing has found a way to actually *answer* Kant's question!
B. *Aaaargh*! Don't fall for it! You're leaving out … *experience!*

I'm captain of the A team (along with Quine, Rorty, Hofstadter, the Churchlands, Andy Clark, Lycan, Rosenthal, Harman, and many others). I think the A team wins, but I don't think it is *obvious*. In fact, I think it takes a rather remarkable exercise of the imagination to see how it might even be possible, but I do think one

(a written version of a debate with David Chalmers, held at Northwestern University, Evanston, IL,. supplemented by an email debate with Alvin Goldman). © Dennett 2001, Feb 15.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-72478-2_26) contains supplementary material, which is available to authorized users.

D. C. Dennett (✉)
Center for Cognitive Studies, Tufts University, Medford, MA 02155, USA

© Springer International Publishing AG 2018
S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_26

can present a powerful case for it. As I like to put it, *we* are robots made of robots—we're each composed of some few trillion robotic cells, each one as mindless as the molecules they're composed of, but working together in a gigantic team that creates *all the action* that occurs in a conscious agent. Turing's great contribution was to show us that Kant's question could be recast as an *engineering* question. Turing showed us how we could trade in the first-person perspective of Descartes and Kant for the third-person perspective of the natural sciences and *answer all the questions* —without philosophically significant residue.

David Chalmers is the captain of the B team, (along with Nagel, Searle, Fodor, Levine, Pinker, Harnad and many others). He insists that he just *knows* that the A team leaves out consciousness. It doesn't address what Chalmers calls the Hard Problem. How does he know? He says he just does. He has a gut intuition, something he has sometimes called "direct experience." I know the intuition well. I can feel it myself. When I put up Turing's proposal just now, if you felt a little twinge, a little shock, a sense that your pocket had just been picked, you know the feeling too. I call it the Zombic Hunch (Dennett 2001). I feel it, but I don't credit it. I figure that Turing's genius permitted him to see that we can *leap over* the Zombic Hunch. We can come to see it, in the end, as a misleader, a roadblock to understanding. We've learned to dismiss other such intuitions in the past–the obstacles that so long prevented us from seeing the Earth as revolving around the sun, or seeing that living things were composed of non-living matter. It still *seems* that the sun goes round the earth, and it still *seems* that a living thing has some extra spark, some extra ingredient that sets it apart from all non-living stuff, but we've learned not to credit those intuitions. So now, do you want to join me in leaping over the Zombic Hunch, or do you want to stay put, transfixed by this intuition that won't budge? I will try to show you how to join me in making the leap.

Are You *Sure* There Is Something Left Out?

In *Consciousness Explained*, (Dennett 1991) I described a method, *heterophenomenology*, which was explicitly designed to be

> the *neutral* path leading from objective physical science and its insistence on the third-person point of view, to a method of phenomenological description that can (in principle) do justice to the most private and ineffable subjective experiences, while never abandoning the methodological principles of science. (*CE*, p72.)

How does it work? We start with recorded raw data. Among these are the vocal sounds people make (what they say, in other words), but to these verbal reports must be added all the other manifestations of belief, conviction, expectation, fear, loathing, disgust, etc., including any and all internal conditions (e.g. brain activities, hormonal diffusion, heart rate changes, etc.) detectable by objective means.

I guess I should take some of the blame for the misapprehension, in some quarters, that heterophenomenology *restricts itself* to verbal reports. Nothing could be further from the truth. Verbal reports are different from all other sorts of raw data precisely in that they admit of (and require, according to both heterophenomenology and the 1st-person point of view) interpretation as speech acts, and subsequent assessment as expressions of belief about a subject's "private" subjective state. And so my discussion of the methodology focused on such verbal reports in order to show how they are captured within the fold of standard scientific ("3rd-person") data. But all other such data, all behavioral reactions, visceral reactions, hormonal reactions, and other changes in physically detectable state are included within heterophenomenology. I thought that went without saying, but apparently these additional data are often conveniently overlooked by critics of heterophenomenology.

From the recorded verbal utterances, we get transcripts (e.g., in English or French, or whatever), from which in turn we devise interpretations of the subjects' speech acts, which we thus get to treat as (apparent) expressions of their beliefs, on all topics. Thus using the intentional stance (Dennett 1971, 1987), we construct therefrom the subject's heterophenomenological world. We move, that is, from *raw* data to *interpreted* data: a catalogue of the subjects' convictions, beliefs, attitudes, emotional reactions, … (together with much detail regarding the circumstances in which these intentional states are situated), but then we adopt a special move, which distinguishes heterophenomenology from the normal interpersonal stance: the subjects' beliefs (etc.) *are all bracketed for neutrality.*

Why? Because of two failures of overlap, which we may label false positive and false negative. False positive: Some beliefs that subjects have about their own conscious states are provably false, and hence what needs explanation in these cases is the *etiology of the false belief.*

For instance, most people—naive people—think their visual fields are roughly uniform in visual detail or grain all the way out to the periphery. Even sophisticated cognitive scientists can be startled when they discover just how poor their capacity is to identify a peripherally located object (such as a playing card held at arm's length). It certainly *seems* as if our visual consciousness is detailed all the way out all the time, but easy experiments show that it isn't. (Our color vision also seems to extend all the way out, but similar experiments show that it doesn't.) So the question posed by the heterophenomenologist is:

Why do people think their visual fields are detailed all the way out? not this question: How come, since people's visual fields are detailed all the way out, they can't identify things parafoveally?

False negative: Some *psychological things that happen in people* (to put it crudely but neutrally) are *unsuspected by those people*. People not only *volunteer* no information on these topics; when provoked to search, they *find* no information on these topics. But a forced choice guess, for instance, reveals that nevertheless, there is something psychological going on. This shows, for instance, that they *are* being influenced by the meaning of the masked word even though they are, as they put it, entirely unaware of any such word. (One might put this by saying that there is a lot of unconscious mental activity—but this is tendentious; to some, it might be held to beg the vexed question of whether people are briefly conscious of these evanescent and elusive topics, but just hugely and almost instantaneously forgetful of them.)

Now faced with these failures of overlap—people who believe they are conscious of more than is in fact going on in them, and people who do not believe they are conscious of things that are in fact going on in them—heterophenomenology maintains a nice neutrality: it characterizes their beliefs, their heterophenomenological world, without passing judgment, and then investigates to see what could explain *the existence of those beliefs*. Often, indeed typically or normally, the existence of a belief is explained by confirming that it is a *true* belief provoked by the normal operation of the relevant sensory, perceptual, or introspective systems. Less often, beliefs can be seen to be true only under some arguable metaphorical interpretation—the subject claims to have manipulated a mental image, and we've found a quasi-imagistic process in his brain that can support that claim, if it is interpreted metaphorically. Less often still, the existence of beliefs is explainable by showing how they are illusory byproducts of the brain's activities: it only *seems* to subjects that they are reliving an experience they've experienced before (*dêja vu*).

> In this chapter we have developed a *neutral* method for investigating and describing phenomenology. It involves extracting and purifying *texts* from (apparently) speaking *subjects*, and using those texts to generate a theorist's fiction, the subject's *heterophenomenological world*. This fictional world is populated with all the images, events, sounds, smells, hunches, presentiments, and feelings that the subject (apparently) sincerely believes to exist in his or her (or its) stream of consciousness. Maximally extended, it is a neutral portrayal of exactly *what it is like to be* that subject–in the subject's own terms, given the best interpretation we can muster….. People undoubtedly do believe that 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. (*CE*, p98)

Is this truly neutral, or does it bias our investigation of consciousness by stopping one step short? Shouldn't our *data* include not just subject's subjective *beliefs about their experiences*, but the *experiences themselves*? Levine, a first-string member of the B Team, insists "that conscious experiences themselves, not merely our verbal judgments about them, are the primary data to which a theory must answer." (Levine 1994)

This is an appealing idea, but it is simply a mistake. First of all, remember that heterophenomenology gives you much more data than just a subject's verbal judgments; every blush, hesitation, and frown, as well as all the covert, internal reactions and activities that can be detected, are included in our primary data. But what about this concern with leaving the "conscious experiences themselves" out of the primary data? Defenders of the first-person point of view are not entitled to this complaint against heterophenomenology, since *by their own lights,* they should prefer heterophenomenology's treatment of the primary data to any other. Why? Because it does justice to both possible sources of non-overlap. On the one hand, if some of your conscious experiences occur unbeknownst to you (if they are experiences about which you have no beliefs, and hence can make no "verbal judgments"), then they are just as inaccessible to your first-person point of view as they are to heterophenomenology. *Ex hypothesi*, you don't even suspect you have them —if you did, you could verbally express those suspicions. So heterophenomenology's list of primary data doesn't leave out any conscious experiences you know of, or even have any first-person inklings about. On the other hand, unless you claim not just reliability but outright infallibility, you should admit that some—just some —of your beliefs (or verbal judgments) about your conscious experiences might be wrong. In all such cases, however rare they are, what has to be explained by theory is not the conscious experience, but your belief in it (or your sincere verbal judgment, etc.). So heterophenomenology doesn't include any spurious "primary data" either, but plays it safe in a way you should approve.

Heterophenomenology is nothing but good old 3rd-person scientific method applied to the particular phenomena of human (and animal) consciousness. Scientists who were interested in taking the first-person point of view seriously figured out how to do just that, bringing the data of the first person into the fold of objective science. I didn't invent the method; I merely described it, and explained its rationale.

Alvin Goldman has recently challenged this claim. In "Science, Publicity and Consciousness" (1997), he says that heterophenomenology is not, as I claim, the standard method of consciousness research, since researchers "rely substantially on subjects' introspective beliefs about their conscious experience (or lack thereof)" (p. 532). In private correspondence (Feb 21, 2001) he has elaborated his claim thus:

> The objection lodged in my paper to heterophenomenology is that what cognitive scientists *actually* do in this territory is not to practice agnosticism. Instead, they rely substantially on subjects' introspective beliefs (or reports). So my claim is that the heterophenomenological method is not an accurate description of what cognitive scientists (of consciousness) standardly do. Of course, you can say (and perhaps intended to say, but if so it wasn't entirely clear) that this is what scientists *should* do, not what they *do* do.

I certainly would play the role of reformer if it were necessary, but Goldman is simply mistaken; the adoption of agnosticism is so firmly built into practice these days that it goes without saying, which is perhaps why he missed it. Consider, for instance, the decades-long controversy about mental imagery, starring Shepard, Kosslyn, and Pylyshyn among many others. If agnosticism were not the tacit order of the day, Kosslyn would have never needed to do his well-known experiments to support subjects' claims that what they were doing (at least if described metaphorically) really was a process of image manipulation. (The issues are not settled yet, of course.) In psychophysics, the use of signal detection theory has been part of the canon since the 1960s, and it specifically commands researchers to control for the fact that the response criterion is under the subject's control although the subject is not himself or herself a reliable source on the topic. Or consider the voluminous research literature on illusions, both perceptual and cognitive, which standardly assumes that the data are what subjects *judge* to be the case, and *never* makes the mistake of "relying substantially on subjects' introspective beliefs." The diagnosis of Goldman's error is particularly clear here: of course experimenters on illusions rely on subjects' introspective beliefs (as expressed in their judgments) about *how it seems to them,* but that *is* the agnosticism of heterophenomenology; to go beyond it would be, for instance, to assume that in size illusions *there really were visual images of different sizes somewhere in subjects' brains (or minds)*, which of course no researcher would dream of doing. Finally, consider such phenomena as *déja vu.* Sober research on this topic has never made the mistake of abandoning agnosticism about subjects' claims to be reliving previous experiences. See, e.g., Bower and Clapper, in 1989, for instance, or any good textbook on methods in cognitive science for the details. (Goldman has responded to this paragraph in a series of emails to me, which I have included in an Appendix.)

A bounty of excellent heterophenomenological research has been done, is being done, on consciousness. See, e.g., the forthcoming special issue of *Cognition*, edited by Stanislas Dehaene, on the cognitive neuroscience of consciousness. It contains a wealth of recent experiments all conducted within the methodological strictures of heterophenomenology, whose resolutely 3rd-person treatment of belief attribution squares perfectly with standard scientific method: when we assess the attributions of belief relied upon by experimenters (in preparing and debriefing subjects, for instance) we use precisely the principles of the intentional stance to settle what it is reasonable to postulate regarding the subjects' beliefs and desires. Now Chalmers has objected (in the debate) that this "behavioristic" treatment of belief is itself question-begging against an alternative vision of belief in which, for instance, "having a phenomenological belief doesn't involve just a pattern of responses, but often requires having certain experiences." (personal correspondence, 2/19/01). On the contrary, heterophenomenology is neutral on just this score, for surely we mustn't *assume* that Chalmers is right that there *is* a special category of "phenomenological" beliefs—that there is a kind of belief that is off-limits to "zombies" but not to us conscious folks. Heterophenomenology allows us to proceed with our catalogue of a subject's beliefs leaving it open whether any or all of them are Chalmers-style phenomenological beliefs or mere zombie-beliefs. (More on this later.) In fact, heterophenomenology permits science to get on with the business of accounting for the patterns in all these subjective beliefs without stopping to settle this imponderable issue. And surely Chalmers must admit that the patterns in these beliefs are among the phenomena that any theory of consciousness must explain.

Let's look at a few cases of heterophenomenology in action (Please refer to the accompanying video "Ramachandran's Video" present at the link mentioned at the beginning of the article). Do you see the motion? You see apparent motion. Does the yellow blob really move? The blob *on the screen* doesn't move. Ah, but does the *subjective* yellow blob in your experience move? Does it *really* move, or do you just *judge* that it moves? Well, it sure seems to move! That is what you judge, right? Now perhaps there are differences in how you would word your judgments. And perhaps there are other differences. Perhaps some of you not only judge that it seems to move, but are made slightly dizzy or nauseated by the apparent motion. Perhaps some people get motion sickness from motion capture and others don't. Perhaps some of you don't even experience the apparent motion at all. Perhaps some of you can use such apparent motion just like real motion, to help disambiguate shapes, for instance, and perhaps you can't. We can explore these variations in as much detail as you like, and can come back to you again and again with further inquiries, further tests, further suggested distinctions.

> You are *not* authoritative about what is happening in you, but only about what *seems* to be happening in you, and we are giving you total, dictatorial authority over the account of how it seems to you, about *what it is like to be you*. And if you complain that some parts of how it seems to you are ineffable, we heterophenomenologists will grant that too. What better grounds could we have for believing that you are unable to describe something than that (1) you don't describe it, and (2) confess that you cannot? Of course you might be lying, but we'll give you the benefit of the doubt. (*CE*, p96-7)

Is there anything about your experience of this motion capture phenomenon that is not explorable by heterophenomenology? I'd like to know what. This is a fascinating and surprising phenomenon, *predicted* from the 3rd-person point of view, and eminently studiable via heterophenomenology. (Tom Nagel once claimed that 3rd-person science might provide us with brute correlations between subjective experiences and objective conditions in the brain, but could never *explain* those correlations, in the way that chemists can explain the correlation between the liquidity of water and its molecular structure. I asked him if he considered the capacity of industrial chemists to predict the molar properties of novel artificial polymers in advance of creating them as the epitome of such explanatory correlation, and he agreed that it was. Ramachandran and Gregory predicted this motion capture phenomenon, an entirely novel and artificial subjective experience, on the basis of their knowledge of how the brain processes vision.)

See next Rensink's change blindness (Please refer to the accompanying video "Rensink's Video" present at the link mentioned at the beginning of the article). (By the way, this is an effect I predicted in *CE,* much to the disbelief of many readers.)

Were your qualia changing before you noticed the flashing white cupboard door? You saw each picture several dozen times, and *eventually* you saw a change that was "swift and enormous" (Dennett 1999; Palmer 1999) but that swift, enormous change was going on for a dozen times and more before you noticed it. Does it count as a change in color qualia?

The possible answers:

A. Yes.
B. No.
C. I don't know

   1. because I now realize I never knew quite what I meant by "qualia" all along.
   2. because although I know just what I have always meant by "qualia", I have no first-person access to my own qualia in this case.

      a. and 3rd-person science can't get access to qualia either!

Let's start with option C first. Many people discover, when they confront this case, that since they never imagined such a phenomenon was possible, they never considered how their use of the term "qualia" should describe it. They discover a heretofore unimagined flaw in their concept of qualia—rather like the flaw that physicists discovered in their concept of *weight* when they first distinguished weight from mass. The philosophers' concept of qualia is a mess. Philosophers don't even agree on how to apply it in dramatic cases like this. I hate to be an old I-told-you-so but I told you so ("Quining Qualia"). This should be at least mildly embarrassing to our field, since so many scientists have recently been persuaded by philosophers that they should take qualia seriously—only to discover that philosophers don't come close to agreeing among themselves about *when* qualia—*whatever* they are—are present. (I have noticed that many scientists who think they are newfound friends of qualia turn out to use the term in ways no self-respecting qualophile will countenance.)

But although some philosophers may now concede that they aren't so sure what they meant by "qualia" all along, others are very sure what concept of qualia they've been using all along, so let's consider what they say. Some of them, I have learned, have *no problem* with the idea that their very own qualia could change radically without their noticing. *They* mean by "qualia" something to which their 1st-person access is variable and problematic. If you are one of those, then heterophenomenology is your preferred method, since it, unlike the first-person point of view, can actually study the question of whether qualia change in this situation. It is going to be a matter of some delicacy, however, how to *decide* which brain events count for what. In this phenomenon of change blindness for color changes, for instance, we know that the color-sensitive cones in the relevant region of your retina were flashing back and forth, in perfect synchrony with the white/brown quadrangle, and presumably (we should check) other, later areas of your color vision system were also shifting in time with the external color shift. But if we keep looking, we will also presumably find yet other areas of the visual system that only come into synchrony after you've noticed. (such effects have been found in similar fMRI studies, e.g. O'Craven et al. 1997).

The hard part will be deciding (on what grounds?) which features of which states to declare to be qualia and why. I am not saying there can't be grounds for this. I can readily imagine there being *good* grounds, but if so, then those will be grounds for adopting/endorsing a 3rd-person concept of qualia (cf. the discussion of Chase and Sanborn in Dennett 1988, or the beer-drinkers in CE 395-6). The price you have to pay for obtaining the *support* of 3rd-person science for your conviction about how it is/was with you is straightforward: you have to grant that *what you mean* by how it is/was with you is something that 3rd-person science could either support or show to be mistaken. Once we adopt any such concept of qualia, for instance, we will be in a position to answer the question of whether color qualia shift during change blindness. And if some subjects in our apparatus tell us that their qualia do shift, while our brain-scanner data shows clearly that they don't, we'll treat these subjects as simply wrong *about their own qualia*, and we'll explain why and how they come to have this false belief.

Some people find this prospect inconceivable. For just this reason, some people may want to settle for option B: No, my qualia don't change—*couldn't* change—until I notice the change. This decision *guarantees* that qualia, tied thus to noticing, are securely within the heterophenomenological worlds of subjects, are indeed constitutive features of their heterophenomenological worlds. On option B, what subjects can say about their qualia fixes the data.[1]

By a process of elimination, that leaves option A, YES, to consider. If you think your qualia did change (though you didn't notice it at the time) *why* do you think this? Is this a theory of yours? If so, it needs evaluation like any other theory. If not, did it just come to you? A gut intuition? Either way, your conviction is a prime candidate for heterophenomenological diagnosis: what has to be explained is how you came to have this belief. The last thing we want to do is to treat your claim as incorrigible. Right?

Here is the dilemma for the B Team, and Captain Chalmers. If you eschew incorrigibility claims, and especially if you acknowledge the competence of 3rd-person science to answer questions that can't be answered from the 1st-person point of view, your position collapses into heterophenomenology. The only remaining alternative, C(2a), is unattractive for a different reason. You can protect qualia from heterophenomenological appropriation, but only at the cost of declaring them outside science altogether. If qualia are so shy they are not even accessible from the 1st-person point of view, then no 1st-person science of qualia is possible either.

I will not contest the existence of first-person *facts* that are unstudiable by heterophenomology and other 3rd-person approaches. As Steve White has reminded me, these would be like the humdrum "inert historical facts" I have spoken of elsewhere—like the fact that some of the gold in my teeth once belonged to Julius Caesar, or the fact that none of it did. One of those is a fact, and I daresay no possible extension of science will ever be able to say which is the truth. But if 1st-person facts are like inert historical facts, they are no challenge to the claim that heterophenomenology is the maximally inclusive science of consciousness, because they are unknowable even to the 1st-person they are about!

[1] Consider Option B for the simpler case raised earlier. Do you want to cling to a concept of visual consciousness according to which your conviction that your visual consciousness is detailed all the way out is *not* contradicted by the discovery that you cannot identify large objects in the peripheral field? You *could* hang tough: "Oh, all that you've shown is that we're not very good at identifying objects in our peripheral vision; *that* doesn't show that peripheral consciousness isn't as detailed as it seems to be! All you've shown is that a *mere behavioral capacity* that one might mistakenly have thought to coincide with consciousness doesn't, in fact, show us anything about consciousness!" Yes, if you are careful to *define* consciousness so that nothing "behavioral" can bear on it, you get to declare that consciousness transcends "behaviorism" without fear of contradiction. See "Are we Explaining Consciousness Yet?" for a more detailed account of this occasionally popular but hopeless move.

David Chalmers as a Heterophenomenological Subject

Of course it *still* seems to many people that heterophenomenology must be leaving something out. That's the ubiquitous Zombic Hunch. How does the A team respond to this? Very straightforwardly: by *including* the Zombic Hunch among the heartfelt convictions any good theory of consciousness must explain. One of the things that it falls to a theory of consciousness to explain is *why some people are visited by the Zombic Hunch.* Chalmers is one such, so let's look more closely at the speech acts Chalmers has offered as a subject of heterophenomenological investigation.

Here is Chalmers' definition of a zombie (his zombie twin):

> Molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely… he is embedded in an identical environment. He will certainly be identical to me *functionally*; he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting…. he will be awake, able to report the contents of his internal states, able to focus attention in various places and so on. It is just that none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. There is nothing it is like to be a Zombie… 1996, p95

Notice that Chalmers allows that zombies have internal states with contents, which the zombie can report (sincerely, one presumes, believing them to be the truth); these internal states have contents, but not conscious contents, only pseudo-conscious contents. The Zombic Hunch, then, is Chalmers' conviction that he has just described a real problem. It *seems to him* that there is a problem of how to explain the difference between him and his zombie twin.

> The justification for my belief that I am conscious lies not just in my cognitive mechanisms but also in *my direct evidence* [emphasis added]; the zombie lacks that evidence, so his mistake does not threaten the grounds for our beliefs. (One can also note that the zombie doesn't have the same beliefs as us, because of the role that experience plays in constituting the contents of those beliefs.) (Reply to Searle)

This speech act is curious, and when we set out to interpret it, we have to cast about for a charitable interpretation. How does Chalmers' justification *lie in* his "direct evidence"? Although he says the zombie *lacks* that evidence, nevertheless the zombie *believes* he has the evidence, just as Chalmers does. Chalmers and his zombie twin are heterophenomenological twins: when we interpret all the data we have, we end up attributing to them exactly the same heterophenomenological worlds. Chalmers fervently believes he himself is not a zombie. The zombie fervently believes he himself is not a zombie. Chalmers *believes* he gets his justification from his "direct evidence" of his consciousness. So does the zombie, of course.

The zombie has the conviction that he has direct evidence of his own consciousness, and that this direct evidence is his justification for his belief that he is conscious. Chalmers must maintain that the zombie's conviction is false. He says that the zombie *doesn't* have the same beliefs as us "because of the role that experience plays in constituting the contents of those beliefs," but I don't see how this can be so. Experience (in the special sense Chalmers has tried to introduce) plays no role in constituting the contents of those beliefs, since *exhypothesi,* if experience (in this sense) were eliminated—if Chalmers were to be suddenly zombified—he would go right on saying what he says, insisting on what he now insists on, and so forth.[2] Even if his "phenomenological beliefs" suddenly ceased to be phenomenological beliefs, he would be none the wiser. It would not *seem to him* that his beliefs were no longer phenomenological.

But wait, I am forgetting my own method and arguing with a subject! As a good heterophenomenologist, I must grant Chalmers full license to his deeply held, sincerely expressed convictions and the heterophenomenological world they constitute. And then I must undertake the task of explaining the etiology of his beliefs. Perhaps Chalmers' beliefs about his experiences will turn out to be true, though how that prospect could emerge eludes me at this time. But I will remain neutral. Certainly we shouldn't give them incorrigible status. (He's not the Pope.) The fact that some subjects have the Zombic Hunch shouldn't be considered grounds for revolutionizing the science of consciousness.[3]

Where's the Program?

That leaves the B Team in a bit of a predicament. Chalmers would like to fulfil the Philosopher's Dream:

> *To prove* a priori, *from one's ivory tower, a metaphysical fact that forces a revolution in the sciences.*

It is not an impossible dream. (That is, it is not logically impossible.) Einstein's great insight into relativity comes tantalizingly close to having been a purely philosophical argument, something a philosopher might have come up with just from first principles. And Patrick Matthew could claim with some justice in 1860 to have scooped Darwin's theory of natural selection in 1831 by an act of pure reason:

> it was by a general glance at the scheme of Nature that I estimated this select production of species as an a priori recognizable fact–an axiom, requiring only to be pointed out to be admitted by unprejudiced minds of sufficient grasp. [see DDI, p49]

[2] "I simply say that invoking consciousness is not necessary to explain actions; there will always be a physical explanation that does not invoke or imply consciousness. A better phrase would have been 'explanatorily superfluous', rather than 'explanatorily irrelevant.'" (Chalmers' second reply to Searle, on his website).

[3] Chalmers seems to think that conducting surveys of his audiences, to see what proportion can be got to declare their allegiance to the Zombic Hunch, yields important data. Similar data-gathering would establish the falsehood of neo-Darwinian theory and the existence of an afterlife.

The Zombic Hunch is accompanied by arguments designed to show that it is *logically possible* (however physically impossible) for there to be a zombie. This logical possibility is declared by Chalmers to have momentous implications for the scientific study of consciousness, but as a candidate for the Philosopher's Dream it has one failing not shared with either Einstein's or Matthew's great ideas: *it prescribes no research program.* Suppose you are convinced that Chalmers is right. Now what? What experiments would you do (or do differently) that you are not already doing? What models would you discard or revise, and what would you replace them with? And why?

Chalmers has recently addressed this very issue in a talk entitled "First-Person Methods in the Science of Consciousness" (*Consciousness Bulletin*, Fall 1999, and on Chalmers' website), but I hunt through that essay in vain for any examples of research that are somehow off limits to, or that transcend, heterophenomenology:

> I take it for granted that there are first-person data. It's a manifest fact about our minds that there is something it is like to be us - that we have subjective experiences - and that these subjective experiences are quite different at different times. Our direct knowledge of subjective experiences stems from our first-person access to them. And *subjective experiences are* ***arguably*** *the central data that we want a science of consciousness to explain.* [emphases added] I also take it that the first-person data can't be expressed wholly in terms of third-person data about brain processes and the like. There may be a deep connection between the two - a correlation or even an identity - but if there is, the connection will emerge through a lot of investigation, and *can't be stipulated at the beginning of the day* [emphasis added]. That's to say, no purely third-person description of brain processes and behavior will express precisely the data we want to explain, though they may play a central role in the explanation. So as data, the first-person data are irreducible to third-person data.

Notice how this passage blurs the distinctions of heterophenomenology. "Arguably?" I have argued, to the contrary, that *subjects' beliefs* about their subjective experiences are the central data. I've reviewed these arguments here today. So, is Chalmers rejecting my arguments? If so, what is wrong with them? I agree with him that a correlation or identity—or indeed, the veracity of a subject's beliefs —"can't be stipulated at the beginning of the day." That is the neutrality of heterophenomenology. It is Chalmers who is holding out for an opening stipulation in his insistence that the Zombic Hunch be granted privileged status. As he says, he "takes it for granted that there are first-person data." I don't. Not in Chalmers' charged sense of that term. I don't stipulate at the beginning of the day that our subjective *beliefs* about our first-person experiences are "phenomenological" beliefs in a sense that requires them somehow to depend on (but not causally depend on) experiences that zombies don't have! I just stipulate that the contents of those beliefs exhaustively constitute each person's (or zombie's) subjectivity.

In his paper on first-person methods, Chalmers sees some of the problems confronting a science of consciousness:

> When it comes to first-person methodologies, there are well-known obstacles: the lack of incorrigible access to our experience; the idea that introspecting an experience changes the experience; the impossibility of accessing all of our experience at once, and the consequent possibility of "grand illusions"; and more. I don't have much that's new to say about these.

> I think that could end up posing principled limitations, but none provide in-principle barriers to at least initial development of methods for investigating the first-person data in clear cases.

Right. Heterophenomenology has already made the obligatory moves, so he doesn't need to have anything new to say about these. I don't see anything in this beyond heterophenomenology. Do you? Chalmers goes on:

> When it comes to first-person formalisms, there may be even greater obstacles: can the content of experience be wholly captured in language, or in any other formalism, at all? Many have argued that at least some experiences are "ineffable". And if one has not had a given experience, can any description be meaningful to one? Here again, I think at least some progress ought to be possible. We ought at least to be able to develop formalisms for capturing the structure of experience: similarities and differences between experiences of related sorts, for examples, and the detailed structure of something like a visual field.

What a good idea: we can let subjects speak for themselves, in the first-person, and then we can take what they say seriously and try to systematize it, to capture the structure of their experience! And we could call it heterophenomenology.

If Chalmers speaks of anything in this paper (remember, it is entitled "*First-person* Methods in the Science of Consciousness") that is actually distinct from 3rd-person heterophenomenology, I don't see what it is. Both there and in his contribution to our debate he mentioned various ongoing research topics that strike him as playing an important role in his anticipated 1st-person science of consciousness—work on blindsight and masking and inattentional blindness, for instance—but all this has long ago been fit snugly into 3rd-person science.

In the debate, Chalmers asserted that a heterophenomenological methodology would not be able to motivate questions about what was going on in consciousness in these phenomena. That is utterly false, of course; these very phenomena were, after all, parade cases for heterophenomenology in *Consciousness Explained.* It is important to remember that the burden of heterophenomenology is to explain, in the end, every pattern discoverable in the heterophenomenological worlds of subjects; it is precisely these patterns that make these phenomena striking, so heterophenomenology is clearly the best methodology for investigating these phenomena and testing theories of them.

I find it ironic that while Chalmers has made something of a mission of trying to convince scientists that they must abandon 3rd-person science for 1st-person science, when asked to recommend some avenues to explore, he falls back on the very work that I showcased in my account of how to study human consciousness empirically from the 3rd-person point of view. Moreover, it is telling that none of the work on consciousness that he has mentioned favorably addresses his so-called Hard Problem in any fashion; it is all concerned, quite appropriately, with what he insists on calling the easy problems. First-person science of consciousness is a discipline with no methods, no data, no results, no future, no promise. It will remain a fantasy.

Appendix: Goldman on Heterophenomenology

Alvin Goldman, responding to the paragraph above about Goldman 1997 (see page 5), entered into an email debate with me, lightly edited by me to avoid repetition and remove material not germane to the topics:

Goldman: First, a brief substantive reply to your points [see above, p5]. When cognitive scientists rely on subjects' reports about visual illusions, I take them to be relying on the veracity of the Ss' judgments (beliefs) about how the stimuli *look* (etc.). That is, after all, what the Ss presumably say, or can be interpreted as saying: "It looks as if such-and-such". And the cognitive scientist takes that to be true, i.e., that it does *look* that way to the S (roughly at the time of report). Similarly, the cognitive scientist obviously does not conclude that Ss who report a *deja vu* experience really did have the same type of experience in his/her past. That could not be ascertained by the subject by introspection, which is restricted to present events. So even if the S's deja vu report implies that he/she believes that a certain event or experience occurred in the past (I am not sure it does imply this), the cognitive scientist does not rely on the accuracy of this belief. However, the cognitive scientist (also) takes the S to report, and to believe, that he/she is currently having a "seems-like-this-happened-to-me-in-the past" experience. And the cognitive scientist *does* trust the S's report of *that*. In other words, the scientist concludes that the S does have (roughly at the time of report) an experience of the type "seems-like-this- happened- to-me-in-the-past".

In the context of the treatment of illusions, I do have to talk more about "looks" or "seems". As your discussion below indicates (and you have frequently said in print), you take "seems" only to express something about a S's *belief*. There is no further fact about S (beyond a belief fact) that is expressed by "It seems to S to be F". I, on the contrary, think that a seeming-state is not merely a belief, but a visual state, an auditory state, or other "perceptual-phenomenal" state. You think (see your discussion [above, p. 5]) that such an alleged state would have to involve "images" of certain sizes in the brain. But that is a totally unwarranted interpretation. Undergoing a perceptual-seeming episode need involve nothing like "sense-data" of the sort you conjure up. Cognitive scientists do not have to commit themselves to anything like that when they say that a S really is undergoing a certain type of perceptual-seeming episode (when the S reports that he is).

DENNETT REPLY interjected: EXACTLY! They don't have to commit themselves to anything like that. They can remain neutral. My example of mental images in the brain was just a for instance. My point was that to go beyond heterophenomenological agnosticism, they'd have to suppose something was implied by their S's judgments (beyond the bare fact that these were their judgments, which is what heterophenomenology happily allows). Now it MAY be that your point about "perceptual-phenomenal" states that go beyond "mere" belief—states will someday be supported somehow. But in the meantime, cognitive science proceeds along merrily, leaving itself strictly neutral about that. And in at least some instances (for instance, sudden hunches of déja vu) the claim that there is anything "perceptual-phenomenal" about the presentiment over and above the inclination so to judge seems particularly dubious. (Ask yourself what deja vu would be like if it didn't have any so-called "phenomenal" stuffing. Isn't that in fact what it is like?) But in any case, cognitive science can and should (and does!) remain strictly neutral about such questions of phenomenality until the case is clearly made. My point for years is that it never has been made, so it counts, so far, as just a set of tempting hunches (versions of the Zombic Hunch) that cognitive science should also be agnostic about. And I know of no research in cognitive science that has violated that neutrality except by accident.

You say that my view is that "There is no further fact about S (beyond a belief fact) that is expressed by "It seems to S to be F"." Not quite. I have challenged people to show any way in which there is such a further fact. My view is that it has not been shown that there is any such further fact (beyond the obvious other "behavioral" facts that accompany such belief facts, typically) and in the meanwhile cognitive science can proceed quite happily in strict neutrality about this. In fact, it had better be neutral about this from the outset, so that it can actually have a standpoint from which it might confirm (or disconfirm) your belief.

GOLDMAN, continued: So what is going on when people have a perceptual-seeming episode (whether during actual perception or during imagery)? You point out, in connection with the Shepard, Kosslyn, and Pylyshyn debate, that cogscientists would never rely on Ss' reports to try to settle that. I reply: That is certainly true! But I would never claim, and have never claimed, that scientists rely on all aspects or all details of what their Ss might say. This is explicitly addressed in my "Science, Publicity, and Consc" (SPC) paper on p. 544, the last page of the article. "Everyone nowadays agrees that introspection is an unreliable method for answering questions about the micro-structure of cognition. For example, nobody expects subjects to report reliably whether their thinking involves the manipulation of sentences in a language of thought. But this leaves many other types of questions about which introspection could be reliable". This point is made again in my JCS paper, "Can Science Know When You're Conscious?" (*Journal of Consciousness Studies* 2000) Here is what I say on p. 4 of that article: "Cognitive psychologists and neuropsychologists would not rely, after all, on their subjects' reports about all psychological states or processes. When it comes to the nonconscious sphere of mental processing--the great bulk of what transpires in the mind-brain--scientists would not dream of asking subjects for their opinions. Moreover, if subjects were to offer their views about what happens (at the micro-level) when they parse a sentence or retrieve an episode from memory or reach for a cup, scientists would give no special credence to these views."

So I fully acknowledge that for a wide range of questions, scientists do not allow their Ss' introspections to settle anything. (Of course, usually the Ss have nothing to offer about what happens at the micro-level.) But for another large range of questions, I claim, they *do* trust their Ss' introspections. (A more precise specification of which questions are which I have not yet tried to give. Nor do I know of anybody who has tried to be precise on this matter.)

DENNETT REPLY interjected: Try me. I have. I have pointed out that they trust their S's introspective reports to be fine accounts of how it seems to them--with regard to every phenomenon in all modalities. And that this exhausts the utility of their S's protocols, which they can then investigate by devising experiments that probe the underlying mechanisms. They "trust" their Ss only after they've discovered, independently, that their statements, interpreted as assertions about objective, 3rd-person—accessible processes going on in their brains, are reliable. In other words, they only "rely on" S's statements when they have confirmed that they can be usefully interpreted as ordinary reliable reports of objective properties.

Ask yourself how things would stand if Pylyshyn's most extreme line of mental imagery had turned out to be true (more than the barest logical possibility, I'm sure you would agree—he was not insane or incoherent to put forward his criticisms). In that case, I submit, everyone would agree that the agnosticism of heterophenomenology had paid off big time; people turn out to be deeply wrong about what they are doing. They think they are manipulating mental images with such and such features when in fact all that is happening in them is X. The fact that it sure seems to them that they are manipulating mental images would then have to be explained by showing how they are caused to have these heartfelt convictions in spite of their now demonstrated falsehood. Now if that was never a possible outcome of the research, what on earth could Pylyshyn have thought he was doing? For that matter, what could Kosslyn have thought he was doing?

GOLDMAN continued: In any case, the main point is that I of course agree that not everything a subject might say, in an introspective spirit, would be regarded as scientific gospel. So some of the things you say about conflicts between scientific practice and my reconstruction of it don't work.

DENNETT REPLY: I didn't say you did claim that they held that everything is regarded as scientific gospel. I said that you claimed that cognitive scientists aren't systematically agnostic. But they are, systematically, so systematically that they don't even both mentioning it, in all the cases I cite in this passage where I discuss your claim.

The proper way to criticize my view is to develop an independent case for "real seeming." A number of people have tried. Nobody has yet succeeded. See, e.g., the essays in the Phil Topics issue of 1994, and my response, "Get Real". But beyond establishing this as a philosophical point, there is the obligation to show that cognitive science has been (or should be) honoring it. When you can show experiments that get misinterpreted, or can't be analyzed, or would never be dreamt up, by people committed to heterophenomenology, then you can claim that I am mistaken in claiming that heterophenomenology both is, and should be, agnostic.

GOLDMAN, next response: I agree that *one* of the key issues is whether there is anything more to visual seeming (e.g.) than belief. At the risk of repeating what others have said (possibly ad nauseum, from your point of you), this just seems like the obvious, straightforward interpretation of what goes on in, e.g., the blindsight patient. The patient doesn't tell his physician that he doesn't *believe* that there are any objects of such-and-such type in the vicinity (in the area of his scotoma). He says that he doesn't see anything in that vicinity [*expressing,* not *reporting* his *belief* that he doesn't see anything in that vicinity; see CE, pp. 305–6–DCD]. We might even arrange for there to be a case where he does have beliefs about the target properties—as a result of somebody else *telling* him about such properties. But he'll still say that he doesn't *see* anything there. And the standard, default, entry-level reaction of the cognitive scientist is to trust that report, to conclude that S really doesn't see anything there. Of course, the scientist might be a little more cautious, since, among other things, the S might be confabulating, or neglecting. But the reason blindsight is an interesting and challenging phenomenon, a phenomenon related to *vision*, is because it's an absence of *seeing*. How do we know about this absence? From the S. From the subjects' reports. So we are basing our conclusions on a trust of the subjects' reports.

DENNETT REPLY interjected: Not so. Anticipating this sort of response in my own discussion of blindsight in CE, I pointed out the problem of trust. See p. 326, where I show why "the phenomena of blindsight appear only when we treat subjects from the standpoint of heterophenomenology" and particularly point to how the phenomenon would evaporate if we concluded that subjects were malingering, or suffering from hysterical blindness. Heterophenomenology is tailor made for dealing with blindsight.

Again, in the deja vu case, it doesn't capture the phenomenon well to describe it as a *belief* that one experienced a similar thing in the past. Rather, it's a phenomenon in which it *feels* like one experienced such a thing in the past; or one has a seeming memory of such a thing. One might not *believe* that it happened at all, but one still *feels* as if it did. Again it's a reliance on the S's report of this phenomenon that makes the observer think that the S has really undergone this phenomenon at the time of report.

DENNETT REPLY interjected: To "feel as if it did"' is to be strongly tempted to judge that it did. Of course the temptation can be overridden once one is no longer naive. And what is the feeling of temptation? Just noticing that one is so tempted to judge!

GOLDMAN next reply: I realize that a "doxological" (or representational) reductionist like yourself will want to reduce feeling states to dispositions-to-believe. A resistor like myself need not deny, of course, that feeling states do have a tendency to produce beliefs. The question is whether there are "categorical" features of feeling states in virtue of which they have that tendency, or whether they are just pure doxological tendency and nothing else. I find the former view more compelling, and don't think that representational reductionism will work across the board. But this is another big issue (admittedly one that is intimately tied to the issue at hand).

DENNETT REPLY: Fine. And isn't it nice that heterophenomenology can proceed with all of its research agenda without our having to settle anything about this "big issue" first! If you're right, the "categorical" features will eventually be confirmed to be important by some as yet unimagined test. (Or if, as I gather your colleague David Chalmers holds, no empirical or "behavioral" test could shed any light on this important but elusive sort of feature, I guess it will have to be some philosophical argument alone that settles the issue. Seems unlikely in the extreme to me.) In the meantime, a 3rd-person science of consciousness can proceed apace. That's what is so good about its neutrality.

GOLDMAN: One last question about "neutrality". In your discussion of blindsight, do you agree that scientists give prima facie credence to a subject who claims to have no sight in a certain area? You stress that they do not uncritically trust these subjects. They want to check to see if there is neurological damage, and they want to rule out the possibility of "hysterical blindness". But don't they give some prima facie credence to the subject's report? Or do you deny this? If you agree that they do this, the question arises as to whether this is "neutrality", or agnosticism. I think not. Most epistemologists would agree that all of our sources of belief or justification are subject to correction from other sources. We don't trust vision uncritically, or memory, etc. But to say this is not to say that we are "agnostic" toward vision or memory. By giving prima facie credibility to each of these sources, we are doing the most that we ever do to any one source (or any one deliverance of a particular source). I would argue that the same holds here. Although the scientist does not uncritically trust a S's introspection (and there's an additional factor here—the S's report might not stem from introspection at all), he does give it prima facie trust. And that is very far from agnosticism. So if heterophenomenology ascribes true agnosticism to scientists, as you claim it does, then it doesn't get matters right.

DENNETT REPLY: As I try to make clear in CE, in the section entitled "The Discreet Charm of the Anthropologist," (pp. 82–3, on "Feenoman") heterophenomenology is NOT the NORMAL interpersonal relationship with which we treat others' beliefs—with its presumption of truth (marked by the willingness of the interlocutor to argue against it, to present any evidence believed to run counter, etc.). That is also true of anthropologists' relationships with their subjects when investigating such things as their religion. Actually, it extends quite far—when the native informants are telling the anthropologists about, say, their knowledge of the healing powers of the local plants, the anthropologists' first concern is to get the lore, true or false—something to be investigated further later. Ditto for heterophenomenology: get the lore, as neutrally and sympathetically as possible. That is a kind of agnosticism, differing in the ways I detail on pp. 82–3 from the normal interpersonal stance, but it is the normal researcher/subject relationship when studying consciousness with the help of S's protocols. If it doesn't fit your (or a dictionary's, or the majority of epistemologists') definition of agnosticism perfectly, I have at least made clear just what kind of agnosticism it is, and why it is the way it is.

As for blindsight, do the researchers give some prima facie credence to the reports? Of course—otherwise they wouldn't even consider investigating them. As I say, their attitude is to take what subjects say as seriously as possible—a policy that is entirely consistent with a kind of agnosticism, of course. The old introspectionism failed precisely because it attempted, unwisely, to give subjects more authority than they can handle; as the years rolled on, more cautious and savvy researchers developed the methodology I have dubbed heterophenomenology. They crafted a maximally objective, controlled way to turn verbal reports (and interpreted button-pushes, etc., etc.) into legitimate data for science. All I have done is to get persnickety about the rationale of this entirely uncontroversial and ubiquitous methodology, and point out how and why it is what it is—and then I've given it an unwieldy name. So when, in my forthcoming Cognition essay, in the special issue on the cognitive neuroscience of consciousness, I point out that the hundreds of experiments discussed in the various pieces in that issue all conform to heterophenomenology, the editors and referees nod in agreement. Of course. It's just science, after all. And it does study consciousness. Obviously—unless you believe that the "easy" problems of consciousness are not about consciousness at all.

Now I have challenged David Chalmers to name a single experiment (in good repute) which in any way violates or transcends the heterophenomenological method. So far, he has not responded to my challenge. My challenge to you is somewhat different: to show that I misdescribe the standard methodology of cognitive science when I say it adopts the neutrality of heterophenomenology.

References

G.H. Bower, J.P. Clapper, Experimental methods in cognitive science, in *Foundations of Cognitive Science*, ed. by M. Posner (MIT Press, 1989)

Chalmers, *The Conscious Mind* (1996)

Dennett, 'Quining Qualia' in *Consciousness in Contemporary Science*, ed. by A.J. Marcel, E.E. Bisiach (CUP, 1988)

Dennett, Intrinsic changes in experience: swift and enormous commentary on palmer. *BBS* **22**(6), December 1999

Dennett, The zombic hunch: extinction of an intuition?, in *Philosophy at the New Millenium*, ed. by A. O'Hear, vol. 48 (Cambridge Univ. Press, Royal Institute of Philosophy Supplement, 2001), pp. 27–43

A. Goldman, Philos. Sci. **64**, 525–545 (1997)

A. Goldman, Can science know when you're conscious? J. Conscious. Stud. (2000)

K.M. O'Craven, B.R. Rosen, K.K. Kwong, A. Treisman, R.L. Savoy, Voluntary attention modulates fMRI activity in human MT/MST. Neuron, **18**, XXX (1997)

S. Palmer, Behav. Brain Sci. **22**(6), December 1999

Chapter 27
Rethinking Life

Eörs Szathmáry

Introduction

There are two foundational issues concerning the understanding of life: the investigation of the nature of living organisms and the elucidation of the principles of evolution.

Considerable advance has been made in the understanding of the nature of organisms and their evolution in the last fifty years. Experiments and theories of the origins of life, developmental and evolutionary genetics have delivered great contributions. Yet it seems that our knowledge of the principles of life is still rather incomplete. We should minimally address the following key open issues:

- Organisms are not in being; they are in happening. How does autocatalytic closure maintain itself in the sea of potential side reactions? How fuzzy is thus the boundary of a living chemical system? What is the role of the side-reaction halo around core metabolism?
- Can we think of the subsystems of a living system as Lego-pieces, or is this picture too mechanistic and naïve? Are the subsystems of a cell coupled in a fashion so that any doublet remains viable, or are they linked so that when one deletes one subsystem, the others also fall apart? Or both views may be relevant, but in different phases of evolution?
- Would evolution come to a halt without planetary and astronomical forcing, or would it continue, without abiotic changes, like the Red Queen? If the latter,

E. Szathmáry (✉)
Evolutionary Systems Research Group, MTA Ecological Research Centre,
Klebelsberg Kuno utca 3, Tihany 8237, Hungary
e-mail: szathmary.eors@gmail.com

E. Szathmáry
Parmenides Center for the Conceptual Foundations of Science,
Kirchplatz 1, 82049 Pullach, Germany

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_27

> would this be open-ended? Life seems to be self-modifying, but where does this stop? If life creates and re-creates its own state space, can we ever be predictive of long-term evolution? How algorithmic is evolution in toto?

I shall give a brief survey of some attempted answers to these exciting questions.

Life Itself

The triumph of molecular biology on the empirical side and the emergence of systems chemistry and the associated theoretical apparatus yielded unprecedented progress. Molecular biology has revealed the basics of the overlap between chemistry and informational operations. Terms such as copying, proofreading, editing, transcription and translation are informational in nature, and rightly so. For Crick information meant the precise determination of sequence, either in nucleic acids or proteins (Crick 1958). Even accepting this definition one should rather say "precise enough"—all such operations in the molecular life of the cell have a finite precision: transcription and DNA replication have accuracy of 10^{-4} and 10^{-10} per digit per operation. This huge difference makes sense: accuracy has its costs in terms of time and energy, and whereas messenger RNA molecules are disposable, DNA for transgenerational inheritance is not.

The elucidation of the genetic code, whereby information stored in nucleic acids can be translated to the amino acids sequences of proteins, is the jewel in the crown of molecular biology. Kurt Gödel, facing the results about the genetic code, is remembered to have said: "vitalism is dead" (Brenner, Sydney, personal communication). I agree with this overall assessment, but not without qualification. The main problem lies in the recognition that living beings with a genetic code are the results of a perhaps of hundreds of millions of years of evolution. In fact, we still do not have universally accepted account for the evolutionary emergence of the genetic code. Many biologists now share the view that life is older than the genetic code. If so, than some more general (and likely deeper) characterisation of life is necessary.

We believe that autocatalysis has played a central role in bridging the domains of chemistry and biology. The simplest expression of autocatalysis has the form:

$$\mathrm{A} + \mathrm{X} \rightarrow 2\mathrm{A} + \mathrm{Y},$$

where A is the autocatalytic agent, X is a set of raw materials, and Y refers to waste. The energetic drive comes from the energy difference between X and Y. The term autocatalysis make sense: one copy of A helps the formation of another copy of A. Replication from the chemical point rests on autocatalysis, and autocatalysis always results in some form of replication (Orgel 1992). For evolution by natural selection to occur, one should have different forms of replicators that are able to propagate their own kind, formally expressed as:

$$A_i + X_i \rightarrow 2A_i + Y_i,$$

where A_i is an autocatalyst of type i. Informational replication means that differences among replicators are heritable. But, as mentioned above, heredity is not exact: occasionally, A_i produces A_j, ($i \neq j$) rather than itself.

There are many, qualitatively autocatalytic chemical systems, beyond nucleic acid replication, embedded in current biological organization. Perhaps the lesser known examples come from metabolism. In some cases identification is easy: the reverse citric acid cycle fixes carbon dioxide in some bacteria (Fig. 27.1). Starting with one molecule of, say, malate, after one turn of the cycle we have two molecules of malate. A less obvious (because more complex) example is the Calvin cycle, where three molecules of 3-phosphoglycerate produce a fourth one (Fig. 27.2). This is fascinating. Even more fascinating would be if such cycles could work, even if less efficiently, without enzymatic aid. Why? Enzymes are proteins, and they act as biocatalysts for most of the reactions of metabolism, including those of the two mentioned cycles. But again, the production of enzymes now rest on the genetic code, and living beings could not have had it in the beginning.

Cooling down our expectations we find that there are very few chemical autocatalytic networks of small molecules that do not require enzymes. A famous case is the formose reaction, an autocatalytic production of sugar molecules at the expense formaldehyde consumed (Fig. 27.3). Unfortunately, in a batch reaction this system converts into tar. This unwelcome fate is shared by many other examples of

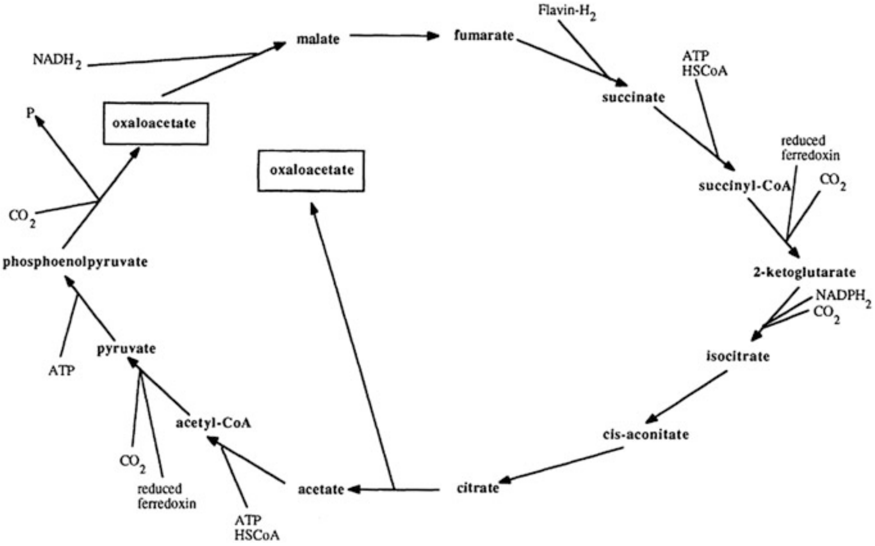


Fig. 27.1 The reductive citric acid cycle. Each step is catalyzed by an enzyme in contemporary organisms. The carboxylation of phosphoenolpyruvate has so far proven impossible to implement under prebiotic conditions (from Szathmáry 1995)

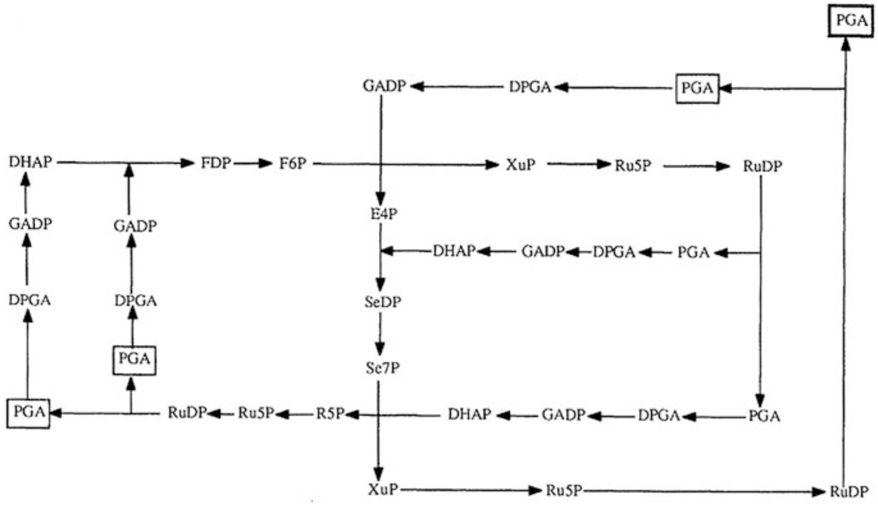


Fig. 27.2 The Calvin cycle is really an autocatalytic network of carbon fixation of plants. External reactants are not shown. PGA is 3-phosphoglycerate. Also here, each step is catalyzed by an appropriate enzyme (from Szathmáry 1995)

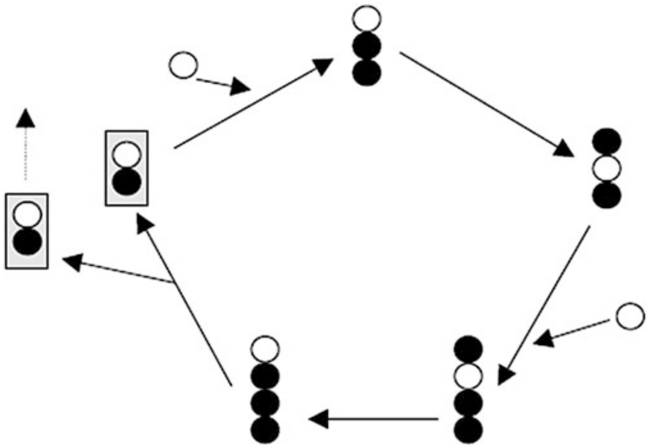


Fig. 27.3 The autocatalytic core of the formose cycle. The single open circle depicts formaldehyde; the marked circle doublet is glycolaldehyde. Larger symbols stand for sugar molecules. This is a non-enzymatic reaction network, with several alternative reaction routes (not shown). The formation of sugars was discovered by Butlerov in 1861 (from Szathmáry 2006)

potential prebiotic relevance. We should add that this does not happen when the system is run under steady state conditions when the access production is washed out from a flow reactor (Decker 1972); we shall return to the significance of this statement.

Irrespective of the prebiotic feasibility of an autocatalytic intermediary metabolism, one should ask: what are the organizing principles of simple life that rest on chemistry?

To my mind the best theoretical understanding of the foundations of individual (organismal) living systems resides in the chemoton theory of Gánti (1971, 1975, 1978, 2003). Biological populations typically consist of individuals, and our notion of “life” refers to both. A mule cannot reproduce, but nobody has ever said that it would thus be not alive. Gánti has offered a phenomenological characterisation of living systems as individuals in terms of his “life criteria” that come in two forms: absolute and potential (although “potentiating” might have been better for the latter). Absolute criteria (such as metabolism) must be satisfied by any living system at any time, whereas potential life criteria (such as reproduction) are necessary only to create a living world. If (as likely) molecular replicators preceded living systems, a living world in this sense may preceded living systems (Griesemer 2003; Griesmer and Szathmáry 2009)!

I digress at this point into a discussion of whether viruses are alive or not, and whether in fact any life criteria could literally be “absolute”. First, definitions are always arbitrary, and if their formulation is internally consistent, one may accept any of them: theories are falsifiable, definitions are not. But there often is a pragmatic difference between alternative definitions: their “fitness” in cultural evolution can be markedly different. One important consideration is whether adopting a particular definition helps us to get to more productive associations and research programs than another. In this sense I agree with Gánti that viewing viruses as live is not very productive. Using an information technological analogy we can say that a virus is to the living cell as a particular programme is to a computer. Such a programme is a set of instructions making the computer copy the former in arbitrary numbers, even at the price of ruining the latter. Computers are functional without such malign programmes, but without the former the latter are completely inactive. Viruses replicate and evolve, but are not alive. A population of mules cannot evolve, although each mule is alive. In this spirit I formulated the concept of units of evolution and units of life (Szathmáry 2002, 2003, 2006): their domains overlap, but not fully (Fig. 27.4).

As usual, science rests in part on abstraction and idealisation. We only know one life for the time being: that which has evolved by natural means here on earth. This poses a special problem that is best explained by a didactic example. Suppose that all living systems would be light blue (they are not, but all of them have nucleic acids, right?). How would you know that this particular trait is a necessary or a contingent feature of life? There are two partial answers that can be given. First, we might one day find elsewhere, or synthesize, systems that are like all common living systems except that they are not light blue. In this case it will make a lot of sense to raise the concept of “being alive” to a higher level of abstraction by dropping “light blueness” as a critical feature. Second, we could reason that there must have been simpler, but already rather complex, systems without this colour.

Now we come back to the issue of the chemical organization of minimal life. The term “minimal” highlights the insight that one cannot have the same

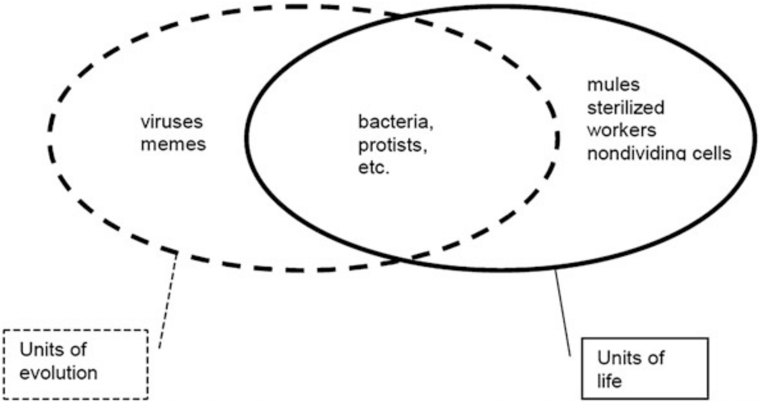


Fig. 27.4 Units of evolution and units of life (from Szathmáry 2002)

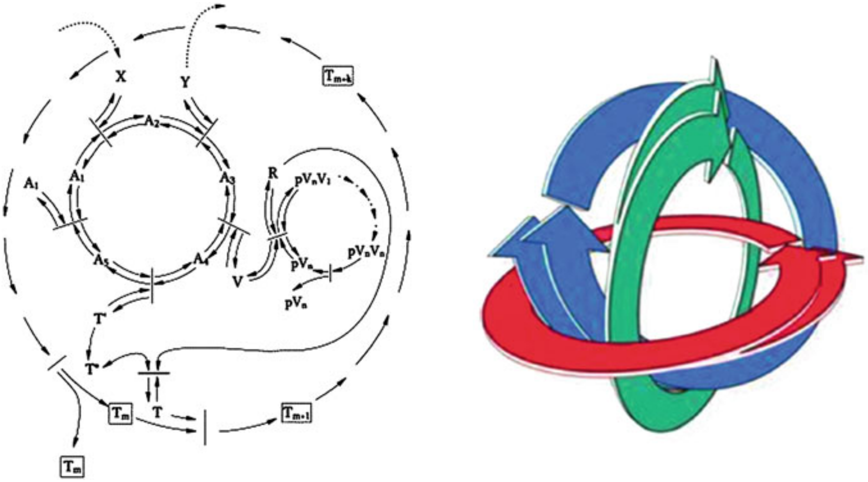


Fig. 27.5 The chemical couplings of the chemoton model (left) and its abstract symbol (right). A_i are the intermediates of the metabolic engine, V is a monomer of the pV_n template polymer, and T_m is a bilayer membrane consisting of *m* molecules of T, X is food and Y is waste

organizational model of cellular and multicellular living beings, since the latter consist of units that are already alive. The chemoton as a minimal model consists of three abstract autocatalytic, qualitatively different chemical systems: a metabolic engine, an informational replicator and a boundary (Fig. 27.5). The system as a whole is also autocatalytic, but (within certain quantitative constraints on its parameters) additionally it can divide in physical space also. This model is also regarded now as a conceptual foundation of the recently emerged field of systems chemistry (von Kiedrowski et al. 2010). A useful definition of the latter field is that it deals with conceptualisation, analysis, synthesis and coupling of different autocatalytic chemical systems.

The composition of the chemoton prompts one to wonder about partial combinatorics of its constituent systems. This is a legitimate question indeed: one can conceptualize and realize systems doublets (Fig. 27.6) instead of the full trinity of the chemoton (Szathmáry et al. 2005). Gánti himself considered the "self-reproducing microsphere" as a doublet of metabolism and membrane growth (Gánti 1978). Others have considered realizing template replicators within membranes (Szostak et al. 2001). Arguably, spontaneous or artificial genesis of doublets seems simpler than that of the chemoton. But, one way or the other, realization of chemoton-like organizations seems to be necessary to paint the whole canvas.

Note the abstract and idealized nature of the chemoton. It is abstract since none of its subsystems is identified at start with any concrete class of compounds. It is idealized because in all likelihood no chemical cycle comprising just five elementary reaction steps could produce the compounds for its own autocatalysis and that of the other subsystems at the expense of the difference between one molecule of food (X) and one molecule of waste (Y). Yet, the logic of the organization is much clearer revealed by such idealized constructions that avoid being bogged down in hundreds of chemical reactions. It is a valid and exciting question to ask: *given* the basic model in Fig. 27.5, what concrete chemical systems could satisfy the constraints of such an organization? Sadly, chemistry is no at the stage to answer this question. This limits our understanding of the hypothetical domain of exobiology (Benner et al. 2004). We only know "one experiment": life on earth. It would thus be invaluable to find *independently evolved* cases of life.

There is a fundamental question to which inspection of the chemoton model is likely to lead us. As it is portrayed in Fig. 27.5, the impression one gets is that the subsystems are assembled as Lego pieces: the chemoton symbol also suggests that the one can take away any one subsystem so that the other two remain coupled and are, presumably, functional. It is at this point where the original idealisation of the chemoton breaks down. The chemistry of any living being is imagined to operate in a tiny domain of chemical space. This chemistry must have autocatalytic closure, which ensures that the organization is maintained by continual recreation, and that it is also reproducible in the biological sense. The recognition of the importance of autocatalysis is important, since it paves the way to the recognition of self-reference as a key aspect of living systems.

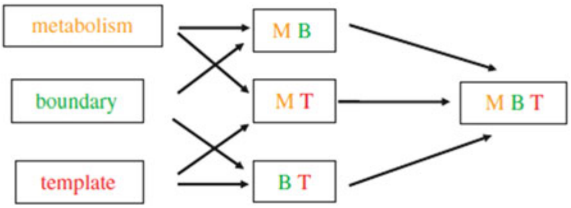


Fig. 27.6 Combination of different autocatalytic systems into system doublets (infrabiological systems) and a triplet (corresponding to the chemoton). From Szathmáry et al. (2005)

Whenever one defines a chemical system as a list of molecules and some of their interactions, one always neglects many molecules and further interactions. Some of the neglected reactions will happen time and time anew. Living systems must and are able to maintain themselves in this sea of side reactions. But depending on the environment, some of the emerging, neglected molecules can have all of sudden a non-negligible (good or bad) effect on the system. These side reactions can thus re-define the system in real time, and can have an important contribution to further evolution. Thus the chemical organization of a living system is not black and white: what we regard as the black hard core is surrounded by fluctuating onion-layers with different shades of grey. Nobody can exhaustively prestate what part of which onion layer will or will not become relevant in the next time instance. The old saying that "living systems are not in being, they are in happening" thus acquires a deeper meaning. This view suggests that there might not be sufficient, entailing laws of the chemistry of biological organization. Full appreciation of the consequences of this updated view for the origin of life is lacking.

It is exactly because of the side reactions that the coupling topology conceived by Gánti might mislead us to some extent. It is perfectly possible that one will never be able to sustain an autocatalytic metabolism without the boundary system. Boundaries have two important roles: to keep the inner components of the system in, and keep harmful reactants out. Only experiments will tell us whether without the boundaries either metabolism or template replication are sustainable (note that sustainability is more than running a reaction for, say, a few minutes). But Decker's experiments with the formose system (Decker 1972) suggest that growing compartments might be critical to prohibit non-enzymatic metabolism from producing (too much) tar. Ultimately, in the terminology of Fig. 27.6, we might find systems {M, B}, {T, B} and {M, B, T} are all feasible self-sustaining combinations but {M}, {T} and {M, T} are not.

This last issue is also related to the question how chemical networks can increase in complexity. As we have seen above, co-opting low-propensity reactions into the organization is one way. There is a more spectacular option: chemical symbiosis (King 1977). Imagine two autocatalytic networks A and B that emerged independently, i.e. in different environments (these can be different vesicle populations). What happens when they meet, e.g. by vesicle fusion? They might annihilate each other by forming tar, for example. The more promising outcome is that they become integrated, which can happen in two possible ways: (i) A and B together form a set of new reactions, out which some act as "glue" to couple the two systems, or (ii) some of the original reactions become unimportant and the remaining parts unite. Note that both changes are heritable without the action of genetic replicators in the conventional sense.

There are two considerations supporting the future applicability, also in experiments, of this view. One is that we might give up the hope of a "one-pot" approach to chemical origin of life (Sutherland 2017). Instead, different chemical systems may have originated in different environments, relegating their further evolution to "hybrid zones". Another fact is that coenzymes (catalytic small molecules that at together with enzymes in contemporary metabolism) seem to be auto- and cross-catalytic for their production (King 1980). This might implicate that (i) metabolism may have indeed started without enzymes, (ii) coenzymes may indeed be relics of ancient metabolism (White 1976), (iii) metabolism is autocatalytic, and finally, (iv) metabolism may have grown more complex by chemical symbiosis. All these ideas are in principle experimentally testable.

One more evidence in favour of the autocatalytic nature of metabolism is in order. One could imagine metabolism as a complex feed-forward network producing complex molecules from simple ones without autocatalytic organization. This would entail that enzymes and food molecules would be enough to kick-start even contemporary metabolism. But even in the case where autocatalytic parts of the grand network are identifiable, they may not be indispensable to launch metabolism. The exciting case is the one where the presence of an autocatalytic seed is obligatory (Fig. 27.7). Inspection of the metabolism of contemporary metabolic networks has revealed (Kun et al. 2008) that Gánti was right: metabolism is always autocatalytic in the latter, strict sense.

The space of possible molecules is indefinitely large (for all practical purposes one could say it is virtually infinite). Thus, even in a laboratory as big as the Earth one cannot physically realize all possible molecules and their reactions: during the unfolding of the chemosphere some domains will not be visited by pure chance.

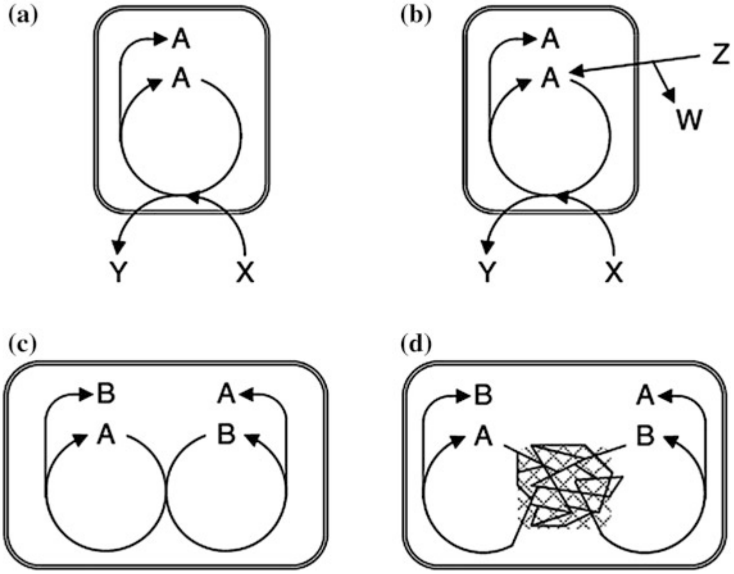


Fig. 27.7 Metabolism with facultative and obligatory autocatalysis. **a** A protocell showing an indispensable autocatalytic core A. **b** A richer medium is able to launch metabolism because Z can be converted to A. **c** The autocatalytic core is composed of A and B, a pair of cross-catalytic molecules. **d** If A and B are embedded in a huge network, it may be difficult to identify the autocatalytic compounds. Inner metabolite: A, food: X and Z, waste: Y and W (from Kun et al. 2008)

What is relevant is the adjacent possible (Kauffman 2000). The chemosphere develops by making steps in the available adjacencies. Chemicals that are not present cannot exert any effect on the dynamics of the system. But once a new molecule appears for the first time in the history of the chemosphere, new interactions and further adjacencies emerge. Thus already the chemosphere is a self-modifying system (Kampis 1991): a feature that biology has inherited rather than discovered.

Open-Ended Evolution of Life

Regarding evolution, one inherits all the limitations of chemistry, but more limitations are surely relevant. Biological heredity today rests on nucleic acids, with virtually unlimited hereditary potential (Maynard Smith and Szathmáry 1995). This feature is necessary but not sufficient condition for evolutionary open-endedness. Suppose that a vast genotype space is mapped to a severely limited set of phenotypes. Were this the case, genotypes could revisit previous states by drift without phenotypic change, exhausting the phenotype space in limited time—hence no progress of any kind, no open-endedness. The basic question is hence what grand view of evolution we are formulating (de Vladar et al. 2017).

"*Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.*"—these are the famous last words of *The Origin of Species* (Darwin 1959). Many would feel uncertain these days whether the "production of higher animals" would indeed (even indirectly) follow. Linked to this, many share the feeling that our theories of evolution still do not match the grandeur of evolution itself. Relevant questions revolve around halting (whether evolution, given no externally forced abiotic change, would come to a halt), open-endedness (whether evolution is in some sense unbounded), and progress (whether something is maximized in evolution). Answers to these questions cannot be independent.

We distinguish three forms of open-endedness (de Vladar et al. 2017). Weak open-endedness allows for the occurrence of novel phenotypes indefinitely. Strong open-endedness entails the continual appearance of evolutionary novelties/innovations. Ultimate open-endedness refers to an indefinite increase in complexity. Open-ended innovation rests on exaptations (predaptations) that generate novel niches. Exaptations may result in new traits and new rules as the dynamics unfolds implying that evolution might not be fully algorithmic.

van Valen (1973) proposed decades ago that species go extinct at a constant rate due to antagonistic coevolution under a zero-sum game assumption (an evolutionary gain in fitness in one species is matched by a distributed decrement among the other species). This is the famous Red Queen mechanism, allegedly supported by the fact that consumable resources are finite (one ounce of nitrogen consumed by one species will be unexploited by all the others). However, evolutionary lag (reduced fitness relative to the maximum possible in a stationary environment) can be due to either a decrease as well as an increase of available resources, since both may require genetic adaptation to overcome the lag. An earlier model had two variables: average lag and species number (Stenseth and Maynard Smith 1984). One possible outcome was maximum diversity and zero lag load and the other was intermediate species diversity with a constant non-zero lag load, implying Red Queen coevolution. In the latter mode species are constantly turned over even without external abiotic forcing.

A fine distinction is in order: continual evolution is not necessarily open-ended. Endless cycling across the same states does not introduce novelty in the long run. The model by Nordbotten and Stenseth (2016) belongs to this category because the phenotype space is bounded and population densities cannot blow up, thus in the long run any part of the state space is potentially visited an infinite number of times. Can one say something useful about open-endedness?

Waddington (1969) made a bold proposal relevant for the quest for open-ended evolution, as follows: "*The complete paradigm must therefore include the following items: A genetic system whose items (Qs) are not mere information, but are algorithms or programs which produce phenotypes (Q * s). There must be a mechanism for producing an indefinite variety of new Q'*s, some of which must act in a radical way, which can be described as 'rewriting the program'. There must also be an indefinite number of environments, and this is assured by the fact that the evolving phenotypes are components of environments for their own or other species. Further, some at least of the species in the evolving biosystem must have means of dispersal, passive or active, which will bring them into contact with the new environments (under these circumstances, other species may have the new environments brought to them).*" (p. 39).

To be sure, no Artificial Life platform satisfies Waddington's conditions. A notion of his proposal, referred to as 'strong open-endedness' later (de Vladar et al. 2017), is crucially expressed as "some of which must act in a radical way which can be described as 'rewriting the program'". This raises the problem of evolutionary innovations and novelties, two seemingly similar but fundamentally different notions. Wagner (2015) distinguishes three cases: (i) the evolutionary innovation of a new functional capacity, (ii) type I novelty as the origin of a novel body part, and (iii) type II novelty as a considerable modification of a pre-existing body part. Wagner states the link as "what is special about life and biology as a science is that we are dealing with an open-ended creative process, and understanding this process of innovation is one of the great intellectual challenges of current biology" (p. 76). The origin of novel functions and novel body parts are distinct processes: whereas insect wings evolved as novel body parts, dinosaur wings evolved as modified forelimbs.

Day (2012) established an intriguing link between theories of evolution and computation. Crucial to his construct is the existence of unlimited heredity of digital replicators, where the number of possible types hyperastronomically surpasses that of individuals present (Maynard Smith and Szathmáry 1995). (Note that infinite state space does not guarantee open-endedness if by construction the theory entails a single equilibrium or a stationary distribution.) Day then observed that the state of an evolving biota can be mapped to the natural numbers, and evolution of the whole biota is a function on the natural numbers. The stunning conclusion is that one can, in principle, have a predictive open-ended evolution theory only if evolution is progressive. The former means that one should be able to decide which parts of genotype space will be visited and which will not be. The latter means that there is a mapping onto the natural numbers so that by evolution one always obtains a larger number. There is no necessary or obvious link between this mapping and, for example, a measure of complexity, however. Also, this notion of progress allows for evolutionary convergence in parts of the biota. Note also that even if a predictive theory is in principle possible, there is no guarantee that we shall ever create one.

On this note, it is important to consider in some detail in what sense evolution *in toto* may be non-algorithmic. In a fascinating manifesto, Longo et al. (2012, p. 1383) write: "*…if we consider the proper biological observable (crystalline, kidney), each phenotypic consequence …has an* a priori *indefinite and unorderable, hence algorithmically undefinable set of potential uses, not pre-definable in the language of physics… we cannot predefine nor, a fortiori, mathematize and algorithmically list those uses ahead of time nor what shall come into existence in the evolving biosphere. In other plain words, we cannot write down equations of motion for these unprestatable, co-constituted, newly relevant observables and parameters in evolution.*" Thus, there are no entailing equations of motion at the level of biology, which in this case corresponds to the meta-level. Hence, there cannot be a 'full' algorithm for evolutionary dynamics.

It is justified to ask where this inability to prestate phenotype space comes from. It was boldly suggested that it is rooted, ultimately, in the inexhaustibility of combinatorial chemical matter (molecules and their interactions: Fernando et al. 2011). Thus we claim that a necessary condition for strong open-ended evolvability is unlimited heredity controlling a combinatorial space that is as least as versatile as chemistry, in turn allowing for an indefinitely large set of phenotypes showing innovations and novelties largely mediated by developmental changes (Jablonski 2005). (It remains to be seen whether chemistry must be carbon-based or not: Benner et al. 2004).

This does not contradict the possibility that we might achieve one day an Artificial Life simulation showing interesting open-endedness by including efficient causes rendered from foundational principles (e.g. a rich-enough chemistry and/or underlying physical laws). Although these causes and principles can allow for open-ended evolution, novelties of increasingly complex organizations might, but will not necessarily, be deducible from them. Post hoc identification does not entail prediction.

Just as one cannot prestate all possible uses of a screwdriver, one cannot prestate all possible exaptations, i.e. evolutionary adaptations that may turn out to be useful, even if initially in a modest way, for a new function (Longo et al. 2012). Exaptations can lead to novel adaptations and generate genuinely novel niches for others. Therefore, although evolution has algorithmic components, it may not be algorithmic in toto.

Finally, we enquire about the relationship between open-endedness and complexity. The major evolutionary transitions have entailed, and enabled further spectacular innovations and novelties (Maynard Smith and Szathmáry 1995) concomitant with an increase in complexity by *any* sensible definition. This prompted us (de Vladar et al. 2017) to indentify the notion of 'ultimate open-endedness'; namely, a "process in which there is the possibility for an indefinite increase in complexity" (Banzhaf et al. 2016, p. 69). Standish (2003) makes the bold suggestion that beyond a threshold level complexity might not limit open-ended evolution of forms, provided the attainability of diverse forms increases exponentially with complexity—a relation that may or may not hold: future work needs to tell us.

The converse of the above discussion is effective open-endedness as opposed to theoretical open-endedness (Banzhaf et al. 2016). The first notion relates to what we can observe in a finite, real world. Consider, for example, the generation of complexity by a major transition in individuality. Since the biosphere as a whole cannot reproduce, there is a ceiling to such transitions somewhere below that level. Therefore, by first principles, there is no theoretical open-endedness in this regard, but it is arguable that effectively there *is* such a phenomenon, because even after almost 4 billion years of evolution (paraphrasing Richard Feynman) "there is plenty of room up there" (de Vladar et al. 2017).

References

W. Banzhaf, B. Baumgaertner, G. Beslon, R. Doursat, A. James, J.A. Foster, B. McMullin, V.V. de Melo, T. Miconi, L. Spector, S. Stepney, R. White, Defining and simulating open-ended novelty: requirements, guidelines, and challenges. Theory Biosci. **135**, 131–161 (2016)

S.A. Benner, A. Ricardo, M.A. Carrigan, Is there a common chemical model for life in the universe? Curr. Opin. Chem. Biol. **8**, 672–689 (2004)

F.H.C. Crick, On protein synthesis. Symp. Soc. Exp. Biol. **12**, 138–163 (1958)

C. Darwin, in *The Origin of Species*, 1st edn. (John Murray London, 1959)

T. Day, Computability, Gödel's incompleteness theorem, and an inherent limit on the predictabilty of evolution. J. R. Soc. Interface **9**, 624–639 (2012)

H.P. de Vladar, M. Santos, E. Szathmáry, Grand views of evolution. Trends Ecol. Evol. **32**, 324–334 (2017)

P. Decker, Open systems which can mutate between several steady states ("Bioids") and a possible prebiological role of the autocatalytic condensation of formaldehyde. Z. Naturforsch. **27b**, 257—263 (1972)

C. Fernando, G. Kampis, E. Szathmáry, Evolvability of natural and artificial systems. Procedia Comput. Sci. **7**, 73–76 (2011)

T. Gánti, *The Principle of Life (in Hungarian)* (Gondolat, Budapest, 1971)

T. Gánti, Organization of chemical reactions into dividing and metabolizing units: the chemotons. Biosystems **7**, 15–21 (1975)

T. Gánti, *The Principle of Life (in Hungarian)*, 2nd edn. (Gondolat, Budapest, 1978)

T. Gánti, *The Principles of Life* (Oxford University Press, Oxford, 2003)

J. Griesemer, E. Szathmáry, Gánti's chemoton model and life criteria, in *Protocells Bridging Nonliving and Living Matter*, ed. by S. Rasmussen, M.A. Bedau, L. Chen, D. Deamer, D.C. Krakauer, N.H. Packard, P.F. Stadler (MIT Press, Cambridge, MA, 2009), pp. 481–512

J.R. Griesemer, The philosophical significance of Gánti's work, in T. Gánti, *The Principles of Life.* (Oxford University Press, Oxford, 2003), pp. 167–186

D. Jablonski, Evolutionary innovations in the fossil record: the intersection of ecology, development, and macroevolution. J. Exp. Zool. **304B**, 504–519 (2005)

G. Kampis, *Self-Modifying Systems in Biology and Cognitive Science* (Pergamon Press, 1991)

S. Kauffman, *Investigations* (Oxford University Press, Oxford, 2000)

G.A.M. King, Symbiosis and the origin of life. Orig. Life **8**, 39–53 (1977)

G.A.M. King, Evolution of the coenzymes. Biosystems **13**, 23–45 (1980)

Á. Kun, B. Papp, E. Szathmáry, Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. Genome Biol. **9**, R51 (2008)

G. Longo, M. Montévil, S.A. Kauffman, No entailing laws, but enablement in the evolution of the biosphere, in *GECCO '12 Proceedings of the 14th Annual Conference Companion Genetic and Evolutionary Computation*, pp. 1379–1392, 2012

J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Freeman & Co., Oxford, 1995)

J.M. Nordbotten, N.C. Stenseth, Asymmetric ecological conditions favor Red-Queen type of continued evolution over stasis. Proc. Natl. Acad. Sci. USA **113**, 1847–1852 (2016)

L.E. Orgel, Molecular replication. Nature **358**, 203–209 (1992)

R.K. Standish, Open-ended artificial evolution. Int. J. Comput. Intell. Appl. **3**, 167 (2003)

N.C. Stenseth, J. Maynard Smith, Coevolution in ecosystems: Red Queen evolution or stasis? Evolution **38**, 870–880 (1984)

J. Sutherland, Opinion: studies on the origin of life—the end of the beginning. Nat. Rev. Chem. **1** (2017). https://doi.org/10.1038/s41570-016-0012

E. Szathmáry, A classification of replicators and lambda-calculus models of biological organization. Proc. R. Soc. Lond. B **260**, 279–286 (1995)

E. Szathmáry, Units of evolution and units of life, in *Fundamentals of Life*, ed. by G. Pályi, L. Zucchi, L. Caglioti (Elsevier, Paris, 2002), pp. 181–195

E. Szathmáry, The biological significance of Gánti's work in 1971 and today, *The Principles of Life* (Oxford University Press, Gánti, 2003), pp. 157–168

E. Szathmáry, The origin of replicators and reproducers. Phil. Trans. R. Soc. Lond. B Biol. Sci. **361**, 1761–1776 (2006)

E. Szathmáry, M. Santos, C. Fernando, Evolutionary potential and requirements for minimal protocells. Top. Curr. Chem. **259**, 167–211 (2005)

J.W. Szostak, D.P. Bartel, P.L. Luisi, Synthesizing life. Nature **409**, 387–390 (2001)

L. van Valen, A new evolutionary law. Evol. Theory **1**, 1–30 (1973)

G. von Kiedrowski, O. Sijbren, P. Herdewijn, Welcome home, system chemists! J. Syst. Chem. **1**, 1 (2010)

C.H. Waddington, Paradigm for an evolutionary process, in *Towards a Theoretical Biology, Vol. 2. Sketches. An International Union of Biological Sciences Symposium*, Aug 1967 (Waddington, C.H., ed.). (Edinburgh University Press, 1969), pp. 106–123

G. Wagner, Evolutionary innovations and novelties: let us get down to business! Zool. Anzeiger. **256**, 75–81 (2015)

H.B. White, Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. **29**, 101–104 (1976)

Chapter 28
Genome Regulation Is All Non-local: Maps and Functions

Basuthkar J. Rao

Basics of Genome Design Principles and Functions

In this chapter, we try to envision genomes both as macroscopic entities that living systems carry in the form of genetic information as well as relevant details of the mechanisms that help maintain genome stability. The human genome contains approximately 6 billion bases of DNA, extending approximately to 2 m of DNA in a cell. A DNA double helical repeat unit scales to 3.5 nm, where the helix exhibits the properties of wormlike chain with persistence length of about 50 nm. Such DNA wormlike chain wraps around a protein-core (histone-octamer) with about 160–240 base pairs of DNA forming Nucleosome Core Particle (NCP). The stretches of DNA that separate two connecting NCPs are called linker DNA, thus leading to a bead-on-a-string type of structure (chromatin) constituting a 10 nm wide fiber. Moreover, the chromosomes within the nucleus are spatially confined, non-randomly, each in its own domain (Ishita et al. 2013; Chakraborty et al. 2015; Sarosh et al. 2016; Mugdha et al. 2016). These domains are much smaller than the typical size of a chromatin fiber in a good solvent; thereby implying extensive folding (compactification) stabilises the structure. The extent of compactification in interphase chromatin is considerably variable between regions harboring active *versus* and less active genes. Even within the different phases of cell cycle, a dramatic variation of compactification is apparent. Chromosome compactification (estimated to be of the order of several thousand fold) reaches the maximum in metaphase chromosomes where chromatin fibres are folded by the help of SMC

B. J. Rao (✉)
Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400005, India
e-mail: bjrao@tifr.res.in

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_28

motor proteins (Kristian et al. 2014; Hirano 2005). The rules of compactification and reversal of the same during cell-type specific gene expression changes (transcriptome) as well as during genome replication (where every DNA base pair of the genome needs to be opened up) remain far from clear.

In most mammals, all interphase chromosomes exist as spatially non-random physical entities called chromosome territories (CTs) (Cremer and Cremer 2001), untethered to the nuclear envelope. These CTs not only occupy non-random positions in the nucleus, but their relative positions with respect to other CTs also persist. This format of interphase chromosomal arrangement, referred to as the radial arrangement (Bickmore 2013; Cremer and Cremer 2010; Misteli 2010) is largely conserved in the mammalian lineage. A high-throughput analysis used across the whole genome for quantitatively assaying contact frequencies of a portion of genome with any other portion (Hi-C assay) (Lieberman-Aiden et al. 2009) has uncovered fractal nature of genome organization with an average fractal dimension of ≈1.08 in fibroblast cells of *Homo sapiens*. Even 3D rheology-based biophysical experiments have revealed that the mammalian interphase nuclei show fractal organization in the genomes (Bancaud et al. 2009). What are fractals and how do they manifest physically both in living and non-living systems?

A fractal is defined as *"a rough or fragmented geometric shape that can be subdivided in parts, each of which is (at least approximately) a reduced-size copy of the whole*" (Mandlebrot 1975). This property is referred to as self-similarity and the term "fractal" was coined by Benoît Mandelbrot in (1975), derived from the Latin word *fractus* meaning "broken" or "fractured". Fractals being self-similar structures exhibit similar features when examined at increasing magnifications. Several natural objects approximate to fractals, which include clouds, mountain ranges, lightning bolts, coastlines, and snowflakes etc. Interestingly, the converse need not be true: not all self-similar objects are fractals—for example, a straight Euclidean line is formally self-similar but has no fractal characteristics. The main features of fractal geometry are the following: Besides being self-similar, it is *too irregular* to be easily described in traditional Euclidean geometric language. There is no characteristic scale for their description. They obey a nonlinear, power law relationship. In a log–log diagram where log of the size is plotted against the values of a variable measured at different size scales shows a slope of a linear regression, which defines the fractal dimension.

The fractal concept seems ubiquitous in nature. The extension of this concept in biology has tremendously improved our understanding of the underlying design principles in biological systems. Fractality is discernible not only in the anatomy of the vascular and pulmonary systems but also in functional processes such as regulation of blood pressure, ion channel kinetics, heart rate variability, allometric scaling growth, allosteric enzyme kinetics, metabolic rates in mammals and population genetics etc. (Thamrin et al. 2010; Lisa 2009). The fractal design underscores the principle that biological structures could be built by rather simple, iterative schemes such that the complex designs could result out of simple recursive rule, applied across all scales, a theme that Darwinian selection has successfully adapted in Biology. Genomes, being large information storage systems, have successfully adapted fractal geometry rules in their designs and functions, whose intricate dynamics is the prime focus of the current article.

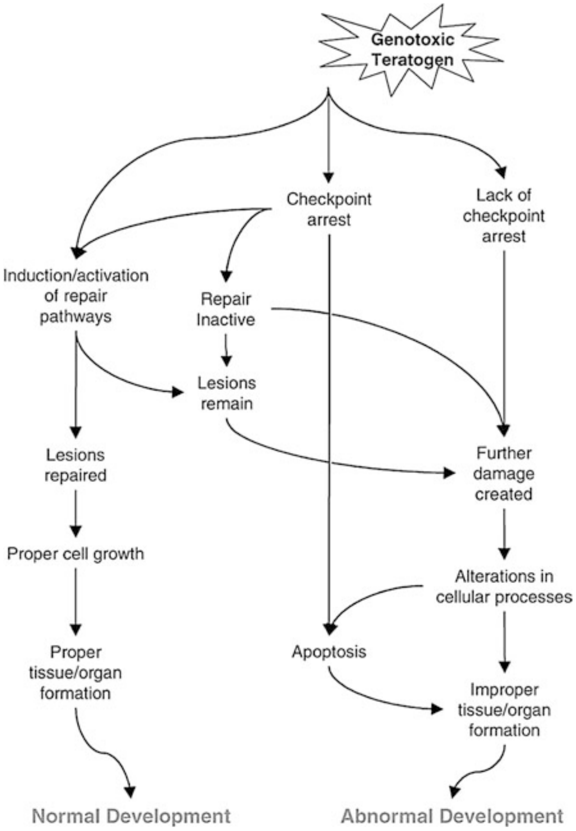
Genomes being large and critical information entities in cells are subject to intricate control mechanisms. The essence of genetic information seems to be its linear continuity of chemical content in a DNA polymer: Even a single break in the DNA helix anywhere in the genome seems to pose serious enough challenge that it leads to cell unviability. Every organism contains a definite sequence of information where internal rearrangements either within the linear scheme of a chromosome or even within the spatial arrangement of CTs in space is not tolerated by the cell, implying that a strict code of conduct is imposed by the mechanisms that maintain genome stability. This is not to mean that DNA polymers in the genome are not subject to spontaneous chemical and mechanical pressures that give rise to high rates of mutations and breaks even in normal cells. It is estimated that a human cellular genome in its normal life cycle exhibits several thousands of DNA changes (mutations and breaks) per every cell cycle. However, the resounding success of genome stability that rendered them the function of genetic material seems to stem from the fact that all living creatures have co-evolved very efficient and fairly elaborate DNA repair (genome maintenance) mechanisms, thereby ensuring that cells hardly experience high genomic stress. These repair responses are classically referred to as Genome homeostasis mechanisms whose details are intricately worked out.

A non-obvious aspect of genome dynamics that requires a special mention relates to the nature of genome designs: Genomes seem to be much more than a linear sequence of chemical information. Supra-chromosomal spatial organization encompassing all CTs in a cell seems to involve a concerted cross-talk among CTs in space such that even non-genomic sub-compartments play critical roles in orchestrating the “nucleography” within the nucleus. A change induced in any part of the genome (either in the form of mutations or chromatin folding or expression changes, etc.) seems to provoke compensatory changes elsewhere in the genome such that newer genomic equilibrium ensues in order to minimize the detrimental changes in the genome. Another facet of genome regulation that has come to light more recently relates to the intimate dialogue between genetics and epigenetics in a cell, where the former refers to DNA information in the genome whereas the latter to the chemical modifications in the histone protein that cover the DNA in chromatin. Epigenetic changes seem to be highly plastic, broad-based in specificity, affect the genome functions (transcriptional, replicational as well as repair related) very rampantly without affecting any sequence changes in DNA. Typically life-style related environmental affects such as dietary inputs, pathological burdens, microbiome-based effects etc. impact genome functions largely through epigenetic changes. In fact, ageing related changes have much more to do with epigenetics than genetics in an organism. So, currently, unraveling the mechanistic evaluation of how genetics and epigenetics cross talk within the genomes of different human populations has become a very important area of research. Nature (genetics) *versus* nurture (epigenetics) debates have ceased and given rise to how Nature-nurture mutually collaborate in maintaining genome fitness.

In order to better appreciate the genome homeostasis mechanisms, we elaborate below some aspects of intricate regulations genomes are subjected to. These features highlight molecular biological as well as protein-centric details that underline highly conserved aspects DNA-repair mechanisms that researchers have uncovered in the recent past. I alert the reader here that a good bit of description below is technical so that accuracy of information content is maintained.

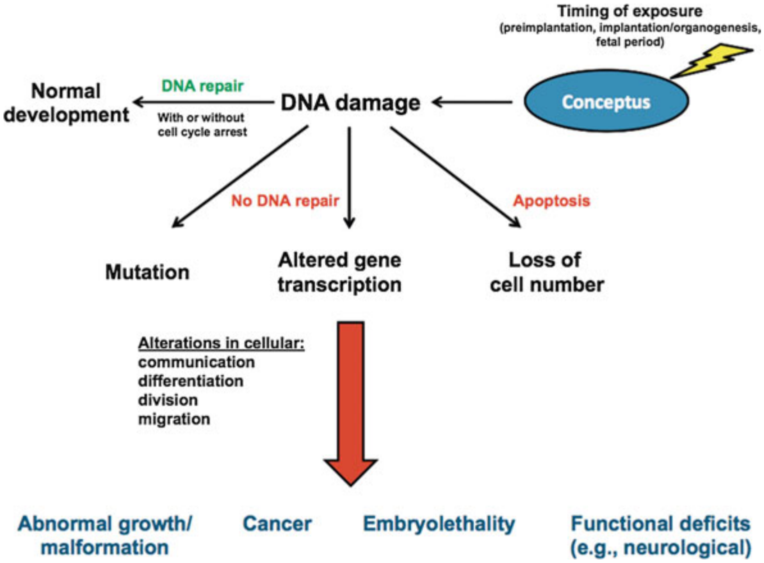
Genome Maintenance Is Essential for Normal Organismal Biology: Intricate Mechanisms

Genotoxic stressors pose a threat to genomic integrity of an organism from early stages of its life cycle including gamete formation, embryonic, fetal and post-natal development (Vinson and Hales 2002). Embryonic development is accompanied by rapid cell proliferation, increased DNA replication and shorter cell cycle, thus increasing the risk of DNA damage during development. The role of DNA repair enzymes during development must be to: (i) counteract endogenous genotoxic agents arising naturally due to extensive DNA metabolism; (ii) actively participate in rapid DNA metabolism due to tremendous cellular proliferation; or (iii) participate in non-repair pathways critical for proper tissue growth and development (Vinson and Hales 2002). DNA damage repair during organismal development is thus of utmost importance and a failure to do so would lead to altered gene expression, cell death and thus finally can be a factor determining the outcome of the conceptus. Developmental toxicity can also manifest into structural malformations, altered growth and functional deficits. Ablation of DNA repair genes has been associated with embryo lethality, enhanced sensitivity toward genotoxic exposures, developmental abnormalities, sterility, and predisposition to disease; thus further emphasizing the importance of DNA damage repair during development (Friedberg and Meira 2006; Khan et al. 2017).



Developmental Stages and Responses to DNA Damage

Two major factors affect the susceptibility of DNA damage during embryonic development: (a) the genotype and (b) stage of development when the embryo is exposed to a damaging agent. Depending on the dose and timing of damage exposure, a cell can respond to DNA damage in various ways. Cells that have incurred damage can repair themselves and proceed with development normally. In a scenario where repair is inefficient or nil, genetic lesions persist and are replicated, resulting in altered gene expression thereby leading to mutations. Errors in coding regions of the DNA can lead to genetic instabilities and may predispose the embryo to genetic disorders and cancers. Alternatively, in cases when exposed to high doses of DNA damage that are irreparable, cells can opt to undergo apoptosis, resulting in drastic reduction in cell numbers within a developing embryo. Altered gene expression, sudden loss of cells can further affect cell division, differentiation, migration, formation of tissues, communication and molecular cross-talk between cellular components. These factors in-turn would give rise to functional deficits and structural malformations in the conceptus.



(Pachkowski et al. 2011).

DNA Damage Response in Gametes

Genome integrity in the gametes ensures faithful transmission of genetic information from one generation to another. Sperm derived genome is known to be much more susceptible to DNA damage as compared to the maternally inherited genome. The oocyte is known to acquire genomic alterations mainly due to a long delay between entry into meiosis and fertilization of the oocyte, while the male germ cell undergoes thousands of mitotic division before the spermatogonia enters the meiotic prophase as opposed to the limited mitotic divisions that the oocyte undergoes. Exposure to DNA damaging agents during gametogenesis further poses a threat to the gametes. Few studies that have investigated DNA repair during gametogenesis have identified genes encoding proteins that are involved in various repair pathways. Studies in rhesus monkeys and human show that DNA damage recognition within the oocytes appears to occur via classical damage sensing proteins such as ATM and ATR followed by cell cycle checkpoint activation via CHEK1, CHEK2, PCNA, MDM2, TP53 and CDKN1B (Wells et al. 2005; Zheng et al. 2005). Expression of genes involved in Base Excision Repair (BER), Nucleotide Excision Repair (NER), Mismatch Repair (MMR), Homologous

Recombination (HR), and Non-homologous End-joining repair (NHEJ) have been documented within oocytes from various species such as rhesus monkeys, mouse, rat, human as well as *C.elegans*; suggesting that the maternal genome has myriad of pathways via which it can repair the DNA damage (Hendrey et al. 1995; Kanungo et al. 1997; Jurisicova et al. 1998; Wells et al. 2005; Zheng et al. 2005). In cases of extensive double strand breaks either due to genotoxic stress or through prolonged persistence of recombination errors, early oocytes (before mid blastula transition) are thought to prefer undergoing apoptosis and do not show a complete repair response (Vinson and Hales 2002; Carroll and Marangos 2013). In contrast to the oocytes, BER is the major repair pathway that dominates during the entire spermatogenesis process. NER and MMR occur at lower propensity in pre-mitotic cells and decline with age in post-mitotic spermatogenic cell types (Richardson et al. 2000; Xu et al. 2005). Interestingly, the status of DNA repair in elongated spermatids is still unclear with only expression of very few BER genes such as *Mpg* and *Apex* detected in these cell types (Aguilar-Mahecha et al. 2001). A recent study by Ermolaeva et al. reported a novel mechanism of delaying germ cell proliferation in *C.elegans* (Ermolaeva et al. 2013). Their results demonstrate that presence of genomic instability in germ cells elevates stress resistance in somatic cells and induces innate immune response (Ermolaeva et al. 2013). Thus, there appears to be a systemic response when germ cells incur DNA damage, further highlighting the importance of faithful transmission of genomic information to progeny cells.

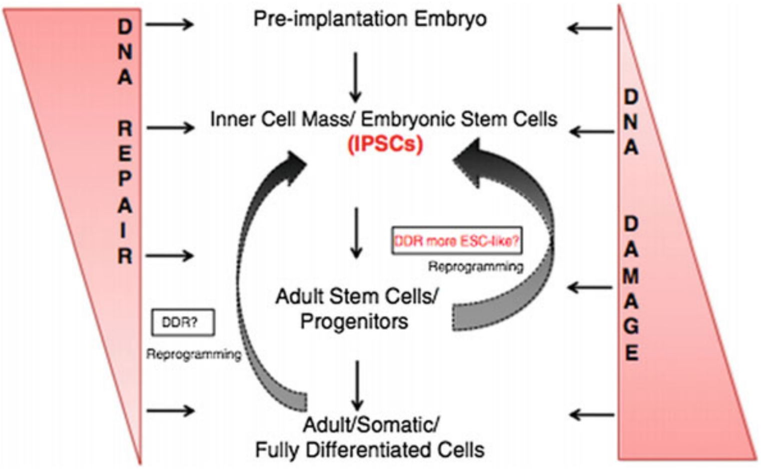
DNA Damage Response in Pre-implanted Embryos

Ovum differs from both somatic cells as well as germ cells in terms of their cell cycle, function and purpose. Genotoxic stress to pre-implantation embryos traditionally is thought to cause lethality and hence DNA repair pathways during this stage of the embryo are the least studied (Jacquet 2012). However, with studies reporting DNA damage mediated genomic instabilities resulting in malformations in the fetus due to exposure to irradiation during pre-implantations stages (Pampfer and Streffer 1989); a genome wide transcriptional analysis of embryos at this stage of development has been carried out. With no activation of the embryonic genes, pre-implanted embryos rely on maternal DNA repair transcripts for maintaining genomic integrity at an early stage (Jaroudi and SenGupta 2007). A maternal deficiency in the DNA MMR enzyme, Pms2, led to the formation of unrepaired replication errors during early cleavage divisions (Vinson and Hales 2002). Damages in paternal chromatin that have persisted from the germ cells are recognized and can repair during fertilization or in early embryos (Barton et al. 2007; Derijck et al. 2008). Before embryonic genome activation, single stranded breaks in an oocyte are predominantly repaired by BER. Studies in mouse and early rat embryos also suggest incidence of moderate levels of MMR and NER. Interestingly, reduction or absence of transcripts involved in BER, MMR and NER is observed as the oocyte matures. Thus, oxidative stress induced single stranded breaks are not of much threat to the matured oocyte post activation of the embryonic genome.

Studies in Rhesus monkey show a presence of many MMR transcripts throughout embryogenesis, but the actual ability of the embryo to combat DNA lesions using this repair mechanism appears to be limited (Zheng et al. 2005). This may be due to various reasons such as overexpression of *MSH3* protein that interferes with MutSα complex formation or low levels of *MLH1* expression (Zheng et al. 2005). Majority of the transcripts for factors involved in BER and NER were presented in varied amounts throughout embryogenesis in rhesus monkeys, while few transcripts for these repair pathways such as *Mpg* and *Cft2h3* have been detected in mouse embryos. In their extensive study, Zheng et al. reported presence of all DSB repair genes except RAD50 in Rhesus monkey embryos. Prevalence of DNA Damage Repair Response (DDR), mostly in form of HR (Chiruvella et al. 2012), has also been documented in human and mouse embryos with presence of genes such as ATM, ATR, BRAC1 and BRAC2. Very few NHEJ transcripts are detected at pre-implantation stages of the embryo, suggesting the importance of error-free repair in these undifferentiated cells (Chiruvella et al. 2012).

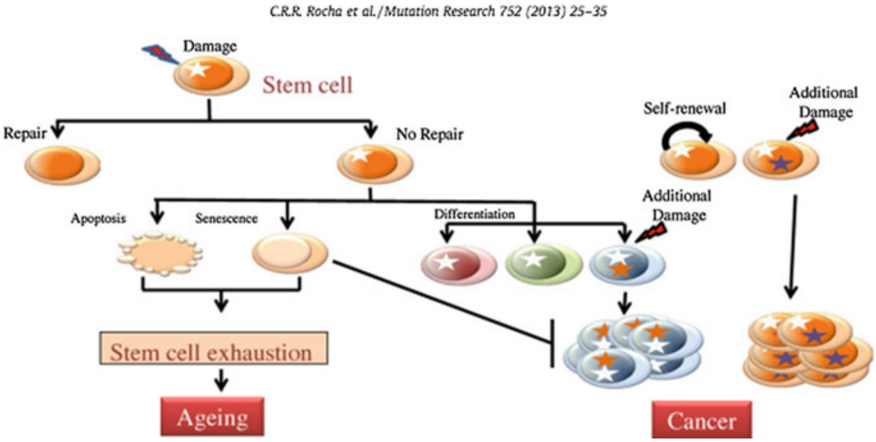
DNA Damage Response in Implanted Embryos

DNA damage repair pathways are in embryos post mid-gestational period starts to resemble the somatic cells. Spatial patterns of expression of repair genes become evident, thus serving as precursors for tissue specific DNA damage regulation. Vinson et al., in their extensive study, highlighted the developmental stage and tissue specific regulation of NER during rat gestation (Vinson and Hales 2002). Transcripts of genes involved in BER, MMR and HR are also present in varied levels at different developmental stages of an embryo (Vinson and Hales 2001a). Specific studies in rats (Craddock et al. 1984; Ozolins and Hales 1997; Vinson and Hales 2001a, b; Ertsey et al. 2004; Lee et al. 2004) and mice (Chiruvella et al. 2012) have demonstrated embryogenesis stage dependent alterations in expression of DNA repair genes. Expression of repair genes in a pre-implanted embryo is also known to be tissue specific. For instance, BER activity is thought to be differential across tissues as expression of APE is observed to be the highest in lives, thymus and the brain at gestational day 14. Further, expression of methyl guanine methyl transferase (MGMT) involved in NER is lower in hepatic cells as compared to that in rat brains (Craddock et al. 1984). Interestingly, the efficiencies of BER and MMR are higher in fetal brain as compared to the adult brain (Marietta et al. 1998; Riis et al. 2002). Thus, DNA repair in during embryogenesis is highly developmental stage as well as tissue specific.


DNA REPAIR
Pre-implantation Embryo
Inner Cell Mass/ Embryonic Stem Cells
(IPSCs)
DDR more ESC-like?
Reprogramming
DDR?
Reprogramming
Adult Stem Cells/ Progenitors
Adult/Somatic/ Fully Differentiated Cells
DNA DAMAGE

DNA Damage Response in Stem Cells

Two broad classes of stem cells exist in mammals: Embryonic stem cells (ESCs) and Adult stem cells or tissue specific stem cells (ASCs). Maintenance of the genome integrity is of utmost importance for both these types of stem cells as a failure to do so would affect their self-renewal abilities or impinge on passage of incorrect information to the progeny during differentiation program (Kenyon and Gerson 2007; Mandal et al. 2011). This would implicate on the tissue specific functions of these cells, thus affecting the organismal health.



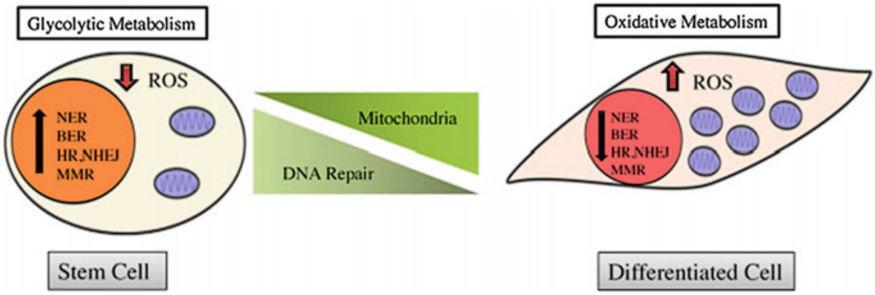
In this regards, ESCs constantly thrive to minimize the accumulation of DNA damage and genomic instabilities, and in order to achieve this these cells have an inherent capacity to prefer undergoing cell-cycle arrest followed by apoptosis and thus have a much lower tolerance threshold to DNA damage as compared to somatic cells (Heyer et al. 2000; Mantel et al. 2007; Desmarais et al. 2012). Both ESCs and ASCs are highly proliferative, have a characteristically shorter G1 phase of the cell cycle and are sensitive to IR (Becker et al. 2006). Since these cells spend most of their time in S-phase of the cell cycle and as a mechanism to protect the genomic information of these so called "mother-cells", majority of the repair occurs at G2/M block and is mediated via error-free HR mechanism or rarely via high fidelity NHEJ pathway (Adams et al. 2010a, b; Momcilovic et al. 2010; Pillai et al. 2010; Bogomazova et al. 2011).

Double-strand breaks (DSBs) in ESCs, like somatic cells, activate ATM signaling, resulting in phosphorylation and nuclear localization of ATM followed by activating its downstream targets such as CHK2, p53, etc. (Momcilovic et al. 2009). However unlike somatic cells, human ESCs undergo a G2/M arrest but show a negligible G1 arrest (Filion et al. 2009). Interestingly, inactivation of ATM in ESCs did not affect phosphorylation of its downstream targets such as CHK2, p53, gamma-H2AX upon IR treatment (Rass et al. 2013). Moreover, ATM null cells did not show any increase in genomic instability upon exogenous DNA damage (Rass et al. 2013). Taken together, these results suggest presence of an ATM independent repair pathway in ESCs. Owing to the faster rate of cell proliferation and the amount of time ESCs spend in S-phase, it is thought replication-induced damages is the major form of insult to the DNA in these cells. ATR is known to repair majority of the replication-induced repair and hence, it can be speculated that in absence of ATM, ESCs may employ ATR-mediated repair (Adams et al. 2010a). However, more studies on ATM-mediated repair or other ATM-independent pathways of repair in ESCs is necessary to completely understand damage signaling in these cells.

Mutations of important HR players such as RAD51, RAD54, MRN complex in ESCs, unlike somatic cells, are lethal (Friedberg et al. 2006; Nagaria et al. 2013); hence increasing the problems for studying the importance of HR in these cells. Although, this observation leads to the fact that unlike adult cells where NHEJ can take over the repair process in absence of HR; majority of the repair mediated activities is carried out solely by HR in ESCs. Error-prone repair via NHEJ appears to be indispensible in these ESCs, increasing the proportions of cells undergoing cell cycle arrest and apoptosis (Heyer et al. 2000; Mantel et al. 2007; Desmarais et al. 2012). All these mechanisms together ensure that the genetic information that is passed on through generations is preserved. Intriguingly, the amount of NHEJ increases in progenitor cells and ASCs is observed to be higher as compared to ESCs. As the primary motive of these cells is increasing cell numbers required for proliferation, these cells are thought to rely on a faster error-prone NHEJ pathway as compared to slower HR (Frosina 2010; de Laval et al. 2013a, b; Giachino et al. 2013). Thus an interesting phenomenon of a shift in relative contributions from HR versus NHEJ changes is observed as development proceeds.

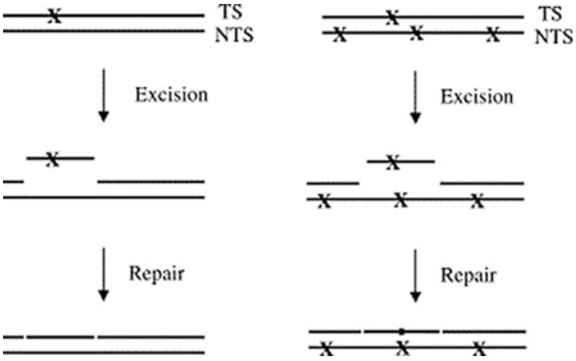
DNA Damage Response in Post-Mitotic Terminally Differentiated Cells

Most of the DNA damage response studies till date have been focused on either proliferative dividing cells or in cancer cells. Post mitotic cells do not undergo division hence it is valid to ask what is to repair within these cells? Diseases such as neurodegeneration and cardiomyopathy highlight the importance of DNA damage repair in these cells. Moreover, recent studies have suggested that DNA damage repair not only controls cellular events such as cell cycle and cell death; but also impinges on the fundamental differentiation program of the cell. The status of DNA repair in differentiated cells differs considerably as compared to stem cell. An overall attenuation of DNA repair pathways (NER, BER, HR and NHEJ) is observed in majority of the cell types post differentiation, with the focus shifting towards guarding the integrity of transcriptionally active genes. (Nouspikel 2009a, b; Fortini et al. 2013; Iyama and Wilson 2013; Lukasova et al. 2013; Rulten and Caldecott 2013). DSBs in differentiated cells occur mainly through DNA replication and dampening of DNA repair pathways allows these cells to economize and conserve cellular resources under stress-free conditions. In order to achieve this, global genome repair is attenuated in post mitotic cells and instead both transcribing and non-transcribing strands of active genes are repaired proficiently (Nouspikel and Hanawalt 2000; Chiruvella et al. 2012; Ramos-Espinosa et al. 2012; Schneider et al. 2012; Lukasova et al. 2013; Rodrigues et al. 2013).

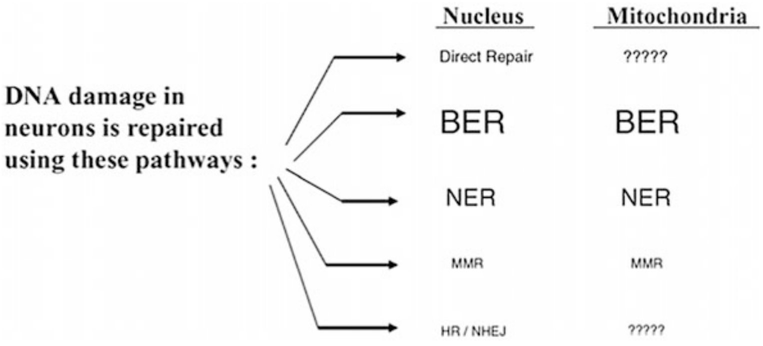


Pathways that repair single stranded lesions including BER and NER are down regulated in post mitotic cells of various human tissues such as striated muscles, neurons and macrophages (Nouspikel and Hanawalt 2000, 2006; Nouspikel et al. 2006; Hsu et al. 2007; Narciso et al. 2007; Nouspikel 2009a, b; Oliver et al. 2013; Sykora et al. 2013a, b). Terminally differentiated myotubes not only have a decreased BER capacity with downregulation of genes such as XRCC1 and DNA ligase I (Kulkarni et al. 2008), but also are resistant to treatments with SSB inducing agents (Fortini et al. 2012). Although post-mitotic neurons display lower survival rate and decreased repair capacities, activities of proteins such as XRCC1 is vital in order to prevent loss of cerebellar interneurons and abnormal hippocampal functions (Lee et al. 2009).

Interestingly, DNA repair in active genes via transcription-coupled repair is unaffected in differentiated cells. In a scenario when global genome repair is attenuated, TCR relies on the non-transcribed strand in order to repair the newly transcribed strand. Thus, it is important for differentiated cells to repair lesions in the non-transcribed strand, as accumulation of damage in these strands would lead to increase in mutations compromising the genomic integrity of expressed genes, further manifesting into metabolic dysfunction and cell death. Differentiated cells have thus adapted a phenomenon termed as differentiation associated repair (DAR) in order to maintain the integrity of the non-transcribed strand (Nouspikel and Hanawalt 2002; Nouspikel 2007, 2009a, b). Although NER is blocked at global repair level, various studies have documented the presence of NER enzymes in differentiated cells. Studies from Nouspikel et al. suggests that these NER enzymes are recruited to non-transcribed strand in differentiated cell, thus assisting DAR (Nouspikel and Hanawalt 2002; Nouspikel 2007, 2009a, b).



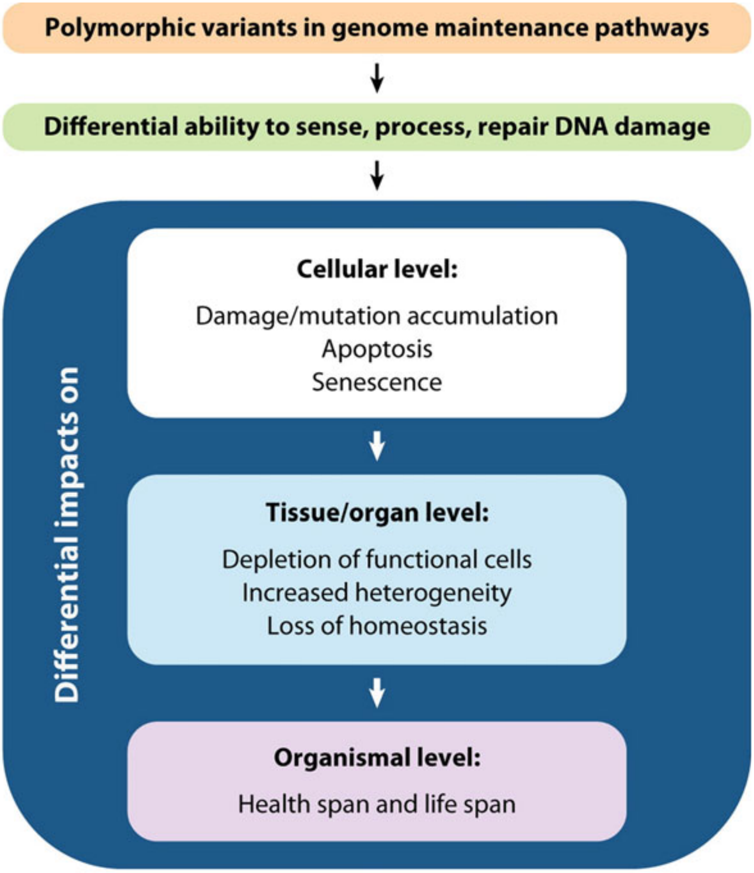
Repair of DSBs in post-mitotic cells is highly tissue specific. In general, none of the terminally differentiated cells including neurons, astrocytes, myotubes, adipocytes hematopoietic cells and granulocytes recruit HR pathway upon induction of DNA damage response (Nouspikel and Hanawalt 2002; Lal et al. 2009). Differentiated cells that have regenerative capacities such as blood, mesenchymal or muscle cells usually are known to suppress DSB repair in response to DSB accumulation (Fortini et al. 2012; Lukasova et al. 2013; Oliver et al. 2013). Recent studies demonstrate that in blood cells specific microRNAs (miR-24) serve as signals to suppress DSB repair response upon DNA damage induction (Lal et al. 2009). Regenerative differentiated cells in contrast to adapting error-prone NHEJ pathway, that would result in viable but malfunctioning cells, prefer to induce apoptosis in response to DSB formation. Conversely, long-lived terminally differentiated cells such as neurons, adipocytes and astrocytes do not have this choice with their limited regenerative capabilities and hence opt for error-prone NHEJ (Kruman 2004; Meulle et al. 2008; Tomashevski et al. 2010; Marinoglou 2012; Ramos-Espinosa et al. 2012; Schneider et al. 2012). While adipocytes repair DSBs faster as compared to its precursor cells due to overexpression of NHEJ pathways genes (Meulle et al. 2008), neurons have been found to activate NHEJ pathway genes only upon re-entry into the cell cycle via suppression of ATM (Kruman 2004; Schwartz et al. 2007; Schmetsdorf et al. 2009; Tomashevski et al. 2010). Terminally differentiated radio-resistant astrocytes lack functional DDR signaling due to repression at transcriptional level (Schneider et al. 2012).



Interestingly, there is a potential cross talk between DDR and differentiation; whereby DDR acts as a signal to control differentiation. Lesions in pre-B cells (Bredemeyer et al. 2008; Sherman et al. 2010) or NSCs (Armesilla-Diaz et al. 2009; Armesilla-Diaz et al. 2009; Tedeschi and Di Giovanni 2009), suppresses self-renewal capacities of these cells and favors differentiation into germinal B-cells and neurons respectively via p53 dependent DDR signaling. Conversely, in skeletal muscle progenitors, genotoxic stress blocks the transcription of differentiated specific genes while DNA is being repaired. DNA damage induced phosphorylation of MyoD suppresses Myo-D dependent transcription, thus negatively regulating differentiation in myoblasts (Simonatto et al. 2011; Simonatto 2013). Thus, DDR, in a cell type and cell cycle phase specific manner, can both positively and negatively regulate differentiation.

DNA Damage-Repair and Ageing

Ageing is a complex phenomenon that is a consequence of multiple causative features. One of the theories of ageing "The Soma Theory" states that during the course of its life span an individual's genome is faced with various types of assaults (Kirkwood 1977) including those as a result of side attacks from free-radicals produced during various biological reactions within the cell (Harman 1956, 1973). Although the organisms body have been equipped with machinery to prevent or repair these damages, evolutionary pressures have forced organisms to invest more energy and resources in maintenance and repair of germ-line cells as compared to that in somatic cells. Recent work by Schuler et al. demonstrate that ageing cells possess multiple 53BP1 clusters which do not colocalize with pKu70, thus supporting the hypothesis that repair in somatic cells is compromised with age (Schuler and Rube 2013). This bias leads to accumulation of non-heritable mutations in the genome, which might be the primary factor influencing ageing and life span of the organism (Harman 1956, 1973; Kirkwood 1977).


Polymorphic variants in genome maintenance pathways
Differential ability to sense, process, repair DNA damage
Differential impacts on
Cellular level:
Damage/mutation accumulation
Apoptosis
Senescence
Tissue/organ level:
Depletion of functional cells
Increased heterogeneity
Loss of homeostasis
Organismal level:
Health span and life span

Vijg J, Suh Y. 2013.
Annu. Rev. Physiol. 75:645–68

Figure: The impact of genetic variation in genome maintenance genes on phenotypes at the cellular, tissue/organ, and organismal levels

Using premature ageing disorders as a model system, data accumulated by various studies suggest that alteration in chromatin structure and accumulation of

DNA damage (in particular DSBs) are the two major contributors of ageing. A recent attempt using systems biology approach on senescent cells has revealed that the complexity of ageing is dependent on interactions between three aspects: (a) Mitochondrial dysfunction, (b) telomere erosion and (c) chromatin structure (Kirkwood 2011). It is thus pertinent to understand DNA damage vis-à-vis chromatin structure, mitochondrial function and telomere attrition in order to unravel the link between ageing and DNA damage.

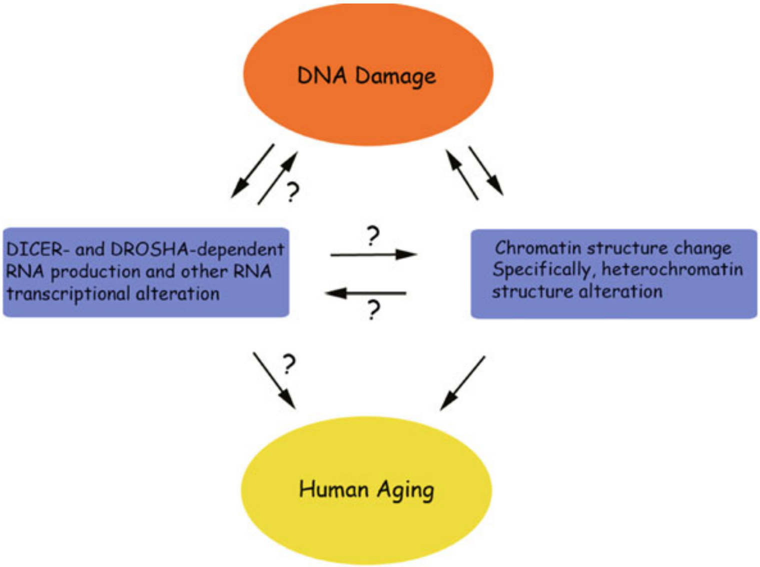
DNA Damage and Chromatin Structure

Various studies have demonstrated the importance of chromatin structure and DNA damage as a probable cause towards organismal ageing (Gong 2013), however few studies have tried to address the cross talk between these two cellular features. In a quest to decipher whether damage leads to structural changes in the chromatin or vice versa, Pegoraro et al. (2009), Pegoraro and Misteli (2009) knocked down the NURD protein complex that governs establishment of heterochromatin. They observed that the absence of functional NURD complex first leads to aberrant chromatin structure followed by induction of DNA damage. These observations suggest that epigenetic and chromatin structure changes drive DNA damage induction (Pegoraro et al. 2009; Pegoraro and Misteli 2009). Alternatively, Schuler and colleagues visualized DNA damage dependent structural defects in hetereochromatic domains (Schuler and Rube 2013).

Although the sequence of events remains debatable, it is definite that distinct chromatin structural alterations impinge on organismal ageing. Specifically, various studies in premature ageing disease patients have demonstrated that the process of ageing is tightly regulated by densely packaged chromatin. Aged organisms as well as premature ageing syndrome cells display deficit in heterochromatin maintenance resulting in prominent loss of heterochromatin structure, marked down regulation of heterochromatin proteins and altered patterns of histone modification (Goldman et al. 2004; Scaffidi and Misteli 2005, 2006; Shumaker et al. 2006; Larson et al. 2012). In turn studies in *C.elegans* have attributed presence of more open chromatin structure to shorter lifespan in worms (Hamilton et al. 2005; Han and Brunet 2012). Interestingly in ageing Drosophila, down regulation of heterochromatin genes is accompanied with increase in ribosomal RNA transcription (Eickbush et al. 2008; Larson et al. 2012). Thus, it can be further extrapolated that decrease in amount of heterochromatin and a concomitant increase in RNA transcription accelerates growth and ageing (Chen et al. 2007; Hansen et al. 2007; Pan et al. 2007).

RNA transcription is also an integral part of cellular response to DNA damage. In a recent study by Francia et al., the authors demonstrated that small RNA molecules produced by RNase type III enzymes DICER and DROSHA are vital to initiate a DNA damage response (Francia et al. 2012). Interestingly, DNA damage-induced RNA can regulate cellular senescence in cultured human and mouse cells, and in living zebrafish larvae (Mudhasani et al. 2008; Francia et al.

2012). However, it is still unclear if these small RNA molecules regulate the maintenance of heterochromatin, thus accelerating senescence and ageing. In summary, organismal ageing appears to be largely the remit of structural changes to chromatin, potentially leading to epigenetically induced transcriptional deregulation, via DNA damage and telomere shortening and/or mitochondrial dysfunction.


DNA Damage
?
?
DICER- and DROSHA-dependent RNA production and other RNA transcriptional alteration
?
Chromatin structure change Specifically, heterochromatin structure alteration
?
Human Aging

Concluding Thoughts

One would often wonder why biological systems have evolved such complex mechanisms to maintain regulation: The explanation is largely evolution-centric which can be elaborated in the following manner. All life forms are sheer accidents of nature where congenial states have put together conditions that favored certain biochemical reactions giving rise to self-perpetuating cycles of material replication. Thermodynamic considerations drove additional condensation reactions which led to functional compartmentalization of biochemical machines, thereby giving rise to primitive cells which were constantly subjected to environmental pressures. It is possible that life forms became extinct multiple times during the onslaught of environmental hardships, followed by selections. In order to shape biochemical machines that function robustly even against seemingly unfavorable conditions, biochemical pathways acquired complex network based cross talk where reactants and products impact the pathways non-linearly via feedback and feed forward regulations. As the cellular viability, adaptability and its evolvability became entirely systemic responses, integration of biochemical networks acquired high significance during successful rounds of evolutionary selections. When complex networks got integrated, multiple options opened up for the systems, which when subjected to Darwinian selection lead to improved fitness. Those systems that showed facile integration of networks at low cost, without sacrificing much of functional outputs as the external conditions varied, got robustly selected. Highly successful networks got conserved across evolution, being used recursively across different themes. Conversely less successful networks were jettisoned or used rather infrequently as the situation demanded. In a system biology centric view, complex regulatory systems that are high in noise-level offered myriads of functional opportunities in the system, a condition that is imminently well suited for evolution, a process that is fundamentally "directionless". The apparent "direction" that emerges during evolution is entirely linked to the "baggage from the past" and the high cost associated with reinventing of a new "wheel" afresh. So the biological systems in general and genomes in particular are caught in the web of "complex regulations of the past", but evolve further due to complexity-driven multiple possibilities endowed with the system. Scale-free designs (fractals), excessive non-linearity associated with the operational functions, less hardwiring and more plasticity in regulation are some of the cardinal features that kept successful biological designs afloat in spite of harsh external conditions. The intricacies we have learned in these regulations are highly unlikely to change and therefore are here to stay.

References

B.R. Adams, S.E. Golding, R.R. Rao, K. Valerie, Dynamic dependence on ATR and ATM for double-strand break repair in human embryonic stem cells and neural descendants. PLoS ONE **5**(4), e10001 (2010a)

B.R. Adams, A.J. Hawkins, L.F. Povirk, K. Valerie, ATM-independent, high-fidelity nonhomologous end joining predominates in human embryonic stem cells. Aging (Albany NY) **2**(9), 582–596 (2010b)

A. Aguilar-Mahecha, B.F. Hales, B. Robaire, Expression of stress response genes in germ cells during spermatogenesis. Biol. Reprod. **65**(1), 119–127 (2001)

A. Armesilla-Diaz, P. Bragado, I. Del Valle, E. Cuevas, I. Lazaro, C. Martin, J.C. Cigudosa, A. Silva, p53 regulates the self-renewal and differentiation of neural precursors. Neuroscience **158** (4), 1378–1389 (2009a)

A. Armesilla-Diaz, G. Elvira, A. Silva, p53 regulates the proliferation, differentiation and spontaneous transformation of mesenchymal stem cells. Exp. Cell Res. **315**(20), 3598–3610 (2009b)

A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, J. Ellenberg, Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. EMBO J. **28**, 3785–3798 (2009)

T.S. Barton, B. Robaire, B.F. Hales, DNA damage recognition in the rat zygote following chronic paternal cyclophosphamide exposure. Toxicol. Sci. **100**(2), 495–503 (2007)

K.A. Becker, P.N. Ghule, J.A. Therrien, J.B. Lian, J.L. Stein, A.J. van Wijnen, G.S. Stein, Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase. J. Cell. Physiol. **209**(3), 883–893 (2006)

W.A. Bickmore, The spatial organization of the human genome. Annu. Rev. Genomics Hum. Genet. **14**, 67–84 (2013)

A.N. Bogomazova, M.A. Lagarkova, L.V. Tskhovrebova, M.V. Shutova, S.L. Kiselev, Error-prone nonhomologous end joining repair operates in human pluripotent stem cells during late G2. Aging (Albany NY) **3**(6), 584–596 (2011)

A.L. Bredemeyer, B.A. Helmink, C.L. Innes, B. Calderon, L.M. McGinnis, G.K. Mahowald, E.J. Gapud, L.M. Walker, J.B. Collins, B.K. Weaver, L. Mandik-Nayak, R.D. Schreiber, P.M. Allen, M.J. May, R.S. Paules, C.H. Bassing, B.P. Sleckman, DNA double-strand breaks activate a multi-functional genetic program in developing lymphocytes. Nature **456**(7223), 819–823 (2008)

J. Carroll, P. Marangos, The DNA damage response in mammalian oocytes. Front. Genet. **4**, 117 (2013)

S. Chakraborty, I. Mehta, M. Kulashreshtha, B.J. Rao, Quantitative analysis of chromosome localization in the nucleus. Methods Mol. Biol. **1228**, 223–233 (2015). https://doi.org/10.1007/978-1-4939-1680-1_17

D. Chen, K.Z. Pan, J.E. Palter, P. Kapahi, Longevity determined by developmental arrest genes in Caenorhabditis elegans. Aging Cell **6**(4), 525–533 (2007)

K.K. Chiruvella, R. Sebastian, S. Sharma, A.A. Karande, B. Choudhary, S.C. Raghavan, Time-dependent predominance of nonhomologous DNA end-joining pathways during embryonic development in mice. J. Mol. Biol. **417**(3), 197–211 (2012)

V.M. Craddock, A.R. Henderson, S. Gash, Repair and replication of DNA in rat brain and liver during foetal and post-natal development, in relation to nitroso-alkyl-urea induced carcinogenesis. J. Cancer Res. Clin. Oncol. **108**(1), 30–35 (1984)

T. Cremer, C. Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet. **2**, 292–301 (2001)

T. Cremer, M. Cremer, Chromosome territories. Cold Spring Harb. Perspect. Biol. **2**, a003889 (2010)

B. de Laval, P. Pawlikowska, D. Barbieri, C. Besnard-Guerin, A. Cico, R. Kumar, M. Gaudry, V. Baud, F. Porteu, Thrombopoietin promotes NHEJ DNA repair in hematopoietic stem cells through specific activation of Erk and NF-kappaB pathways and their target IEX-1. *Blood*, 2014 Jan 23, vol. 123(4), pp. 509–519 (2013). https://doi.org/10.1182/blood-2013-07-515874. Epub 2013 Nov 1

B. de Laval, P. Pawlikowska, L. Petit-Cocault, C. Bilhou-Nabera, G. Aubin-Houzelstein, M. Souyri, F. Pouzoulet, M. Gaudry, F. Porteu, Thrombopoietin-increased DNA-PK-dependent DNA repair limits hematopoietic stem and progenitor cell mutagenesis in response to DNA damage. Cell Stem Cell **12**(1), 37–48 (2013)

A. Derijck, G. van der Heijden, M. Giele, M. Philippens, P. de Boer, DNA double-strand break repair in parental chromatin of mouse zygotes, the first cell cycle as an origin of de novo mutation. Hum. Mol. Genet. **17**(13), 1922–1937 (2008)

J.A. Desmarais, M.J. Hoffmann, G. Bingham, M.E. Gagou, M. Meuth, P.W. Andrews, Human embryonic stem cells fail to activate CHK1 and commit to apoptosis in response to DNA replication stress. Stem Cells **30**(7), 1385–1393 (2012)

D.G. Eickbush, J. Ye, X. Zhang, W.D. Burke, T.H. Eickbush, Epigenetic regulation of retrotransposons within the nucleolus of Drosophila. Mol. Cell. Biol. **28**(20), 6452–6461 (2008)

M.A. Ermolaeva, A. Segref, A. Dakhovnik, H.L. Ou, J.I. Schneider, O. Utermohlen, T. Hoppe, B. Schumacher, DNA damage in germ cells induces an innate immune response that triggers systemic stress resistance. Nature **501**(7467), 416–420 (2013)

R. Ertsey, C.J. Chapin, J.A. Kitterman, L.M. Scavo, Ontogeny of poly(ADP-ribose) polymerase-1 in lung and developmental implications. Am. J. Respir. Cell Mol. Biol. **30**(6), 853–861 (2004)

T.M. Filion, M. Qiao, P.N. Ghule, M. Mandeville, A.J. van Wijnen, J.L. Stein, J.B. Lian, D.C. Altieri, G.S. Stein, Survival responses of human embryonic stem cells to DNA damage. J. Cell. Physiol. **220**(3), 586–592 (2009)

P. Fortini, C. Ferretti, E. Dogliotti, The response to DNA damage during differentiation: pathways and consequences. Mutat. Res. **743–744**, 160–168 (2013)

P. Fortini, C. Ferretti, B. Pascucci, L. Narciso, D. Pajalunga, E.M. Puggioni, R. Castino, C. Isidoro, M. Crescenzi, E. Dogliotti, DNA damage response by single-strand breaks in terminally differentiated muscle cells and the control of muscle integrity. Cell Death Differ. **19** (11), 1741–1749 (2012)

S. Francia, F. Michelini, A. Saxena, D. Tang, M. de Hoon, V. Anelli, M. Mione, P. Carninci, F. d'Adda di Fagagna, Site-specific DICER and DROSHA RNA products control the DNA-damage response. Nature **488**(7410), 231–235 (2012)

E.C. Friedberg, A. Aguilera, M. Gellert, P.C. Hanawalt, J.B. Hays, A.R. Lehmann, T. Lindahl, N. Lowndes, A. Sarasin, R.D. Wood, DNA repair: from molecular mechanism to human disease. DNA Repair (Amst) **5**(8), 986–996 (2006)

E.C. Friedberg, L.B. Meira, Database of mouse strains carrying targeted mutations in genes affecting biological responses to DNA damage Version 7. DNA Repair (Amst) **5**(2), 189–209 (2006)

G. Frosina, The bright and the dark sides of DNA repair in stem cells. J. Biomed. Biotechnol. **2010**, 845396 (2010)

C. Giachino, L. Orlando, V. Turinetto, Maintenance of genomic stability in mouse embryonic stem cells: relevance in aging and disease. Int. J. Mol. Sci. **14**(2), 2617–2636 (2013)

R.D. Goldman, D.K. Shumaker, M.R. Erdos, M. Eriksson, A.E. Goldman, L.B. Gordon, Y. Gruenbaum, S. Khuon, M. Mendez, R. Varga, F.S. Collins, Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome. Proc. Natl. Acad. Sci. USA **101**(24), 8963–8968 (2004)

L.W.E. Gong, S.Y. Lin, Chromatin Remodeling in DNA Damage Response and Human Aging. Licensee InTech, 2013

B. Hamilton, Y. Dong, M. Shindo, W. Liu, I. Odell, G. Ruvkun, S.S. Lee, A systematic RNAi screen for longevity genes in C. elegans. Genes Dev. **19**(13), 1544–1555 (2005)

S. Han, A. Brunet, Histone methylation makes its mark on longevity. Trends Cell Biol. **22**(1), 42–49 (2012)

M. Hansen, S. Taubert, D. Crawford, N. Libina, S.J. Lee, C. Kenyon, Lifespan extension by conditions that inhibit translation in Caenorhabditis elegans. Aging Cell **6**(1), 95–110 (2007)

D. Harman, Aging: a theory based on free radical and radiation chemistry. J. Gerontol. **11**(3), 298–300 (1956)

D. Harman, Free radical theory of aging. Triangle **12**(4), 153–158 (1973)

J. Hendrey, D. Lin, M. Dziadek, Developmental analysis of the Hba(th-J) mouse mutation: effects on mouse peri-implantation development and identification of two candidate genes. Dev. Biol. **172**(1), 253–263 (1995)

B.S. Heyer, A. MacAuley, O. Behrendtsen, Z. Werb, Hypersensitivity to DNA damage leads to increased apoptosis during early mouse development. Genes Dev. **14**(16), 2072–2084 (2000)

P.H. Hsu, P.C. Hanawalt, T. Nouspikel, Nucleotide excision repair phenotype of human acute myeloid leukemia cell lines at various stages of differentiation. Mutat. Res. **614**(1–2), 3–15 (2007)

T. Hirano, Philos. Trans. R. Soc. Lond. B Biol. Sci. **360**(1455), 507–514 (2005). SMC proteins and chromosome mechanics: from bacteria to humans

S.M. Ishita, M. Kulashreshtha, S. Chakraborty, U. Kolthur-Seetharam, B.J. Rao, Chromosome territories reposition during DNA damage-repair response. Genome Biol. **14**(12), R135 (2013)

T. Iyama, D.M. Wilson 3rd, DNA repair mechanisms in dividing and non-dividing cells. DNA Repair (Amst) **12**(8), 620–636 (2013)

P. Jacquet, Developmental defects and genomic instability after x-irradiation of wild-type and genetically modified mouse pre-implantation and early post-implantation embryos. J. Radiol. Prot. **32**(4), R13–R36 (2012)

S. Jaroudi, S. SenGupta, DNA repair in mammalian embryos. Mutat. Res. **635**(1), 53–77 (2007)

A. Jurisicova, K.E. Latham, R.F. Casper, R.F. Casper, S.L. Varmuza, Expression and regulation of genes associated with cell death during murine preimplantation embryo development. Mol. Reprod. Dev. **51**(3), 243–253 (1998)

J. Kanungo, R.S. Cameron, Y. Takeda, J.A. Hardin, DNA-dependent protein phosphorylation activity in Xenopus is coupled to a Ku-like protein. Biol. Bull. **193**(2), 147–152 (1997)

J. Kenyon, S.L. Gerson, The role of DNA damage repair in aging of adult stem cells. Nucleic Acids Res. **35**(22), 7557–7565 (2007)

C. Khan, S. Muliyil, C. Ayyub, B.J. Rao, The initiator caspase Dronc plays a non-apoptotic role in promoting DNA damage signalling in D. melanogaster. J. Cell Sci. **130**(18), 2984–2995 (2017). https://doi.org/10.1242/jcs.200782. Epub 2017 Jul 27

T.B. Kirkwood, Evolution of ageing. Nature **270**(5635), 301–304 (1977)

T.B. Kirkwood, Systems biology of ageing and longevity. Philos. Trans. R. Soc. Lond. B Biol. Sci. **366**(1561), 64–70 (2011)

J. Kristian, T. Kanno, K. Shirahige, C. Sjögren, The maintenance of chromosome structure: positioning and functioning of SMC complexes. Nat. Rev. Mol. Cell Biol. **15**, 601–614 (2014)

I.I. Kruman, Why do neurons enter the cell cycle? Cell Cycle **3**(6), 769–773 (2004)

A. Kulkarni, D.R. McNeill, M. Gleichmann, M.P. Mattson, D.M. Wilson 3rd, XRCC1 protects against the lethality of induced oxidative DNA damage in nondividing neural cells. Nucleic Acids Res. **36**(15), 5111–5121 (2008)

A. Lal, Y. Pan, F. Navarro, D.M. Dykxhoorn, L. Moreau, E. Meire, Z. Bentwich, J. Lieberman, D. Chowdhury, miR-24-mediated downregulation of H2AX suppresses DNA repair in terminally differentiated blood cells. Nat. Struct. Mol. Biol. **16**(5), 492–498 (2009)

K. Larson, S.J. Yan, A. Tsurumi, J. Liu, J. Zhou, K. Gaur, D. Guo, T.H. Eickbush, W.X. Li, Heterochromatin formation promotes longevity and represses ribosomal RNA synthesis. PLoS Genet. **8**(1), e1002473 (2012)

H.M. Lee, Z. Hu, H. Ma, G.H. Greeley Jr., C. Wang, E.W. Englander, Developmental changes in expression and subcellular localization of the DNA repair glycosylase, MYH, in the rat brain. J. Neurochem. **88**(2), 394–400 (2004)

Y. Lee, S. Katyal, Y. Li, S.F. El-Khamisy, H.R. Russell, K.W. Caldecott, P.J. McKinnon, The genesis of cerebellar interneurons and the prevention of neural DNA damage require XRCC1. Nat. Neurosci. **12**(8), 973–980 (2009)

E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**, 289–293 (2009)

G.A. Lisa, The fractal geometry of life. Riv. Biol. **102**, 29–59 (2009)

E. Lukasova, Z. Koristek, M. Klabusay, V. Ondrej, S. Grigoryev, A. Bacikova, M. Rezacova, M. Falk, J. Vavrova, V. Kohutova, S. Kozubek, Granulocyte maturation determines ability to release chromatin NETs and loss of DNA damage response; these properties are absent in immature AML granulocytes. Biochim. Biophys. Acta **1833**(3), 767–779 (2013)

P.K. Mandal, C. Blanpain, D.J. Rossi, DNA damage response in adult stem cells: pathways and consequences. Nat. Rev. Mol. Cell Biol. **12**(3), 198–202 (2011)

B.B. Mandelbrot, Stochastic models for the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands. Proc. Natl. Acad. Sci. USA **72**, 3825–3828 (1975)

C. Mantel, Y. Guo, M.R. Lee, M.K. Kim, M.K. Han, H. Shibayama, S. Fukuda, M.C. Yoder, L.M. Pelus, K.S. Kim, H.E. Broxmeyer, Checkpoint-apoptosis uncoupling in human and mouse embryonic stem cells: a source of karyotpic instability. Blood **109**(10), 4518–4527 (2007)

C. Marietta, F. Palombo, P. Gallinari, J. Jiricny, P.J. Brooks, Expression of long-patch and short-patch DNA mismatch repair proteins in the embryonic and adult mammalian brain. Brain Res. Mol. Brain Res. **53**(1–2), 317–320 (1998)

K. Marinoglou, The role of the DNA damage response kinase ataxia telangiectasia mutated in neuroprotection. Yale J. Biol. Med. **85**(4), 469–480 (2012)

A. Meulle, B. Salles, D. Daviaud, P. Valet, C. Muller, Positive regulation of DNA double strand break repair activity during differentiation of long life span cells: the example of adipogenesis. PLoS ONE **3**(10), e3345 (2008)

T. Misteli, Higher-order genome organization in human disease. Cold Spring Harb. Perspect. Biol. **2**, a000794 (2010)

O. Momcilovic, S. Choi, S. Varum, C. Bakkenist, G. Schatten, C. Navara, Ionizing radiation induces ataxia telangiectasia mutated-dependent checkpoint signaling and G(2) but not G(1) cell cycle arrest in pluripotent human embryonic stem cells. Stem Cells **27**(8), 1822–1835 (2009)

O. Momcilovic, L. Knobloch, J. Fornsaglio, S. Varum, C. Easley, G. Schatten, DNA damage responses in human induced pluripotent stem cells and embryonic stem cells. PLoS ONE **5** (10), e13410 (2010)

R. Mudhasani, Z. Zhu, G. Hutvagner, C.M. Eischen, S. Lyle, L.L. Hall, J.B. Lawrence, A.N. Imbalzano, S.N. Jones, Loss of miRNA biogenesis induces p19Arf-p53 signaling and senescence in primary cells. J. Cell Biol. **181**(7), 1055–1063 (2008)

K. Mugdha, I.S. Mehta, P. Kumar, B.J. Rao, Chromosome territory relocation during DNA repair requires nuclear myosin1β recruitment to chromatin mediated by ϒ-H2AX signaling. Nucleic Acids Research 2016 Jun 30. pii: gkw573

P. Nagaria, C. Robert, F.V. Rassool, DNA double-strand break response in stem cells: mechanisms to maintain genomic integrity. Biochimica et Biophysica Acta (BBA)—Gen. Subj. **1830**(2), 2345–2353 (2013)

L. Narciso, P. Fortini, D. Pajalunga, A. Franchitto, P. Liu, P. Degan, M. Frechet, B. Demple, M. Crescenzi, E. Dogliotti, Terminally differentiated muscle cells are defective in base excision DNA repair and hypersensitive to oxygen injury. Proc. Natl. Acad. Sci. USA **104**(43), 17010–17015 (2007)

T. Nouspikel, DNA repair in differentiated cells: some new answers to old questions. Neuroscience **145**(4), 1213–1221 (2007)

T. Nouspikel, DNA repair in mammalian cells: Nucleotide excision repair: variations on versatility. Cell. Mol. Life Sci. **66**(6), 994–1009 (2009a)

T. Nouspikel, DNA repair in mammalian cells: So DNA repair really is that important? Cell. Mol. Life Sci. **66**(6), 965–967 (2009b)

T. Nouspikel, P.C. Hanawalt, Terminally differentiated human neurons repair transcribed genes but display attenuated global DNA repair and modulation of repair gene expression. Mol. Cell. Biol. **20**(5), 1562–1570 (2000)

T. Nouspikel, P.C. Hanawalt, DNA repair in terminally differentiated cells. DNA Repair (Amst) **1** (1), 59–75 (2002a)

T. Nouspikel, P.C. Hanawalt, DNA repair in terminally differentiated cells. DNA Repair **1**(1), 59–75 (2002b)

T. Nouspikel, P.C. Hanawalt, Impaired nucleotide excision repair upon macrophage differentiation is corrected by E1 ubiquitin-activating enzyme. Proc. Natl. Acad. Sci. USA **103**(44), 16188–16193 (2006)

T.P. Nouspikel, N. Hyka-Nouspikel, P.C. Hanawalt, Transcription domain-associated repair in human cells. Mol. Cell. Biol. **26**(23), 8722–8730 (2006)

L. Oliver, E. Hue, Q. Sery, A. Lafargue, C. Pecqueur, F. Paris, F.M. Vallette, Differentiation-related response to DNA breaks in human mesenchymal stem cells. Stem Cells **31**(4), 800–807 (2013)

T.R. Ozolins, B.F. Hales, Oxidative stress regulates the expression and activity of transcription factor activator protein-1 in rat conceptus. J. Pharmacol. Exp. Ther. **280**(2), 1085–1093 (1997)

B.F. Pachkowski, K.Z. Guyton, B. Sonawane, DNA repair during in utero development: a review of the current state of knowledge, research needs, and potential application in risk assessment. Mutat. Res. **728**(1–2), 35–46 (2011)

S. Pampfer, C. Streffer, Increased chromosome aberration levels in cells from mouse fetuses after zygote X-irradiation. Int. J. Radiat. Biol. **55**(1), 85–92 (1989)

K.Z. Pan, J.E. Palter, A.N. Rogers, A. Olsen, D. Chen, G.J. Lithgow, P. Kapahi, Inhibition of mRNA translation extends lifespan in Caenorhabditis elegans. Aging Cell **6**(1), 111–119 (2007)

G. Pegoraro, N. Kubben, U. Wickert, H. Gohler, K. Hoffmann, T. Misteli, Ageing-related chromatin defects through loss of the NURD complex. Nat. Cell Biol. **11**(10), 1261–1267 (2009)

G. Pegoraro, T. Misteli, The central role of chromatin maintenance in aging. Aging (Albany NY) **1** (12), 1017–1022 (2009)

P. Ramos-Espinosa, E. Rojas, M. Valverde, Differential DNA damage response to UV and hydrogen peroxide depending of differentiation stage in a neuroblastoma model. Neurotoxicology **33**(5), 1086–1095 (2012)

E. Rass, G. Chandramouly, S. Zha, F.W. Alt, A. Xie, Ataxia telangiectasia mutated (ATM) is dispensable for endonuclease I-SceI-induced homologous recombination in mouse embryonic stem cells. J. Biol. Chem. **288**(10), 7086–7095 (2013)

L.L. Richardson, C. Pedigo, M. Ann Handel, Expression of deoxyribonucleic acid repair enzymes during spermatogenesis in mice. Biol. Reprod. **62**(3), 789–796 (2000)

B. Riis, L. Risom, S. Loft, H.E. Poulsen, Increased rOGG1 expression in regenerating rat liver tissue without a corresponding increase in incision activity. DNA Repair (Amst) **1**(5), 419–424 (2002)

P.M. Rodrigues, P. Grigaravicius, M. Remus, G.R. Cavalheiro, A.L. Gomes, M.R. Martins, L. Frappart, D. Reuss, P.J. McKinnon, A. von Deimling, R.A. Martins, P.O. Frappart, Nbn and atm cooperate in a tissue and developmental stage-specific manner to prevent double strand breaks and apoptosis in developing brain and eye. PLoS ONE **8**(7), e69209 (2013)

S.L. Rulten, K.W. Caldecott, DNA strand break repair and neurodegeneration. DNA Repair (Amst) **12**(8), 558–567 (2013)

N.F.* Sarosh, I.S. Mehta, B.J. Rao*, Spatial arrangement of chromosomes in human interphase nuclei is self-organized by inter-chromosomal systemic couplings. Nat. Sci. Rep. **6**, 36819 (2016). https://doi.org/10.1038/srep36819

P. Scaffidi, T. Misteli, Reversal of the cellular phenotype in the premature aging disease Hutchinson-Gilford progeria syndrome. Nat. Med. **11**(4), 440–445 (2005)

P. Scaffidi, T. Misteli, Lamin A-dependent nuclear defects in human aging. Science **312**(5776), 1059–1063 (2006)

S. Schmetsdorf, E. Arnold, M. Holzer, T. Arendt, U. Gartner, A putative role for cell cycle-related proteins in microtubule-based neuroplasticity. Eur. J. Neurosci. **29**(6), 1096–1107 (2009)

L. Schneider, M. Fumagalli, F. d'Adda di Fagagna, Terminally differentiated astrocytes lack DNA damage response signaling and are radioresistant but retain DNA repair proficiency. Cell Death Differ. **19**(4), 582–591 (2012)

N. Schuler, C.E. Rube, Accumulation of DNA damage-induced chromatin alterations in tissue-specific stem cells: the driving force of aging? PLoS ONE **8**(5), e63932 (2013)

E.I. Schwartz, L.B. Smilenov, M.A. Price, T. Osredkar, R.A. Baker, S. Ghosh, F.D. Shi, T.L. Vollmer, A. Lencinas, D.M. Stearns, M. Gorospe, II. Kruman, Cell cycle activation in postmitotic neurons is essential for DNA repair. Cell Cycle **6**(3), 318–329 (2007)

M.H. Sherman, A.I. Kuraishy, C. Deshpande, J.S. Hong, N.A. Cacalano, R.A. Gatti, J.P. Manis, M.A. Damore, M. Pellegrini, M.A. Teitell, AID-induced genotoxic stress promotes B cell differentiation in the germinal center via ATM and LKB1 signaling. Mol. Cell **39**(6), 873–885 (2010)

D.K. Shumaker, T. Dechat, A. Kohlmaier, S.A. Adam, M.R. Bozovsky, M.R. Erdos, M. Eriksson, A.E. Goldman, S. Khuon, F.S. Collins, T. Jenuwein, R.D. Goldman, Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. Proc. Natl. Acad. Sci. USA **103**(23), 8703–8708 (2006)

M. Simonatto, L. Giordani, F. Marullo, G.C. Minetti, P.L. Puri, L. Latella, Coordination of cell cycle, DNA repair and muscle gene expression in myoblasts exposed to genotoxic stress. Cell Cycle. **10**(14), 2355–2363 (2011)

M. Simonatto, F. Marullo, F. Chiacchiera, A. Musaro, J.Y. Wang, L. Latella, P.L. Puri, DNA damage-activated ABL-MyoD signaling contributes to DNA repair in skeletal myoblasts. Cell Death Differ. **20**(12), 1664–1674 (2013)

P. Sykora, D.M. Wilson 3rd, V.A. Bohr, Base excision repair in the mammalian brain: Implication for age related neurodegeneration. Mech. Ageing Dev. **134**(10), 440–448 (2013a)

P. Sykora, J.L. Yang, L.K. Ferrarelli, J. Tian, T. Tadokoro, A. Kulkarni, L. Weissman, G. Keijzers, D.M. Wilson 3rd, M.P. Mattson, V.A. Bohr, Modulation of DNA base excision repair during neuronal differentiation. Neurobiol. Aging **34**(7), 1717–1727 (2013b)

A. Tedeschi, S. Di Giovanni, The non-apoptotic role of p53 in neuronal biology: enlightening the dark side of the moon. EMBO Rep. **10**(6), 576–583 (2009)

C. Thamrin, G. Stern, U. Frey, Fractals for physicians. Paediatr. Respir. Rev. **11**, 123–131 (2010)

E.D. Tichy, R. Pillai, L. Deng, L. Liang, J. Tischfield, S.J. Schwemberger, G.F. Babcock, P.J. Stambrook, Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. Stem Cells Dev. **19**(11), 1699–1711 (2010)

A. Tomashevski, D.R. Webster, P. Grammas, M. Gorospe, I.I. Kruman, Cyclin-C-dependent cell-cycle entry is required for activation of non-homologous end joining DNA repair in postmitotic neurons. Cell Death Differ. **17**(7), 1189–1198 (2010)

R.K. Vinson, B.F. Hales, Expression of base excision, mismatch, and recombination repair genes in the organogenesis-stage rat conceptus and effects of exposure to a genotoxic teratogen, 4-hydroperoxycyclophosphamide. Teratology **64**(6), 283–291 (2001a)

R.K. Vinson, B.F. Hales, Nucleotide excision repair gene expression in the rat conceptus during organogenesis. Mutat. Res. **486**(2), 113–123 (2001b)

R.K. Vinson, B.F. Hales, DNA repair during organogenesis. Mutat. Res. **509**(1–2), 79–91 (2002)

D. Wells, M.G. Bermudez, N. Steuerwald, A.R. Thornhill, D.L. Walker, H. Malter, J.D. Delhanty, J. Cohen, Expression of genes regulating chromosome segregation, the cell cycle and apoptosis during human preimplantation development. Hum. Reprod. **20**(5), 1339–1348 (2005)

G. Xu, G. Spivak, D.L. Mitchell, T. Mori, J.R. McCarrey, C.A. McMahan, R.B. Walter, P.C. Hanawalt, C.A. Walter, Nucleotide excision repair activity varies among murine spermatogenic cell types. Biol. Reprod. **73**(1), 123–130 (2005)

P. Zheng, R.D. Schramm, K.E. Latham, Developmental regulation and in vitro culture effects on expression of DNA repair and cell cycle checkpoint control genes in rhesus monkey oocytes and embryos. Biol. Reprod. **72**(6), 1359–1369 (2005)

Chapter 29
A Philosophical Perspective on a Metatheory of Biological Evolution

Virginia M. F. G. Chaitin and Gregory J. Chaitin

Metabiology

I would like to introduce you to metabiology, a metatheory of biological evolution. Metabiology is a new area of research proposed by the mathematician Gregory J. Chaitin, initially in 2009. These are his main publications (Chaitin 1987, 2009, 2012a, b, 2013):

- "Evolution of mutating software," *EATCS Bulletin* **97** (February 2009), pp. 157–164
- "Life as evolving software," in H. Zenil, *A Computable Universe*, World Scientific, 2012, pp. 277–302
- *Proving Darwin: Making Biology Mathematical*, Pantheon, New York, 2012

Proving Darwin was first published in English, but you have it now available in other languages: Italian, Japanese, Spanish and Chinese.

Talk Overview

I'm going to try to make this talk as swift as possible. We're going to start with the motivation for metabiology and a conceptual description. Next comes the epistemic critique where we're going to make a connection between metabiology and the neo-darwinian synthesis. We want to know how much metabiology relates to biology as it is seen today.

Talk given by the first author at Nanyang Technological University in Singapore, March 2016. Video in YouTube; search for "chaitin" and "nanyang." This work is based on the first author's 2009 UFRJ doctoral thesis on permeable, pluralistic epistemology and mimetic concept migration, and on her UFRJ post-doc research analyzing metabiology using the interdisciplinary methodology developed in her thesis.

V. M. F. G. Chaitin (✉) · G. J. Chaitin
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
URL: https://independent.academia.edu/VirginiaChaitin

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_29

Why is that? Metabiology is an interdisciplinary area of study. As an epistemologist specializing in interdisciplinary studies, when I examine an interdisciplinary framework, I would like to see if it will actually help you with the individual disciplines that have been incorporated into this framework. So this is one of the methodological criteria that I like to adopt when I address interdisciplinary research.

And in the end I will give some examples of the diversity of this interdisciplinary new area of research and how it connects in unexpected ways to other disciplines.

Why Metabiology?

> "*Can we prove mathematically that evolution through random mutations and natural selection is capable of producing the complexity and diversity of life-forms that populate our planet?*"—G. Chaitin

This was the initial motivation of a mathematician who thought that it would be interesting if we could actually prove that Darwinian evolution can work through random mutations and natural selection. This is the starting point for metabiology.

Then, you may wonder, why is it *meta* biology? I'm going to present two different reasons for calling metabiology a metatheory.

Here is the first reason, that was thought of by the mathematician Gregory J. Chaitin. Metabiology is not about randomly mutating DNA but about randomly mutating software programs. So it is one step removed from biology. In that sense it's a metatheory not a biological theory.

Setting the Stage for Metabiology

Here is the initial setting for metabiology. These are the different disciplines that are involved in creating this new area of research:

- neo-darwinian synthesis
  (Darwin's theory + Mendelian genetics + Population genetics)
- Molecular biology
- Algorithmic information theory
- Evolutionary developmental biology (*Evodevo*)
- Metamathematics
- Digital philosophy
- Computability theory

In particular, I would like to call your attention to one of the items in the list, metamathematics, a somewhat mysterious field dealing with incompleteness, uncomputability and randomness.

Metabiology: What Kind of Math?

We are very familiar with mathematical models in many areas of research. But what I would like to signal from the very beginning is that metabiology uses a different kind of mathematics, post-modern mathematics. In this connection, let me recommend Tasić's book *Mathematics and the Roots of Postmodern Thought* published by Oxford University Press (Tasić 2001), a comprehensive discussion of the transition between the modern and the post-modern mode of thought, including mathematics.

We are not really trying to produce a traditional quantitative mathematical model for biology. Instead we are taking advantage of certain aspects of mathematics that are not widely taught and are not widely known. Among those aspects are the fact that there is logical irreducibility in mathematics.

These are mathematical techniques that are post-Gödelian and post-Turing. Mathematics is not a reductionist area of knowledge. When one is doing mathematical modeling, the steps in a process are usually computable, but one can also use uncomputable steps involving Turing oracles for thinking about phenomena mathematically.

In fact, metabiology postulates organisms having an unbounded appetite for acquiring new information from oracles, resulting in an unbounded drive for creativity.

This constellation of ideas involving novelty, creativity and non-mechanical thought, also includes the idea, very dear to Gregory J. Chaitin, that mathematics is quasi-empirical and that it sometimes progresses by doing experimental math instead of searching for proofs.

Oracles for Uncomputable Steps

So post-modern mathematics deals with uncomputability. The discovery of the **uncomputable** in mathematics, a truly fundamental discovery, is in Turing's famous 1936 paper (Turing 1936), "On computable numbers, with an application to the *Entscheidungsproblem*," which was published in the *Proceedings of the London Mathematical Society*.

Oracles appear in mathematics for the first time in a lesser-known work of Turing's, his 1939 paper (Turing 1939), "Systems of logic based on ordinals," also in the *Proceedings of the London Mathematical Society*.

Why does metabiology need to use oracles?

Oracles bring information from outside the current system. This **new information** is not in the current organism. And this new information cannot be deduced or calculated from the information already in the system.

In particular, we need to consult an oracle in order to be able to decide if a mutated organism is fitter than the original organism. In normal biology, the environment would decide. In our model this is an uncomputable step.

So for the uncomputable steps—so that we don't get stuck when you can't actually compute a result—you have what are called Turing oracles. Oracles are not usually thought of as something you can use in scientific research. They are not usually associated with mathematics. However they do exist in mathematics. They are specifically defined, and for us here in the metabiological realm oracles are especially desirable, because they bring new information into the system that we are working on. Which means that if you take advantage of uncomputability, you can actually feed your system with information from outside. And this is a very important feature of metabiology, which will change the way metabiology communicates with biology, as we shall see later.

This new information from the oracle is intimately connected with the notion of fitness in this metabiological model. Our fitness is going to be less about adaptation, less about competition, and more about the creation of novelty in the biological realm. This is particularly interesting and I would even say useful, if you want to explain the diversity of life-forms. How is it that nature came up with so many life-forms? And creativity is part of the picture. We believe that it is a very important part of the picture that is not usually mentioned.

On Metamathematics

Okay, so a little more on metamathematics. When we're talking about biological creativity, and we want to express that mathematically, what we are doing is taking advantage of the creativity in mathematics to express mathematically the creativity in biology.

So this is I would say a fundamental difference of this mathematical proposal, and it is a different kind of mathematical modeling, therefore.

Emil Post, unlike Gödel and Turing, is not very well remembered for his work (Davis 1994). But in fact he was the thinker who concluded from Gödel's incompleteness and Turing's uncomputability results the very astonishing and difficult to accept fact about mathematical reality that it is actually an essentially creative area of study:

> "*...perhaps the greatest service the present account could render would stem from its stressing of its final conclusion that* ***mathematical thinking is, and must be, essentially creative****.*"—Emil Post, "Absolutely unsolvable problems and relatively undecidable propositions—account of an anticipation," 1941, p. 378 in *Solvability, Provability, Definability: The Collected Works of Emil L. Post*, Birkhäuser, 1994 (Davis 1994).

And we're not talking about creativity for finding out solutions for problems. We're talking about creativity that expands the mathematical realm, expanding mathematical knowledge. So it's a creativity that brings about novelty within the mathematical realm. And **metabiology posits a connection between creativity in mathematics and creativity in biology**.

Initial Conceptual Framework

We will now explicate our initial conceptual framework. We do not have a population in metabiology. We have one single organism:

- one organism = computer program
- random algorithmic mutations

And, of course, since metabiology links mathematics, computer science and information theory with biology, it takes advantage of the well-known analogy between DNA and a computer program. You can even elaborate on that and think that a gene would be like a subroutine in this program:

- DNA ↔ program
- gene ↔ subroutine

So for the metabiological evolutionary model, you have one organism and that is one computer program. This computer program will be subjected to a random mutation. And then you will have a new, mutated organism. Is this organism fitter, or not, than the previous organism? Because we are concerned with evolving these life-forms, these computational life-forms:

- fitter organism calculates bigger number
- sequence of fitter randomly mutated organisms

The criterion that we are using is if this new program calculates a number, the fitter organism will calculate a bigger number. Of course, if the mutated organism-program doesn't calculate a number, then it can't be fitter than the original organism.

But there is a very interesting aspect here due to the incompleteness and uncomputability results that metabiology takes advantage of. When comparing the two organisms, the previous organism and the mutated organism, you will have to bring information from outside the system to actually be able to decide which is fitter.

And at this point you can imagine that metabiology not only asks if the mutated organism is fitter than the previous one, it also asks if new genes are coming into the system due to the mutation, as new subroutines. So this is where the variability of life-forms in metabiology is coming from.

Interdisciplinarity

And now, what is the epistemic theoretical framework we shall use for looking at metabiology? Due to influence from Feyerabend, we propose an epistemology that is most definitely not canonical. This epistemology is **pluralistic** because it embraces different forms of knowledge, different ways of organizing experience, and how you create pictures about the world and deal with the world through these pictures.

More precisely, pluralistic epistemology views knowledge as a combination of **permeable** conceptual frameworks, rationality rules, epistemic goals, methods and techniques.

We are much concerned with the conceptual frameworks of these different forms of looking into the world, of these different worldviews. And we are interested not just with how the concepts vary among the different disciplines and along the history of a single discipline, we want to know what is happening with the methods and the techniques for acquiring knowledge, for testing knowledge, and for advancing knowledge in these different knowledge systems.

In particular, pluralistic epistemology focuses on non-isomorphic or **mimetic conceptual migrations**. This involves new meanings for existing concepts, new concepts altogether, and also migrations of entire conceptual neighborhoods. Another essential feature of pluralistic, permeable epistemology is its concern with exchanging **epistemic goals, methods** and **techniques** between disciplines.

So when you examine an interdisciplinary area of research or an interdisciplinary research program, you want to look not only at the conceptual framework but also at how you organize your concepts and your results. What are the procedures for obtaining your empirical data? What do you consider valid data?

These are different aspects of my critique of metabiology.

Since I was present while metabiology was being developed, I was naturally curious how metabiology would look through these epistemic spectacles. I wanted to understand not only how the conceptual framework changes, but also the reinterpretation of methods, tools and techniques that change along with the framework. Because we are using metabiology to blend mathematics and biology in a way that is quite novel. And I wondered, will it actually work? Can a mathematical proof be relevant to biology?

Fertile Interdisciplinarity

An epistemically fertile interdisciplinary area of study is one in which the original frameworks, research methods and epistemic goals of individual disciplines are **combined** and **recreated** yielding novel and unexpected prospects for knowledge and understanding.

Interdisciplinarity is now widespread in academic research. But I realize, and maybe so do you from your own experience, that there are different levels of interdisciplinarity, and there are different kinds of interdisciplinary studies. Some merely juxtapose different disciplines but don't actually merge them. What I believe is the true richness of interdisciplinary research is that different disciplines should blend into each other so that, ideally, a new discipline emerges. This is altogether different from teamwork in which people with different backgrounds work together, but then go back to their original departments, uncontaminated.

In summary, what is desirable is *fertile interdisciplinarity*, which occurs when different fields infect each other with interesting concepts or methods, yielding either a different approach to the original disciplines, or perhaps even creating an entirely new discipline, which is what we hope that metabiology may do.

Unfolding Conceptual Frameworks

We would now like to discuss the **fertility** of *mimetic* (non-isomorphic) migrations, which occurs when a metaphor or analogy extends to a new **conceptual vicinity** and makes sense in the new context. **Semantic fertility** occurs when the mimetic migration creates a new vocabulary or new meanings for an existing vocabulary. **Epistemic fertility** occurs when the mimetic migration relates to existing questions and poses possible new answers, poses entirely new questions and possible answers, or explores entirely different intellectual paths and **shifts paradigms**.

So when a new conceptual framework for interdisciplinary research is proposed, one should identify if it assigns new meanings to concepts, and also if it achieves *epistemic fertility*. Such interdisciplinary research takes advantage of methodologies, techniques and vocabularies used in the individual disciplines that comprise it. Epistemic fertility may even lead to a paradigm shift, if one is extremely fortunate. We in fact feel, as you will see, that metabiology is a paradigm shift in the making.

Mimetic Conceptual Framework

As we have seen, metabiology combines ideas from many different disciplines. But perhaps, as a first-order approximation, metabiology may be thought of as the result of merging only two fields: the so-called "modern synthesis" or neo-darwinism, and algorithmic information theory.

In the **modern synthesis** or **neo-darwinism** one studies how DNA determines and controls the organism; evolution through random mutations; and survival of the fittest, involving adaptation, competition and differential reproduction rates.

In **algorithmic information theory** one studies computer programs, looks at statistical and mathematical properties of these programs, and deals with the mysterious phenomenon of uncomputability.

But when fields merge, a new vocabulary is required. Initially we considered a single organism, a computer program subjected to random algorithmic mutations, and fitter organisms calculate bigger numbers, resulting in a sequence of increasingly fit randomly mutated organisms.

That was before. Here is our new vocabulary. And this is where the semantic fertility of interdisciplinarity starts to manifest itself.

The computer program will now be called a **metabiological organism**, so we're already creating a new conceptual framework. The process we are now considering is not just darwinian evolution, it's **metabiological creativity**. Information content, a concept that comes from algorithmic information theory, was added to the conceptual vicinity of evolution, and now plays a key role because **evolution = new information content**, giving rise to a sequence of increasingly informed, increasingly sophisticated metabiological life-forms.

And of course—this is a technical but important detail—in metabiology evolution is a **hill-climbing random walk in software space**, so that you don't have stagnation, you are always searching for new information content. Hence **metabiological evolution is open-ended**. Which means that the metabolical evolutionary process is always searching for novelty, for diversification; there cannot be any stagnation.

This is not what happens in the biosphere, where adaptation and stagnation definitely occur. However, this is not what metabiology is about. Metabiology attempts to explain the diversity, richness and sophistication of life-forms. How come the biological realm is so creative? This is why metabiology posits a strictly hill-climbing evolution.

Two Definitions of Metabiology

At this point we should delimit the scope of metabiology more precisely by presenting two different ways G. Chaitin defines the subject:

- "*Metabiology is a field parallel to biology dealing with the random evolution of artificial software (programs) rather than natural software (DNA).*"
- "*The goal of metabiology is to find the simplest pythagorean life-form that has hereditary information and evolves according to a fitness criterion.*"

As was already pointed out, that in metabiology software is evolving, not DNA, accounts for the "meta" in metabiology. We are **making biology mathematical at a meta-level**.

Nevertheless, metabiology retains biological principles. It proposes a pythagorean life-form that has hereditary information which evolves according to a fitness criterion. The mathematical fitness criterion is to calculate a bigger number, which seems rather strange. But we will see how it relates metaphorically to what happens in biology further along.

Fertility at the Conceptual Level

So we have already established that we did have some interesting initial results in metabiology at the conceptual level because we're not dealing with the original concepts anymore. Now we are thinking of life as evolving software, more precisely:

- *Life as randomly mutating and evolving software*

Also, we can think that Nature is in some way programming our metabiological organism, without a programmer though, because we don't need a programmer for metabiological evolution to take place:

- *Nature is programming without a programmer!*

And this relates to **digital philosophy** (Pagallo 2005; Longo and Vaccaro 2013), an unexpected item in the list in section "Setting the Stage for Metabiology" where we enumerate those fields feeding into metabiology. Digital philosophy is actually what is proposing a new ontological framework, which is looking at the universe as a computer, a big computer that is performing a computation in which the state of the universe at the next moment is calculated as a result of the state of the universe at the previous moment. And metabiology fits within this framework. So it's looking at evolution in a very procedural way.

Here we are also changing our ontological framework from what used to be the important aspects, energy, metabolism and continuous math, to an ontological substrate of information, algorithms and discrete math. And instead of searching for simulations of evolution, we are searching for mathematical proofs about the efficacy of evolution.

In summary, we have the following migrations (translations between conceptual frameworks):

- *Basic substance*: from biochemicals to information
- *Basic process*: from metabolism to algorithms
- *Language*: from continuous pre-Turing math to discrete post-Turing math
- *Methods*: from simulations to mathematical proofs.

Exchanging Reasoning Procedures and Epistemic Goals

Here is a general **validity criterion** for exchanging **methods** and **techniques** between disciplines. It's valid when the research results driven from this exchange of methods and techniques lead to:

- interpretations that make sense on the conceptual level
- meaningful questions and/or answers, or
- new research problems and/or possible solutions

Now here is the challenge that metabiology faces: Once we have some mathematical proofs and an appropriate conceptual setting, and we try to interpret it biologically, does it make any sense? Will it bring about new questions for biology? Or answers for questions that biology hasn't answered yet? Different questions or different answers? And eventually new research problems.

Fertility at the Epistemic Level

Our first cut at discussing these issues is to realize that metabiology changes an epistemic goal. Instead of simulations, one wants proofs. And this means that one is not looking for shifts in the frequencies of genes. What is important is to know where do the new genes come from.

So in metabiology we have a change in epistemic goals, methods and techniques from explaining gene frequency shifts given selective pressure *using mathematics instrumentally* (**simulations**), to explaining biological creativity *using mathematics to generate and express novelty* (**proofs**). In particular, we want to know:

- Where do **new genes** come from?
- Why the increase in **conceptual complexity**?
- Why new **information content**?

So this is mathematically interesting, that instead of using math instrumentally, only for simulations and calculations, we are working mathematically on a more conceptual level where in a proof you try to understand how ideas can fall into place and give a certain result, in a mathematical setting. Furthermore, we are now explaining biological creativity through mathematical creativity and we are associating this creativity with conceptual complexity and information content, which are concepts that come from algorithmic information theory—a substantial and unexpected confluence of ideas.

How does this constellation of ideas, expressed with the kind of precision that only mathematics can achieve, apply in biology?

But, before discussing this, we need to explain better why this change in epistemic goal was necessary in metabiology.

About Creativity in Biology

Here is a very interesting quote that states why is it that metabiology does not work with simulations:

> "*It's a theory that does not give you a way to simulate creativity, but it gives you a way to prove theorems about it. Creativity is by definition something we* ***don't know*** *how to do.*"—G. Chaitin (Jałochowski 2015)

We don't want to simulate creativity. After all, creativity is something that you don't know how to do before you do it. If one wants to be creative, if it's real creativity, it does not follow a recipe. You will only realize what you have done, how you did it, afterwards. It is not like, "I want to be creative today. I'll wake up, I'll do this, that and the other, and then I will be very creative." That's not how creativity happens. So this is why here in this setting we do not think about simulating creativity.

About Creativity in Randomness

And there is also another very interesting detail. The metabiological model of evolution uses *random* algorithmic mutations. We are well aware that some current biological theories attribute less importance to randomness in the evolutionary process, and instead emphasize those aspects of evolution that are influenced by the environment, and this is directed, not random. Indeed, randomness has a "bad reputation," in the sense that it is frequently associated with lack of purpose, lack of intelligence, and lack of meaning. If this were what randomness is, positing that something is random is not really worth much when one is searching for an explanation.

However, if you think about something that is random in the sense that it is not recognizable by any pattern that one knows, then this means that you are leaping into real novelty. The point we would like to make is that randomness can be a way to achieve something that has not been thought of yet or perhaps did not even come into being before this creative act.

In other words, we are thinking of randomness as a **leap into uncharted territory** or even into previously non-existent territory, rather than as lack of purpose, intelligence, and meaning.

In effect, in metabiology, the act of mapping creates new territory. The world as seen through metabiology is open-ended, in *statu nascendi*, not closed.

Metabiology in a Nutshell

Having done the requisite preparatory analysis, we are now in a position to define metabiology like this:

> "*Metabiology is a mathematical expression of the interaction between randomness and uncomputability giving rise to novelty, increasing conceptual complexity in the form of new information content.*"—V. Chaitin

I like to say that it's a mathematical expression, not a mathematical model, of the interaction between **randomness**—that's something that brings a really creative, novel possibility—and **uncomputability**—which brings the oracular aspects—giving rise to novelty. And increasing the **conceptual complexity** in the form of new **information content**.

The above quoted definition is unfortunately a bit long, and it brings together concepts from radically different areas. It's a very interdisciplinary sentence. For it to make sense we have to be aware that each one of these concepts is now functioning in a new territory, not in the original field from which it emerged. And this is the richness of thinking in such an interdisciplinary manner.

Metabiology and the History of Ideas

We've talked about randomness as a source of creativity, rather than lack of purpose and meaning. Indeed, Darwin was severely criticized for replacing God, a teleological principle, with dice.

This puts us squarely in the intellectual cross-fire between two major intellectual traditions. Not a problem! Our permeable pluralistic stance welcomes intellectual hybrid vigor.

Here, very succinctly, are the two waring traditions, which in our view actually complement each other:

- Randomness and atoms in the void: **Nature does not have an** a priori **purpose**—Democritus, Lucretius, Laplace, Darwin, Boltzmann, Dawkins…

  analysis—reductionism—statistical laws—mechanisms

- Holism, Gaia theory, teleology, *Romantische Naturphilosophie*: **Nature is intelligent and does have a purpose** — Aristotle, Goethe, Lamarck, Wallace, Teilhard de Chardin…

  synthesis—emergence—self-organization—organisms

Metabiology absorbed and combined elements from both of these intellectual traditions or modes of thought. These could only be *mathematically* absorbed and *mathematically* combined, because metabiology is expressed in post-modern math, itself a hybrid of mechanical, closed algorithmic math with the open-ended creativity of incompleteness, undecidability, uncomputability and irreducibility.

Second Reason for the *Meta* in Metabiology

We can now give the previously promised additional reason for the "meta" in metabiology, epistemically grounded on the fact that metabiology does not model specific biological phenomena, it mathematically expresses **general principles** that guide the process of darwinian evolution, namely, random mutations + selection. For this reason, metabiology can relate to **different biological paradigms** that involve these general principles.

As we will see, due to its use of oracles, metabiology manages to simultaneously combine darwinian and lamarckian elements (epigenetics). Metabiology was initially inspired by neo-darwinism, but later escaped its confines. It does so on three levels:

- *On the initial conceptual level*: The emphasis is on mutations aimed at **creativity**, not on the current model of biological mechanism built out of mutations, competition, adaptation and perpetuation—the survival-of-the-fittest pantheon.
- *On the mathematical model*: Metabiology employs **algorithmic mutations**, not point mutations; it models **open-ended** evolution—not adaptation nor stagnation.

- *On the drive of living organisms*: Metabiological organisms are not driven by the urge to survive and perpetuate, they are driven by creative, **innovative experimentation**.

Now it is time to begin a dialogue between metabiology and biology. From this point on, we're going to systematically examine and enumerate one by one the varied facets of the epistemic fertility of metabiology. These will be collected in tables.

Conceptual Frameworks and Neighborhoods

Here are some of the important concepts in metabiology, and each of them comes with an associated conceptual framework plus a neighborhood of related concepts:

- evolution
- algorithmic information
- uncomputable steps—oracle
- environment
- selection
- mutation distance
- fitness

These are all examples of non-isomorphic or *mimetic* concept migration, because their meanings change in the new metabiological context. These reassignments of meanings are covered in Tables 29.1, 29.2, 29.3 and 29.4.

Table 29.1 Darwinian perpetuation—metabiological innovation

Biology	Metabiology
Natural software: DNA–RNA (base 4: AGCT)	Metabiological software: computer programs (base 2: 0, 1)
Organism results from biological processes involving DNA–RNA and environment	Metabiological organism is the software itself
Darwinian challenge: competition, survival, adaptation to a changing environment + passing your genes to the next generation	**Metabiological challenge**: solving a mathematical problem that requires creative, uncomputable steps
"Selfish **genes**"—reproduction and perpetuation	**Process** of random algorithmic mutations—**no perpetuation**

Table 29.2 Selection—fitness—evolution

Metabiology	Biology
Single organism	Populations of organisms
Unit of selection: current organism	Unit of selection: “selfish” gene, organism, group or species
Metabiological selection → evolution increasing **information content** of algorithmic life-form	Natural selection → evolution increasing **sophistication** of biological life-forms
Fitness grows with **conceptual complexity**: it is a global, fixed measure (computational capacity)	**Fitness** may or may not grow with **complexification** of life-form: it is a local, variable, contextual measure
Metabiological evolution: random algorithmic mutations + strictly hill-climbing random walk in software space	Evolution: random mutations + adaptation to environment + Red Queen hypothesis (Leigh Van Valen)
Metabiological creativity: measured in **bits of information**	Biological creativity: measured by the **diversity of life-forms**

Table 29.3 Algorithmic mutation—mutation distance

Metabiology	Biology
Random **algorithmic** mutations	Random **point** mutations
Mutation distance = **conceptual** complexity of the mutation—the size of the smallest program that carries out the mutation (correspondence with **biological complexity** is not rigorously defined)	Mutation distance = number of **base-pairs** changed
“Time” for a given mutation distance to occur is measured in number of tries (correspondence to chronological time is not defined)	Mutation time measured chronologically is inferred from empirical data (fossils)
Speciation: occurs when mutation distance crosses an arbitrary threshold, i.e., a sufficiently great jump in the conceptual complexity of the genome	**Speciation**: happens by accumulating random point mutations driven by environmental change and/or isolation

Table 29.4 Mutation—oracle—environment

Metabiology	Biology
Oracle is responsible for “metabiological **selection**”: checking viability and selecting the mutated organism	**Environment** is responsible for darwinian “natural **selection**”
Random algorithmic mutations	Canonical neo-darwinism: random **indel/point mutations**
Random algorithmic mutations are checked for their viability by the **oracle**	Epigenetics: mutations are not random, depend on the **environment**
Single organism—vertical gene transfer	Horizontal gene transfer—bacteria
Algorithmic mutation	Retrovirus insertion

Evolution as Increasing Information Content

What is evolution, after all? What is the defining feature of evolution? A very substantial epistemically fertile contribution of metabiology is that it unearths this fundamental question. In neo-darwinism stagnation is a possible outcome of evolution. In metabiology, it is not.

So the real issue in understanding neo-darwinism mathematically—which is what metabiology is all about—is the tension between innovation and stagnation:

- **Evolution**: How does the increase in **information content** measured in bits relate to the **increasing sophistication** of life-forms?
- **Stagnation**: Occurs if/when perfect adaptation is achieved and subsequently there are no major changes in the environment.

There are perfectly adapted, ancient life-forms: e.g., the horse-shoe crab. These correspond to stable ecological niches. And please note that computer simulations of evolution like Tierra and Avida all eventually stagnate.

On the other hand, there is the Leigh Van Valen *Red Queen hypothesis* discussed in Chaitin (2012b), according to which evolution is an arms race with other life-forms, an example of which is the rapid development of bacterial immunity to antibiotics. In other words, as the Red Queen says in Lewis Carroll's *Through the Looking Glass*, you have to run as fast as you can to stay in the same place.

Which is the most important driver of evolution, environmental changes such as drifting tectonic plates, or the arms race with ones predators and prey? Well, normally the changing environment gets most of the credit, but the genetic arms race seems to be the explanation for sexuality, a nearly universal feature of contemporary organisms that is otherwise rather difficult to justify theoretically, as was admitted by John Maynard Smith and his disciple Richard Dawkins.

In metabiology evolution is unending; there can be no stagnation. And evolution is measured by the increasing information content of organisms, which may lead to new explanations for their increasing sophistication and diversity. Turning now from metabiology to neo-darwinism, our metabiological analysis of evolution suggests the following important question: **Can biological creativity be measured in bits of information content?**

About Exchanging Epistemic Criteria

In mathematics what counts is the **beauty** of the proofs, there really is an aesthetic criterion for truth. And this is reflected in fields of science that have been mathematized. For example, James Clerk Maxwell made his equations more symmetrical and discovered electromagnetic waves. And we may be facing a similar situation with the algorithmic mutations that are used in metabiology.

The preliminary version of metabiology (Chaitin 2009), following neo-darwinism too closely, employed point mutations. The math was frightfully ugly.

The current version of metabiology (Chaitin 2012a) uses algorithmic mutations, which are powerful, global, **arbitrary algorithmic transformations of an organism**. This seems rather unbiological, but the math is extremely pretty.

Therefore, an important *Question*: How does this mathematical suggestion from metabiology relate to mutations in biology? How powerful can real mutational mechanisms actually be?

Algorithmic Mutations and Biology

Algorithmic mutations are very powerful and they include gradual localized changes like point mutations as well as major global evolutionary leaps.

For an example of the latter, consider the transition from **uni-cellular to multi-cellular** life-forms. This amounts to changing a main program into a subroutine that is called many times, which is not a very big change in a piece of software.

Could algorithmic mutations help us to understand the mysterious Cambrian explosion? Could they possibly explain why the intermediate forms of organisms are frequently missing in the fossil record? Could it be that the major transitions in evolution are a result of algorithmic rather than point mutations?

Perhaps algorithmic mutations can do more faster, in fewer generations, and in less chronological time, than point mutations can.

But are they real, or only a mathematical fantasy? Perhaps they are actually real. In fact, can algorithmic mutations be related to retroviruses which insert themselves into the genome?

How powerful random mutations can actually be is an interesting question to consider in order to decide whether or not metabiology sheds any light on real biology, or as G. Chaitin would say, whether it remains an ideal in the platonic world of ideas but unrealized in the messy real world.

Mutation Distance and Biology

In metabiology the *mutation distance* between organisms A and B is defined to be the size in bits of the smallest mutation M that algorithmically transforms A into B. That is, it is the size in bits of the smallest computer program M such that $B = M(A)$. In other words, M is the simplest program that takes A as input and produces B as output.

Equivalently, the mutation distance may be defined as $-\log_2$ of the probability that a randomly chosen mutation M converts A into B, because in metabiology a K-bit mutation M is tried with probability 2^{-K}.

The mutation distance is also equivalent to the *relative information content* of B given A, an important concept in algorithmic information theory that has an almost immediate biological reinterpretation.

Please note that this is a directed distance, it is not symmetrical. The distance from A to B may be very different from the distance from B to A. For instance, if B is contained in A, it may be very easy to eliminate the rest of A to get B, but much harder to expand B into A.

How is the distance between two genomes normally measured? Well, by a symmetric distance metric, the *Ulam distance*, which is roughly the number of point mutations or indels it takes to go from one to the other, i.e., the number of bases that need to be changed and/or inserted or deleted.

So mutation distance is measured in terms of differences in the algorithmic information content of the sequence of bits in metabiological software organisms, while Ulam distance measures the similarities and differences in the sequence of bases in actual DNA.

Taking a hint from metabiology, should we measure the information content and conceptual complexity (Chaitin 2015a) of DNA? This could lead to an additional criterion for classification of life-forms: mutation distance not Ulam distance, not number of bases changed.

Such distance measures are useful in *cladistics*, which deduces evolutionary trees from genome base sequences, based on the assumption that related genomes cannot be far apart.

Oracle Acts as Environment for Organisms and Mutations

The normal information-theoretic view of evolution is that it increases the information in the genome about the environment—that is, the mutual information between the genome and the environment—until the organisms are fully adapted to their environment, at which point evolution stagnates until there are changes in the environment.

However, metabiology focuses on unending biological creativity and on the consequences of viewing DNA as software, which leads to a completely different information-theoretic perspective on evolution, but still incorporating information from the environment.

How do we extract information from the environment in metabiology?

Given two organisms, because of Turing's halting problem we need to use an oracle to decide which organism is fitter, because we have to eliminate organisms that never halt before we can compare the numbers that they calculate to see which program calculates the bigger number. This is the fitter organism, and in metabiology it is the basis for the next evolutionary step.

Similarly, before applying a random algorithmic mutation to an organism, we need to use an oracle to eliminate mutations that never halt and therefore never produce a mutated organism.

So when selecting a metabiological organism, the oracle plays the role of the environment as in canonical neo-darwinism. And when selecting an algorithmic mutation, the oracle plays the role of the environment as in epigenetics.

In this manner metabiology simultaneously manages to relate to different biological paradigms. As we have said, metabiology is a meta-theory.

Other Settings for Metabiology

Other settings for metabiology have been and are being explored by G. Chaitin's talented student Felipe Abrahão.

In particular, he has come up with computable versions of metabiology in which there are **no oracles**, a remarkable piece of work which earned him a UFRJ doctoral degree and which has been published in a World Scientific volume edited by Mark Burgin and Cristian Calude (Abrahão 2016).

Although the oracle has been eliminated, in its place there is a hierarchy of degrees of computational power, in which the environment is higher in the hierarchy than the individual organisms. In this sense, the computational version of metabiology retains the essential feature of bringing in new information from outside the current system of life-forms, things which they cannot compute by themselves.

He is now looking at the evolution of a population of metabiological organisms interacting through a network, leading to the emergence of new properties of this algorithmic population network (Abrahão et al. 2017).

Metabiology and Digital Philosophy—What Kind of Nature-Computer?

Some final considerations.

Recall that digital philosophy (Pagallo 2005; Longo and Vaccaro 2013) attempts to view the universe as a giant computation.

The metabiological models of Chaitin and Abrahão are certainly in the spirit of digital philosophy, because they are discrete software models, but they raise some important issues.

Is every natural process a computation in a computable universe or does/may the world as a computation view involve uncomputable steps? In other words, in this universe are there any oracles?

Could there in fact be different levels (hierarchies) of computability and uncomputability in the digital philosophy world as a function of the conceptual complexity of the phenomena at each level?

Perhaps digital philosophy needs to evolve too.

Concluding Remarks

We believe we have shown that metabiology is both semantically and epistemically fertile, and raises many issues for future research.

Here is one of these issues.

Metabiology is most definitely not about "selfish genes" or "survival of the fittest." Nobody is selfish, nobody survives in metabiology, it's all about creativity. Metabiological organisms do not want to adapt to their environment, they want to be creative.

To the extent that metabiology provides a theoretical foundation or *grund* for biology, it says something about the human self-image, and it informs the post-humanism debate. For if we can think of ourselves as creative beings instead of as bags of selfish genes struggling to perpetuate themselves (Chaitin 2015b; Chaitin et al. 2014) this makes the human being more significant and resonates with our perception that life is meaningful.

Certainly metabiology is only an embryonic theory and much remains to be done. However, we hope you will agree with us that these ideas are beginning to show some promise. Thank you for giving me a chance to explain all of this. Thank you very much for your kind attention!

References

F.S. Abrahão, The 'paradox' of computability and a recursive relative version of the Busy Beaver function, in *Information and Complexity*, ed. by M. Burgin, C.S. Calude (World Scientific, Singapore, 2016), pp. 3–15

F.S. Abrahão, K. Wehmuth, A. Ziviani, Algorithmic networks: central time to trigger expected emergent open-endedness (2017), https://arxiv.org/abs/1708.09149

G.J. Chaitin, *Algorithmic Information Theory* (Cambridge University Press, UK, 1987)

G.J. Chaitin, Evolution of mutating software. EATCS Bull. **97**, 157–164 (2009)

G.J. Chaitin, Life as evolving software, in *A Computable Universe*, ed. by H. Zenil (World Scientific, Singapore, 2012a), pp. 277–302

G.J. Chaitin, *Proving Darwin: Making Biology Mathematical* (Pantheon, New York, 2012b)

G.J. Chaitin, From Turing to metabiology and life as evolving software, in *Alan Turing: His Work and Impact*, ed. by S.B. Cooper, J. van Leeuwin (Elsevier, Waltham, MA, 2013), pp. 763–764

G.J. Chaitin, V.M.F.G. Chaitin, F.S. Abrahão, Metabiología: los orígenes de la creatividad biológica, in *Investigación y Ciencia*, Jan 2014, pp. 74–80

G.J. Chaitin, Conceptual complexity and algorithmic information. La Nuova Critica **61–62**, 9–28 (2015a)

V.M.F.G. Chaitin, Metabiology, interdisciplinarity and the human self-image. La Nuova Critica **61–62**, 115–124 (2015b)

M. Davis, *Solvability, Provability, Definability: The Collected Works of Emil L. Post* (Birkhäuser, Boston, 1994)

K. Jałochowski, *Gregory and Virginia Chaitin: Against Method*, one-hour TV program that was broadcast in Poland in 2015 in which Gregory and Virginia Chaitin discuss creativity at their philosophical retreat on the tropical island of Paquetá. Available in a DVD and at https://alexanderstreet.com

G.O. Longo, A. Vaccaro, *Bit Bang. La nascita della filosofia digitale* (Apogeo Education, Milano, 2013)

U. Pagallo, *Introduzione alla filosofia digitale Da Leibniz a Chaitin* (Giappichelli, Torino, 2005)

V. Tasić, *Mathematics and the Roots of Postmodern Thought* (Oxford University Press, New York, 2001)

A.M. Turing, On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. **42**, 230–265 (1936)

A.M. Turing, Systems of logic based on ordinals. Proc. London Math. Soc. **45**, 161–228 (1939)

Chapter 30
On How Epistemology and Ontology Converge Through Evolution: The Applied Evolutionary Epistemological Approach

Nathalie Gontier

Outline

Philosophy traditionally distinguishes epistemology (the map) from ontology (the territory). Epistemologies provide knowledge on the ontological state of certain aspects of the world. Cosmologies are epistemological frameworks that concentrate on the nature of matter, space, and time. Traditionally, matter, space, and time are made intelligible through hierarchy theories that describe the ontological layeredness of the cosmos; and causality theories that render mechanical explanations for this layeredness (sections 1-2). In classic cosmologies, the map and the territory are considered different from one another. Ancient scholars maintain realist positions on how their maps reference the world but such a first philosophy is currently refuted (3). Socio-anthropological schools question any linkage between the map and the territory, and understand epistemology as an outcome of sociocultural practices, while traditional evolutionary epistemological schools maintain hypothetical realist positions. By adhering to adaptationist and Neodarwinian views on evolution, organisms are considered hypothetical theories on the outer world (4). Here, we go further by demonstrating that organisms are not just theories about the world but spatiotemporally real entities (5). Organisms evolve knowledge and reproduce it into their offspring, and through processes such as symbiosis and niche construction, they acquire and extend knowledge onto other organisms and onto their niches (6). Life builds realities and it enables for a realist position where the evolving map equals the evolving territory. We revise traditional evolutionary epistemology accordingly (7). The conclusion is that truth and reality are spa-

N. Gontier (✉)
Applied Evolutionary Epistemology Lab, Centre for Philosophy of Science,
University of Lisbon, Lisbon, Portugal
e-mail: nlgontier@fc.ul.pt
URL: http://appeel.fc.ul.pt

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_30

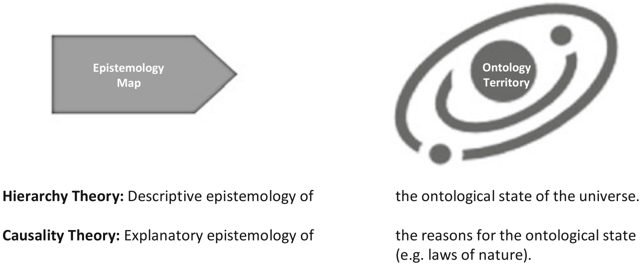
tiotemporally bounded and prone to change in congruence with the organisms that build it and the niches they construct (8).

The Map (Epistemology) and the Territory (Ontology)

Philosophy is traditionally divided into two subdisciplines, *ontology* or the study of existence (that what is), and *epistemology* or the study of knowledge (how we humans come to know that what is) (Ferrier 1854: 44–46). Following the metaphors of this anthology, epistemology provides a *map,* or a *theoretical* or *methodological* means to conceptualize or draw the map of the *territory,* which traditionally refers to the cosmos and all the entities it contains. The entities that exist and the processes that unfold between them are the *object of knowledge*. They are what is being mapped or investigated epistemologically (Fig. 30.1).

Epistemological frameworks on the cosmos underlie the formation of *cosmologies* which are philosophical, religious, ideological or scientific *worldviews* on the nature of *matter*, *space* and *time* (Gontier 2011, 2016b). Cosmologies are illustrated in *cosmographies* which are descriptive and sometimes explanatory *diagrams* or maps that visualize how (aspects of) matter, space and time arrange in the cosmos. Classic examples include Ancient Middle and Far Eastern Wheels of Time or Chains of Being or Medieval Scales of Nature (Barsanti 1992; Lovejoy 1936; Gontier 2011) (Fig. 30.2).

Modern cosmographies include scientific diagrams of entities existing on a *micro-scale* (atomic particles, chemical elements, RNA and DNA molecules, or amino acids); *meso-scale* (trees, webs and networks of life, Fig. 30.3); and *macro-scale* (diagrams of the solar system or the universe).



Hierarchy Theory: Descriptive epistemology of the ontological state of the universe.

Causality Theory: Explanatory epistemology of the reasons for the ontological state (e.g. laws of nature).

Fig. 30.1 Schematic of the classic ontology/epistemology divide

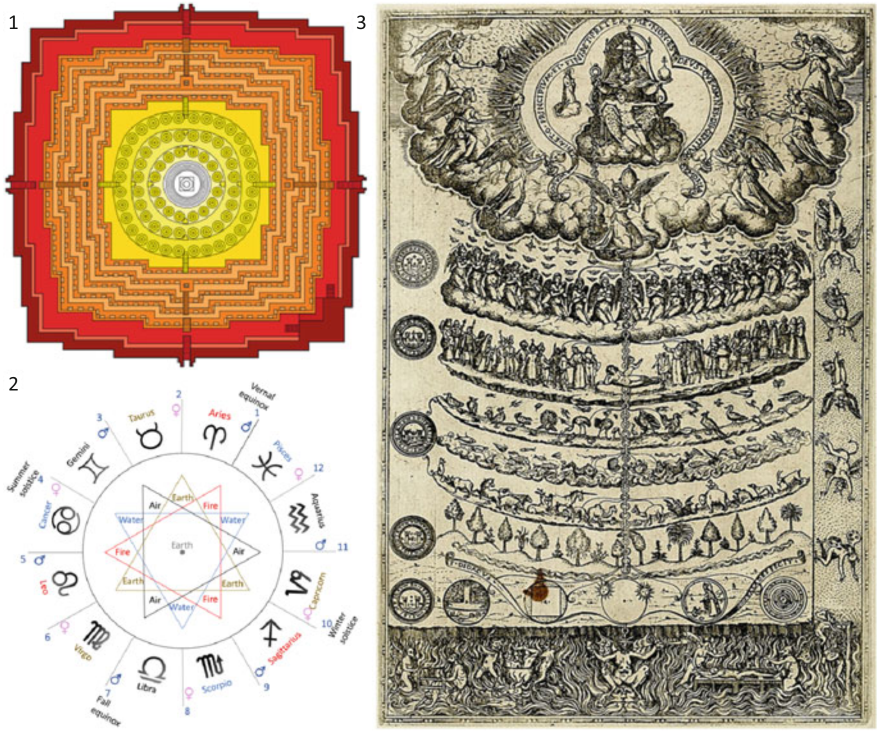
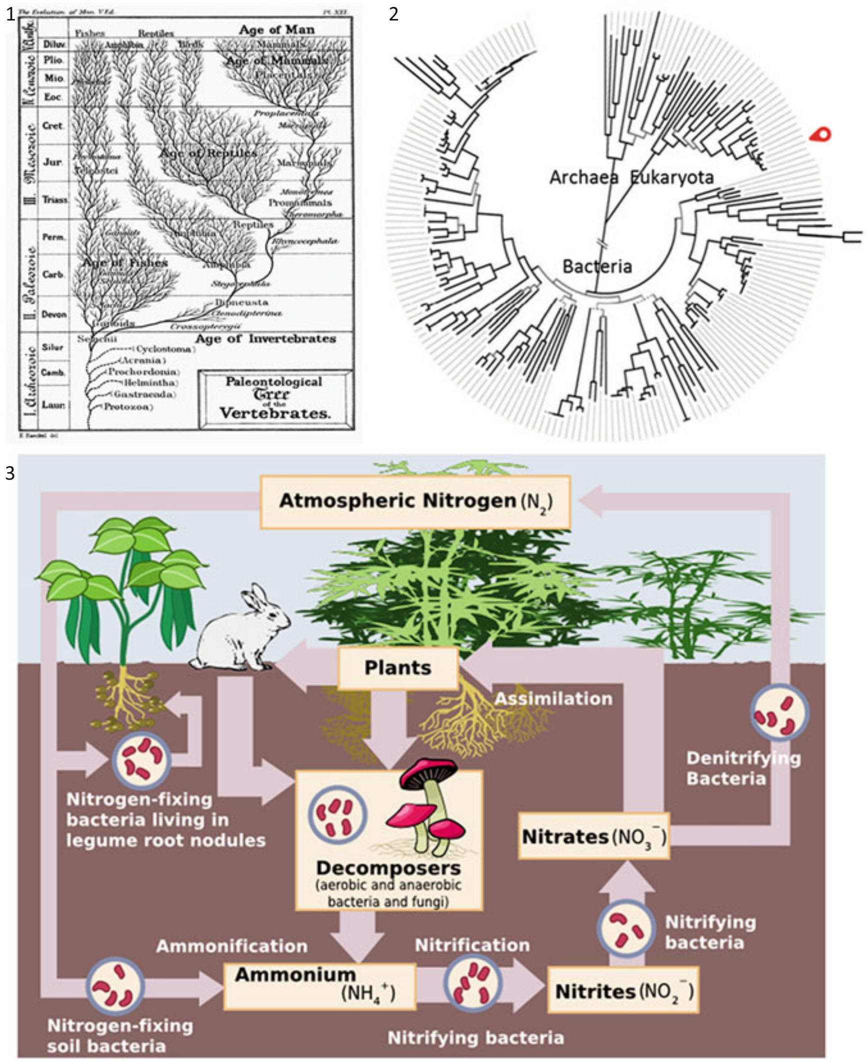


Fig. 30.2 Cosmographies of ancient Far Eastern (1) Middle Eastern (2) and Judeo-Christian cosmologies (3) that demonstrate the hierarchical nature and the underlying causes of the cosmos. (1) Representation of the floor plan of the Buddhist Borobudur temple, located in Indonesia, in the form of a *mandala* or wheel of time. The yellow inner circles represent the realm of formlessness (chaos), the orange middle layers represent the realm of form (the permanent), and the red outer layers represent the realm of desire (the temporal world). (2) The tropical zodiac. It represents a chain of being (the constellations are presumed to be chained animals or gods); a wheel of time (because the zodiac provides a calendar of the Platonic great year and a 360-day year); and a causal explanation for the rotation of the star signs around a geocentric earth (based upon the four elements). (3) A Judeo-Christian reinterpretation of Aristotle's chain of being that was based upon his three-soul theory and determined by the returning cycle of coming (generation) and becoming (decay). For Christians, the chain forms a single and unilinear ladder or stairway to heaven, going from the least to the most perfect beings. The level of perfection is "measured" by the distance that exists between beings and the deity that resides in heaven. The strands of the ladder go from plants over land, water and air animals, to humans and saints. The Christian deity stands above creation and outside of matter, space and time, and is surrounded by angels. Underneath the ladder, we find the underworld that is ruled over by the devil. On the right, we see some falling angels on their way to hell (Credits: (1) Image by David1010, made available on Wikipedia, https://upload.wikimedia.org/wikipedia/commons/8/8b/Borobudur_Mandala_ka.svg; (2) own work; (3) Image by Diego Valadés for *Retorica Christiana* (1579: 218), digitalized by the Getty Research Institute and available under creative commons at https://archive.org/details/rhetoricachristi00vala)



Hierarchy and Causality Theory

Cosmologies and cosmographies associate with *ontological hierarchy theories* that *describe* how matter can be *classified* in space and over time and with *metaphysical* or *ontological causality theories* that explain the *reasons* for this hierarchical order (Gontier 2015b, 2016b).

◀**Fig. 30.3** Scientific diagrams that represent aspects of the living world. (1) Haeckel's (1874) paleontological tree of vertebrates, set in quadrant I of the Cartesian coordinate system (the left columns represent the y-axis marking time, the right column represent the x-axis delineating space). The image provides a chronology of when fishes, reptiles and mammals first originate in the geological time scale. Haeckel's diagram depicts common descent of vertebrates from invertebrates, extinction (the end of lineages), and speciation (the ramification of lineages that mark the rise of new species). These processes are explained by natural selection theory. (Credits: The image comes from the 1879 English translation of the work, and is made available under a creative commons license at https://en.wikipedia.org/wiki/Timeline_of_human_evolution#/media/File:Age-of-Man-wiki.jpg). (2) Adoption of Bork & co-workers' tree of life (Ciccarelli et al. 2006). This unrooted (unhistorical) tree depicts the evolutionary distance between 191 extant species whose whole genomes have been sequenced. The distance is measured by comparing genetic divergence of 31 genes held in common by all these species, and that are involved in the translation of the genetic code. The diagram represents the *most likely* phylogenetic relationship that exists between the species. The tree demonstrates that eukaryotes (multicellular life forms) are more closely related to Archaea than to Bacteria. Archaea and Bacteria are both prokaryotes (unicellular organisms), but they are genetically distinct from one another making some scholars doubt they share a single common ancestor. The red dot on the right marks the location of our species (*Homo sapiens*) on the tree. When making their diagrams, neither Haeckel nor Bork and colleagues took horizontal gene transfer, hybridization, or symbiosis into account, which are processes that can also cause genetic divergence or convergence between species. (3) A cyclic network diagram that demonstrates the important role bacteria, fungi, animals and plants play in the earth's nitrogen cycle. The processes are explained by ecology and symbiosis theory that detail how distinct organisms interact amongst themselves and with the abiotic environment. (Credits: Image by Johann Dréo and made available on Wikipedia under a creative commons license at https://en.wikipedia.org/wiki/Nitrogen_cycle#/media/File:Nitrogen_Cycle.svg.)

Hierarchy Theories

Hierarchy theories provide *descriptions* of the ontological state of the universe (Fig. 30.1). The classic Greek hierarchy, for example, divides the cosmos into an embedded micro-, meso-, and macrocosmos. Other examples, that also root our current division of the sciences, include Hutton and Spencer's distinction between the inorganic (physico-chemical), organic (biological) and superorganic (the sociocultural, ecological and universal); or Julian Huxley's division of the world into a physical, biological and psychosocial level (Gontier 2015b). The fact that different ontological hierarchy theories exist demonstrates that any claim made on ontology remains an epistemological endeavor.

Different cosmologies often apply *different classificatory principles* to build the ontological hierarchy and the criteria used are a means to separate, compare and understand different cosmologies. Far and Middle Eastern cosmologies (Fig. 30.2 (1)) classify the world into *realms* and differentiate between chaos (that what has no form), the permanent (what has a lasting from), and the temporary (what has a form that will generate and decay). Ancient Greeks continue these ideas and their classifications rest on the conjectured *soul* entities have. Inanimate matter has no soul, plants have a vegetative one, animals a sensitive (mobile) one, and humans have a rational soul (Barnes 1984). Judeo-Christian scholars continue Greek classification and build scales of nature that hinge on the presumed level of *perfection* entities have, which is "measured" by how close or distant they are from their deity (Fig. 30.2(2)). 19th century natural history scholars classify entities chronologically, based upon their first appearance in (calendrical) time and space (geographical and geological location). From these chronologies they derive notions of (ontological levels or *amounts* of) *progress* and *complexity*, which are used as additional criteria to classify and understand the order in the world (Fig. 30.3(1)).

Today, biologists classify species by their *level* of *evolutionary relatedness* which is measured by the *amount* of *genetic distance* that exists between organisms. One level down the hierarchy, chemists continue to classify matter based upon the elements that make them up; and another level down, quantum physicists investigate the subatomic level where time and space as we know it dissolve. Biologists and physicists also continue to use *complexity* criteria, as well as additional criteria of *optimality*, *likelihood* and *parsimony* (economy). These are quantitative measurements that enable an examination of how "probable" their cosmologies and cosmographies on the ontological state of the universe are (Fig. 30.3(3)). Note that measuring in terms of optimality, likelihood and parsimony involves a switch from "certain," "real," and "true" to *uncertainty* on how the map links to the territory.

Causality Theories

While hierarchy theories attempt to provide *descriptions* of the ontological state of the universe, causal theories attempt to provide *explanations* for it (Fig. 30.1). *Metaphysics* is a term used synonymously with ontology. It was introduced by Latin scholars to refer to a series of texts written by Aristotle that became classified after his work on Physics (Barnes 1984). Because these works discuss the presumed reasons for the underlying order of the physical world, what Aristotle called a "first philosophy," *meta*-physics can also be understood as the study of that what underlies, brings forth, or enables the physical hierarchy. This pertains to matters of *causality*.

Ontologically, Aristotle distinguishes material from formal, efficient and final causes and he assumes the existence of a primary cause to the cosmos which he calls the unmoved mover (Barnes 1984). It steers the souls in their returning cycles of coming and becoming. The unmoved mover has no cause and undergoes no change but is ultimately responsible for all movement that occurs within the cosmos. This includes the returning cycles of coming and becoming over time, and all motion of matter in space. Judeo-Christian cosmologies take over his metaphysical worldview, and reason that the unmoved mover and the cosmos it sets in motion are created by a deity. Any movement in the world occurs according to divine will.

For natural history scholars, the world abides by constant (unmoving or unperturbed) physical and biological forces, laws or mechanisms that uniformly determine the past, present and future in straight-line causal trajectories (understood in Newtonian physics and set in a Cartesian coordinate system). Laws are constant irrespective of the phenomena to which they apply.

Today, the sciences question the notion of uniformity and instead think in terms of contingency and non-linear dynamics that they model in vector and Hilbert spaces (Eldredge 1985; Gould 1989; Prigogine 1980; Smolin 1997). Classic hierarchies that depict a linear order of entities in space and over time are making room for network diagrams that demonstrate how entities *interact* and generate processes in an "extended present" (Gontier 2016b).

Modern sciences redefine causality. Non-linear dynamics and contingencies make scholars question whether constant forces, laws and mechanisms are independent entities that have existence in the world "out there". Instead, they favor *contextual process accounts* of nature. An apple will fall from a tree, unless you happen to break that fall by catching it. An explanation of the apple's fall thus needs to take the surroundings into account.

Natural selection is traditionally understood as conditional upon the existence of genetic and organismal variation, heredity, and environmental selection (Darwin 1859). These are all *processes* or *phenomena* that occur in the world (Whitehead 1929; Campbell 1974; Hull 1988; Gontier 2017). If these processes do not occur in tandem, then natural selection does not exist; and if this cycle does not repeat over long periods of time, then evolution by means of natural selection does not occur.

Cultural evolutionary theories have demonstrated that many more processes are selective. Cultural evolution (Campbell 1974; Mesoudi 2016; Bradie 2017) occurs through variation in ideas, beliefs, rituals or material artifacts that are the subject of differential learning and teaching, resulting in the retention of some of that variation in cultural tradition over others. Though the phenomena studied by biologists and anthropologists differ, the processes whereby the living and the sociocultural realms change are both selective, and both lead to a pattern of descent with modification. Selection subsequently does not manifest a law that exclusively occurs within the biological domain. Instead, many different processes that involve different entities and phenomena are selective. This implies that phenomena and processes, and not abstract laws, determine the nature of selection.

Summarizing, the goal of epistemology is to acquire knowledge on the ontological state or order of the cosmos by finding its hierarchical structure and the causes that generate this hierarchy. Such knowledge is quintessential because it enables us to understand the world and navigate within it, but we remain bounded by epistemology.

No First Philosophy

Historical research demonstrates that scholars have visualized and conceptualized the territory by different maps. Different epistemologies make us realize that *a map is not identical to the territory*. Stated otherwise, there is no one-to-one correspondence between the map and the territory. Rather, any map provides a *view* or *window* to the territory, or the map merely highlights specific aspects thereof. Why there is no straightforward one-to-one correspondence between the map and the territory subsequently becomes an independent research question.

The traditional ontology/epistemology distinction was made before the recognition that we live in an evolving world that forms part of an expanding universe and possibly a multiverse. Ontologically, it assumes:

(1) the existence of matter, space and time;
(2) the existence of one singular, hierarchically embedded cosmos; and
(3) the existence of causality, i.e. reasons that are formulated by causes, laws or mechanisms, for why the cosmos is what it is.

Epistemologically, it assumes that humans can gain absolute and true knowledge on the hierarchical and metaphysical or underlying causes of the cosmos.

These ideas track back to the ancients and the early natural history scholars that work from within a *paradigm* we now call *realism*. It is one of the oldest theoretical schools of thought developed by human beings, and one that makes the most "common sense".

However, numerous scholars have now demonstrated that many of the assumptions traditional realists made are *biased* toward and *informed* by how we, as historical, biological, cognitive, social and cultural beings, perceive the world. These ideas go back to scholars such as Hume, Kant, Herder, Husserl, Freud, Durkheim, Boas and Kroeber. How we perceive or conceptualize the world *phenomenologically* does not always, if ever, correspond with how the cosmos really is (James 1909).

We perceive ourselves as individual beings, though our bodies house three times as many micro-organisms than human cells. We perceive matter but not the (sub)-atomic particles that make them up, nor the processes that exist amongst matter and energy (Whitehead 1929). We experience our material existence as organized in time and space or place while modern sciences have demonstrated that the mass of matter and energy are interchangeable, and the independent existence of space and time is questioned and substituted by the notion of spacetime.

It makes us realize that our senses, our thinking and our languages by which we formulate cosmologies are often biased. They are biased toward our *Zeitgeist* and *Heimat*, what we today designate as our social and cultural upbringing or folk psychology (Stich 1983). And they are biased toward our evolved biological constitution (Lorenz 1941; Campbell 1974; Popper 1963) and the constraints it imposes upon our cognitive-perceptual apparatus (Goldman 2006; Bechtel 1988).

This makes us conclude that:

(1) unfalsifiable knowledge on the territory is hard to come by (Popper 1972); and
(2) the ontological state of the world, the traditional subject of ontology, is often defined differentially depending upon the epistemological paradigm (Kuhn 1962), the research program (Lakatos 1978), or the language (Quine 1951) one works in.

The *first problem* acknowledges that we can no longer assume a straightforward one-to-one correspondence between our knowledge of reality and how reality is. This makes us face an additional problem. Namely, how we can measure and compare how and to what extent our knowledge does corresponds to reality. It requires an investigation into the *content* of epistemologies, how "true," "valid," "confirmable," "testable," "(un)falsifiable," "likely," "parsimonious," and "optimal" our epistemic theories and methodologies are when they make ontological claims. Answers continue to be sought by philosophers (of science), but they are nowadays also sought by scientists.

The *second problem* asks about the origin, history and nature of epistemology beyond the content of theories and methodologies whereby it approaches ontological problems. This requires an investigation of epistemology from within the historical, sociocultural, cognitive and biological sciences.

It is important to note that in both accounts epistemology is investigated from within the sciences.

Studying and testing the content of scientific knowledge or the sociocultural and cognitive-biological act of doing epistemology from within the sciences involves a rejection of a first philosophy and an acknowledgement that there is no "God's eye view" to the world or a divine language whereby we can express matters of fact. Instead, we recognize that our languages evolved naturally, and that the knowledge we acquired is fallible, contemporary and prone to change in association with the progress made within the sciences.

The Origin, History and Nature of Epistemology

Turning to the second problem, we can roughly distinguish between two different schools of thought: the socio-anthropological school of knowledge, and the evolutionary epistemological school of knowledge. Their names foretell how they understand the study of knowledge.

The Socio-Anthropological School of Knowledge

This school goes back to scholars such as Wittgenstein and Foucault, and recognizes:

(1) that we cannot prove that our epistemological languages, diagrams, theories or methodologies refer to the word;
(2) that "regimes of truth" are partially biased, if not fully determined by human social, political, economic and cultural factors; and

(3) that the act of science or a more broader form of knowledge-seeking is a sociocultural *activity* influenced by “language games” that need to be studied as such.

Socio-anthropologists subsequently understand knowledge, not as a relation between individual human knowers and the world, but as *a relation between different human knowers* (Munz 1993). Knowledge is neither the imprint of the world upon our senses as empiricists used to think, nor an object of the mind as rationalists proposed. Knowledge is *capital* or the property of sociocultural and linguistic groups (Fig. 30.7). How and if knowledge relates to the external world (often interpreted as a physical one) becomes secondary, with most scholars in this school originally concluding that it is impossible to transcend our sociocultural and linguistic roots whereby we investigate the world. Rather, humans live in a super-organic (Sapir 1917) or super-physical world distinct from the physical and biological realm, and that superorganic structure functions as one “superorganism” (Spencer 1876).

Historically, this stance traces back to 18th century romantic movements that culminate in 19th century nationalist schools for the homeland, home culture and home language, against all others. Some of its worst outcomes include solipsism, xenophobia, racism, ethnocentrism and ethnic cleansing associated with the two world wars.

Knowing the terrors early natural history thinking had led to, and in opposition to the latter views, socio-anthropological schools from the 1950s onward opposed the nationalist schools. Going back to scholars such as Herder, Boas, Kroeber, and Whorf, socio-anthropological schools put forward historical particular, relativistic, post-modern, post-structural, post-colonial and overall deconstructionist schools of thought that often make claims against science.

It is important to emphasize that these schools fight against science as it was defined in modern times, during the Enlightenment. This references a period in time determined by Newtonian and Cartesian mechanics in physics, and unilinealism or orthogenesis adhered to in early sociology, anthropology or biology. These schools all assumed that matter in motion, biological organisms, or the history of humans, their knowledge, their languages and their cultures, follow inescapable “straight-line trajectories” or “developmental laws” toward “progress” and “increasing complexity”. These claims were presuppositions protruded by anthropocentric and Eurocentric ideas that have now been proven unwarranted and plainly false. To differentiate these unjustified theories from scientifically-grounded forms of natural history research on the natural origins of organisms, societies, cultures, languages and sciences, the older views have been renamed *historicism* (Popper 1957) and *evolutionism* (Sahlins 1970).

Questioning one epistemological framework however does not need to result in a rejection or complete abandonment of science, which is what some sociologists and anthropologists of science ended up doing. Questioning the scientific endeavor altogether brings forth the following two issues:

(1) It does not accord with the progress science makes (Laudan 1977); and
(2) It underestimates its very own claim about the power human beings have in developing epistemological frameworks as well as languages, societies and cultures.

Regarding the *first issue*, a mere comparison of older with current paradigms demonstrates that humans have gone well beyond the knowledge acquired by the ancients. Knowledge not only increases, especially the medical sciences demonstrate that certain, although most certainly not all problems can be solved. And we have been able to develop new ways by which we study organisms, languages, cultures and human history, which prove that the older ideas are indeed biased and false. This does mark progress because we can make use of science to rule out false theories. Rejecting this latter claim would place early racial claims on par with current genetic evidence that proves that on average, all humans differ only 0,02% from one another. This demonstrates that we all belong to the same species and thus that humans cannot be differentiated into distinct races. Though both claims are theories, and both are *incommensurable* because different methods and paradigms are applied (Kuhn 1962), current knowledge proves that the older ideas are false and the current correct.

Regarding the *second issue*, the social turn toward epistemology developed as a claim against a first philosophy and against science. But it has failed to see the knowledge they themselves have provided about the social and cultural act of what it means to do epistemology. For they have brought forth an epistemology of their own, one that demonstrates how epistemology indeed results from linguistic, sociocultural, and historical group endeavors. This can not only be studied, it can be studied from within the current historical, linguistic, sociocultural and anthropological sciences. Data can be quantified, new methodologies have developed, and theories can be construed.

The Evolutionary Epistemological Schools of Knowledge

Evolutionary epistemologists agree with most socio-anthropological claims on human knowledge, and go further by asking *how knowledge evolved in all biological organisms, and how they as groups construct their environments*. Evolutionary epistemology no longer understands knowledge as confined to cognition or language and as unique to humans. Rather, it examines:

(1) how all organisms acquire knowledge (or perform the act of epistemology);
(2) what the content of organismal knowledge is;
(3) how, over the course of evolution, they reproductively and socio-culturally produce, acquire, transmit and extend that knowledge into their progeny, onto other organisms, and into their environments.

The evolutionary epistemological school of knowledge goes back to scholars including Hume, Descartes, Kant, and Quine. They reasoned that the expectations we have about the world, the mathematical systems whereby we calculate the world, the languages we use to refer to the world, and the causal relationships we humans tend to abstract from our observations, can be better made sense of from within the field of psychology or what we now call the neuro-cognitive sciences.

In line with the rise and diversification of the evolutionary biological sciences, evolutionary epistemologists today ascertain that evolution is the *precondition* for all cognitive, communicative, and sociocultural knowledge that biological individuals and groups acquire, produce or transmit and extend into their environments (Bradie 1986; Gontier 2006; Wuketits 2006).

Different Evolutionary Theories Engender Different Epistemologies

Evolutionary sciences are diverse and there exist different evolutionary schools. In this part we detail how adherence to one school over another also brings diversity into the evolutionary epistemologies proposed.

Different Evolutionary Schools

The *Modern Synthesis* adheres to a Neodarwinian framework and examines how environmental selection acting upon genotypes and phenotypes brings forth new species. Organisms passively undergo selection that unidirectionally comes from an active, selecting environment.

Developmental biologists examine eukaryotic organismal development from conception until death. Extending the phenotype (Dawkins 1989), they *internalize* selection (Levin and Lewontin 1985; Gould 1977) and demonstrate, on the one hand, that organismal development occurs through a complex network of interactions occurring within the body and between gene-regulatory systems, organs, neurons, vascular, lymphoid and hormonal systems (Griffiths and Gray 1994); and on the other, the physico-chemical, biotic and sociocultural environments. This results in multilevel selection theory (Lewontin 1970; Okasha 2005) as well as evolutionary developmental and epigenetic schools that examine how the environment can alter the organism and its future generations and vice versa (Wolpert 2009; Jablonka and Lamb 2006; Hallgrimson and Hall 2011). It calls for a dualist (Craver and Bechtel 2007; Bechtel 2011) and dialectic view (Levins and Lewontin 1985) on how genes, organisms and environments relate to and interact with one another; and it brings forth the notion of *epi-genetic inheritance*, which refers to changes in gene expressions and protein functions induced by the environment.

Paleontologists investigate the evolutionary history of species as it presents itself in the fossil record which is calculated in a geological time scale, and macroevolutionary scholars study above species phenomena and investigate the causal impact the abiotic world has on life, through, for example, meteor impacts or climate change (Eldredge 1985).

Ecologists such as Van Valen (1973) demonstrate that in so far as selection occurs in the outer environment, that environment is by and large made up of other organisms. This raises questions on within and between group competition and selection (Maynard Smith 1964; Wynne-Edwards 1986), as well as how groups or colonies of the same species often behave as superorganisms (Wilson 2005), that sometimes have extended minds. Much of the latter is calculated by cost-benefit equations as they developed within kin selection and rational choice theory.

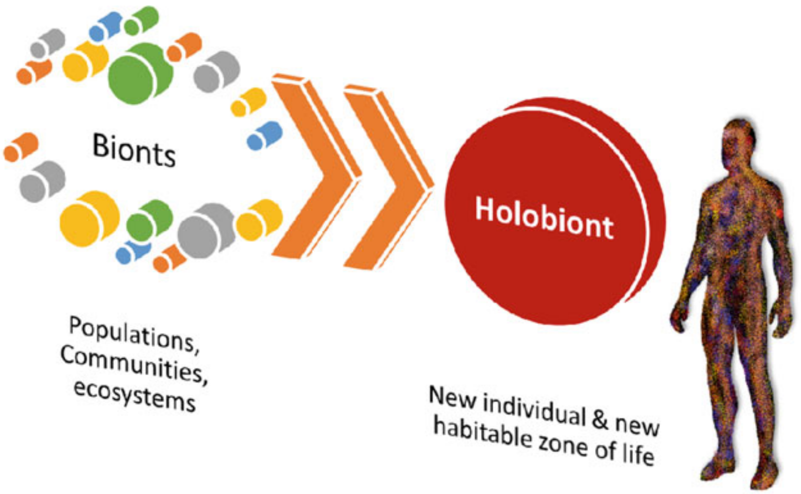
Symbiologists (Margulis 1991; Margulis and Sagan 2000) investigate how biological individuals often interact mutualistically with organisms only distinctly related to them and how they form ecological associations that have an impact that reaches well beyond the biotic environment. Life, for example, is responsible for over 90% of the oxygen present in the earth's atmosphere, and life can induce climate change. Interactions between organisms are called symbiosis and the interacting organisms are called symbionts. Symbiotic associations can underlie the formation of new tissues, organs, traits, or even new individuals called holobionts (Fig. 30.4).

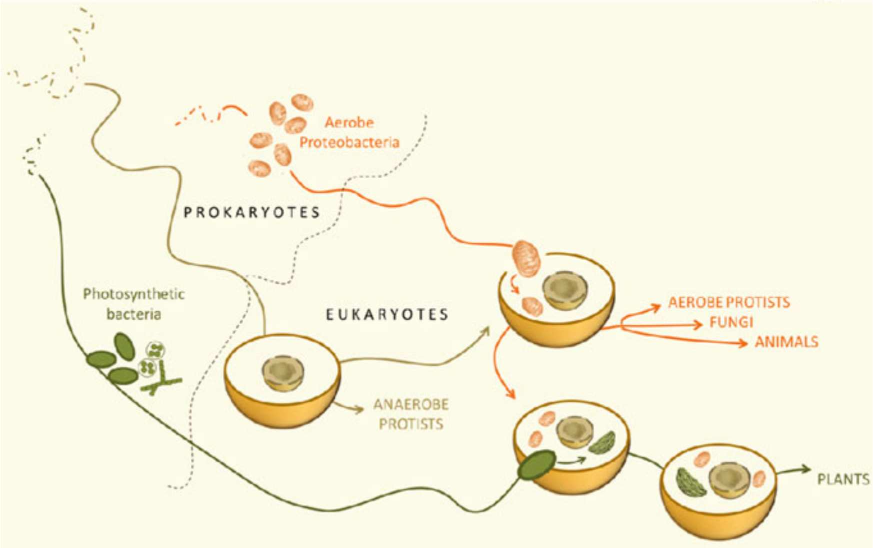
Holobionts are new biological individuals comprised of different organisms (bionts) that simultaneously function as new habitable zones of life for those bionts. A human being, for example, is not a single organism but an entire ecological community consisting of bacteria, viruses, and sometimes fungi that live in- and onside its body. Our bodies provide a new habitable zone of life for our microbiome, and our microbiome mutualistically returns the favor by underlying vital functions such as digestion.

Independently living unicellular organisms and symbionts of eukaryotes often exchange genes amongst themselves and with the host through processes of *horizontal gene transfer* (Zhaxybayeva and Doolittle 2011). Such transfer is called horizontal because it occurs during ontogeny and no (vertical) reproduction is required to acquire the genes. When horizontally acquired genes enter the nucleus, they can be passed on vertically via host reproduction.

Several organisms also directly pass on their symbionts to their progeny. *Wolbachia*, for example, are parasitic microbes that live inside several insect species. In fruit flies, the microbes can penetrate the female eggs, leading to maternal transmission of the *Wolbachia* species. *Wolbachia* can impact the reproductive success and survival of its fruit fly hosts (Faria and Sucena 2015).

When symbiosis becomes hereditary, it is called *symbiogenesis* (Fig. 30.4). When symbiogenesis or lateral gene transfer occurs, it results in *evolution through reticulation* that is characterized by lineage crossing or blending of lineages leading to a web or network instead of a tree of life. Other forms of reticulate evolution include hybridization which also enables expansion into new ecological territories,


Bionts
Holobiont
Populations, Communities, ecosystems
New individual & new habitable zone of life



thereby enabling hybrids to extend their habitable zones of life (Anderson 1949), and it enables rejuvenation of the genome.

Finally, Rousseau's observation that humans build their sociocultural environments has been extended toward other biological organisms under the label *niche construction*. Niche construction theory was first introduced within the field of ecology by Lewontin (Gould and Lewontin 1979; Levins and Lewontin 1985; Lewontin 2000). Beyond humans, all organisms often interact with the environment in ways that are specific to the organism, and all organisms actively participate in construing their and other organisms' niches. Niche construction calls out for the recognition that inheritance extends the germline, it can be ecological. And

◀**Fig. 30.4** Symbiosis and symbiogenesis. Symbiosis is an ecological phenomenon that refers to the fact that many different species live in close association with one another, either inside or onside of one another, and permanently or temporary. Symbiosis underlies the formation of holobionts that function as new biological individuals (top image). Symbiotic relations can take on many forms, ranging from mutual and beneficial to detrimental for one or all. Many of these symbiotic relations, such as the acquisition of our microbiome, are necessary for good health but only occur during and after birth. Nonetheless, symbiosis can become hereditary and lead to symbiogenesis which is evolution through symbiosis. Symbiogenesis delineates the process whereby new tissues, organs or species evolve by permanently incorporating members of older species. Symbiogenesis has played an important role in the formation of the nucleated cell and the origin of the four eukaryotic kingdoms that include the protists, fungi, animals and plants (bottom image). Aerobe proteobacteria penetrated early eukaryotic cells and evolved into mitochondria that are present in most protist, all fungi and animal kingdoms. Some early eukaryotes in addition incorporated cyanobacteria that evolved into chloroplasts present in all plant cells and chloroplasts were acquired multiple times over through secondary and tertiary symbiotic events. In all cases, the bacteria lost their identity and individuality and became part of the body of the holobiont, as cellular organelles. Nonetheless, their ancestors still roam earth today, as individuals

ecological inheritance (Odling-Smee 1988) typifies both biological and sociocultural evolution (Laland et al. 1995).

Summarizing, there exist different views on what evolution is, how it occurs, and who does the evolving. Many of the above theories originally developed outside the Modern Synthesis which is the standard paradigm that explains how evolution occurs by means of natural selection. New mechanisms and processes have been introduced, and attempts are made to extend the Modern Synthesis in order to include schools such as Eco-Evo-Devo that combine insights from ecology, development and evolution (Pigliucci 2009). The various processes whereby life exchanges information horizontally and reticulately are being grouped into new reticulate evolutionary paradigms that emphasize the important role symbiosis, symbiogenesis, hybridization and infectious heredity play in evolution (Gontier 2015a). Most of all, and causally, it calls out for a pluralistic stance: evolution occurs by a variety of distinct mechanisms and processes that often occur simultaneously. Your gene expressions might be altered by your environment and you might be incorporating new genes through lateral gene transfer acquired from one of your symbionts.

Varied Evolutionary Epistemologies

Evolutionary epistemologies are equally diverse and depend upon the evolutionary views adhered to. In fact, evolutionary epistemologies evolve with them. Many of the founders of evolutionary epistemology (Lorenz 1941; Campbell 1974; Skinner 1986) actively participated in founding (comparative) behavioral, ethological, cognitive and sociobiological evolutionary sciences.

The research programs have now been incorporated into these sciences that study how cognition, behavior and communication evolves in all biological species, how organisms embody that cognition, and how it relates to the organism's external environment. For classic evolutionary epistemologists, the question how evolved organisms relate to an outer, physical world remains meaningful. Traditional fields study organismal traits exclusively from within Neodarwinian schools of thought that emphasize adaptationist views. Adaptation is a term first introduced by Lamarck and "literally (refers to) the process of fitting an object to a pre-existing demand …" by assuming that "organisms adapt to their environment because the external world has acquired its properties independently of the organism" (Lewontin 2000: 43). Supporting that selection occurs from the environment onto organisms, traditional evolutionary epistemologists understand organisms as unfalsified *conjectures* or *theories* about the world that somewhat corroborate to it (Campbell 1974; Popper 1963). This enables and endorsement of *hypothetical realist* views. Epistemology understood as evolved knowledge continues to be different from ontology or the world as it is in itself, and the question becomes how the evolved theories or hypotheses that come in the form of organisms refer to the outer world.

Today, due to advances in eco-eco-devo, evolutionary epistemologists endorse radical constructivist (Riegler 2006) and non-adaptationist views (Wuketits 2006), as well as moderate (instead of hypothetical) realist views on how knowledge relates to the outer world (Clark and Chalmers 1998; Munz 1993; Ruse 1989). In moderate realist views, the mind and organismal bodies function as media or mediators between organisms and the environment. In radical constructivist views, the mind has priority in constructing an experiential world of its own that does not necessarily relate to an outer world. And from within non-adaptationist views, knowledge is understood as a *relation between organisms* in the same sense as socio-anthropologists and socially-oriented philosophers of science understand it as a relation between human knowers. How this knowledge relates to an outer, physical world then becomes secondary.

In the remainder of this work, we shall extend upon these traditions and go further than moderate realist, constructivist and non-adaptationist views by demonstrating that the relation between epistemology (in the form of organisms) and ontology (as an "outer world") becomes superfluous. Organisms reconstruct the earth, not just in their minds, they *embody* that knowledge in their anatomy and cognition, and they *extend* it onto their progeny and into the niches they construct. Ever since life evolved, life has rebuilt earth inside out, recycling existing matter, energy and space made in previous moments in time, into a living earth, up to the point that earth no longer exists as a purely physical "outside" entity. If that abiotic entity once existed, it now exists no more. Rather, it evolved into a living planet through the organisms that reconstruct it from its subatomic particles onward by reproducing and constructing new material life forms as well as extended and equally material niches.

Organisms and the environments they build (epistemology understood as evolved knowledge) are what is real (ontologically), and the relation is *exclusive* because there is no outer abiotic earth anymore. Our living planet is not just hypothetically real, it is spatiotemporally real, or stated otherwise variant in time and space.

Organisms build biologically-informed or evolved realities or bio-realities that include the construction of local environmental and sociocultural niches. The living earth evolves in congruence with these expanding (generating or speciating) and contracting (degenerating or perishing) bio-realities that are dependent upon organismal and species survival, reproduction and extinction as well as the ecological materializations they bring forth in time and space (or spacetime). Epistemology, *understood not as theories but as the evolution of embodied knowledge in organisms and their extended niches that underlie bio-reality formation*, therefore *equals* ontology, the current living world. One might call this position *radical spatiotemporal realism*, but I prefer to understand it as the outcome or consequence of *applying evolutionary frameworks* to matters of *epistemology* that show that epistemology equals ontology, which I call *applied evolutionary epistemology*.

A New Cosmology

We started this chapter by demonstrating how cosmologies render epistemologies on the cosmos by providing theories on the nature of *matter*, *space* and *time*. Thus far this has involved a consideration of how matter occupies space which results in *hierarchy theories*, and how matter extends over time which results in *causality theories*. But the cosmologies developed so far are static and do not take evolution of either the map or the territory into account.

Today, we know that *matter* is equivalent to *energy*, space and time are joined into a four-or-more dimensional *spacetime*, and there is growing support that our cosmos forms part of a *multiverse*. How we have conceptualized matter, space and time is therefore not (completely) true.

Matter and energy, we find in the organisms that constantly recycle and rebuild a new earth out of an older one, through the acts of consumption, reproduction and expulsion. Living organisms constantly generate new matter and energy that they extend into their progeny, onto other organisms and onto the environments they rebuild and construct anew. As such life regenerates or *re-cycles* earth (its old *spaces* it occupied in the past), and we build a new earth (or new spaces in time, or new *space times*).

Advances made in modern physics as well as socio-anthropological and evolutionary epistemological schools demonstrate that we have outlived the classic epistemology/ontology divide. It is no longer useful to us, because there is no single static cosmos "out there" that organisms acquire knowledge on or adapt to. What is real evolves which makes reality variant in space and time. What is true at one point is therefore not necessarily true at another, which makes knowledge spatiotemporal or local. In so far as organisms embody and extend their knowledge into their progeny and onto their environment, they make reality happen every day. We make the living earth happen every day. However local and variant reality and knowledge might be, they are both real, and what is more, they are equivalent. In this part, we demonstrate how *epistemologies simply are ontologies*, which makes any distinction between them unsustainable.

Thinking Through the Consequences of Symbiosis and Niche Construction for Ontological Hierarchy Theories and Causality Theories

Niche construction and symbiology theories make a straightforward link between epistemology and ontology, or the organisms and the niches they construct on the one hand, and the outer physical world on the other, problematic. Here we think through the consequences of symbiosis and niche construction for the construction of bio-realities.

But before we do, we need a note on niche construction. Niche construction theory was redefined by Odling-Smee (1988) and Laland et al. (1995) as a form of adaptability or a capacity to become adapted to the outer environment. This view is now incorporated into the new evolutionary sciences that include evolutionary psychology, evolutionary linguistics, evolutionary anthropology, evolutionary sociology and evolutionary archeology. This move is rather unfortunate. Lewontin (2000), who coined the term, defined niche construction as a capacity for organisms to develop a world of their own, distinct from what exists "out there," or better yet, what existed before constructing organisms entered the scene. It enables survival *despite* the environment organisms are born into.

Contrary to this view, current niche construction theories emphasize adaptation or adaptability of organisms to existing sociocultural or biotic niches that are local in scope. It underestimates the very claim made about the important role generations of organisms have in actively building a world of their own, and it recalls the problem also socio-anthropological schools face. They too underestimate the creative force of humans in actively construing their sociocultural and linguistic environments and in lieu focus on deconstructing science. A consequence is that they understand organisms to primarily conform or adapt to a given and somewhat stable biological or sociocultural environment, which are the niches constructed, and only in a later phase can individuals modify it. It underestimates the creative power organisms have in continuously bringing forth new niches, new bionts and new holobionts.

However, an organism-based construction results in new realities that are different from the older ones and that surpass the older in both space and time. They do not infiltrate existing structures or fit on top of older structures, they *replace* older structures. Niche construction theory can fare much better by abandoning both its notions of adaptation and adaptability. These are non-evolutionary because they accept an outer, somewhat stable world. Adaptation or superorganic realms are concepts belonging to older cosmologies, they are not part of the new worldview that is developing. For the same reason, we shall also surpass Levins and Lewontin's (1985) Hegelian and Marxist dialectic position.

Turning to reticulate evolution, it conflicts the traditional views on the genealogical (gene or replicator-based) and ecological (phenotype or interactor-based) *hierarchy* (Tëmkin and Eldredge 2015). The genealogical and chronologically linear hierarchy traditionally goes from genes to cells, organisms, species and higher taxa. However, reticulate evolution crisscrosses and jumps between levels of such a hierarchy, often instantly creating new genealogical hierarchies that take on the form of holobionts at any level of an existing hierarchy. A holobiont is often made up of bionts belonging to the three different domains of life; hybridization can occur between distinct sub-species and species belonging to different genera, families or orders; and lateral gene transfer occurs within and between prokaryotes and eukaryotes. Reticulate evolution instantly alters existing genealogical hierarchies creating new ones that have their own trajectories. And it alters spatiotemporal ecological hierarchies that traditionally line up as going from organisms to populations, communities, ecosystems and the biosphere. One holobiont is an entire ecological space or habitable zone of life for the bionts that make it up.

So far scholars have only studied life in space and over time, but not in spacetime. Linear and single hierarchies induce discussions on arrows in time (Gould 1989; Prigogine 1980), on how major transitions between levels of a linear hierarchy occur (Maynard Smith and Szathmáry 1995), and on how *causality* occurs; upward, which brings forth reductionist worldviews (Dawkins 1983), downward (Campbell 1974), which brings forth holistic views, or through a combination of both (Bechtel 2011; Lewontin 2000) which brings forth cyclic or dialectic views (Fig. 30.5).

Upward causation correlates with linear hierarchies that describe and explain events over *time* (in chronologies, for example, or *genealogies*). The focal level is the level of study, and in upward causation, the focal level is explained by going down one level of the hierarchy. Suppose the focal level is the organism. To explain how it originates in time, Neodarwinians go down one level of the hierarchy to genes and examine how they form organisms (e.g. Dawkins 1983). Organisms in turn bring forth species. This gives a straight-line and irreversible trajectory, and when investigating the history of life, it makes sense that species cannot precede organisms that build them, and organisms cannot precede genes that underlie organismal form.

In *downward causation*, the focal level is explained by the level above the focal level. It associates with holistic views, and it investigates matter in *space* or what we may call an *extended present*. Examples include *ecological hierarchies*. Suppose the focal level is again an organism. To explain group selection, which remains a controversial theory, scholars go one or two levels up the hierarchy to populations and communities and examine how they can cause (groups of) individuals to be decimated, to go extinct, or, to be favored in the inter-organismal struggle for existence. This can only happen when (different) groups, populations,

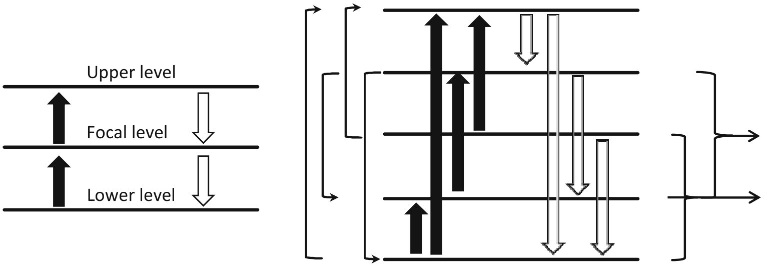


Fig. 30.5 Traditional versus new hierarchical views on ontology and causation. *Left*, the traditional way whereby scholars understand hierarchies as either undergoing upward causation (marked in the black arrows) or downward causation (marked by the white arrows). *Right*, we depict how we are going beyond classic notions of causality. So far, focal levels are only explained by the levels close to them, one level up, one level down, or through a combination of both. For one, there is no reason that either up- or downward causation cannot extend their influences on more than one level up or down the hierarchy (depicted by the same arrows as in the figure on the left), or by simply skipping some levels in its causal influence (depicted by the black arrows on the left of the levels). Secondly, and what is typical about symbiosis and other types of reticulate evolution, is that it jumps in between levels of the chronological or genealogical hierarchies, instantly creating new ones (depicted on the right by the accolades). The picture shows that it cannot be depicted comprehensively in traditional hierarchical lineups, which is why scholars are turning to network diagrams

and communities already exist, which requires a study in space or in an extended present. Similarly, suppose the focal organism represents a human child learning to write his language. It learns it from its teachers that are part of his community, and the child can only learn how to write his language because the community already has a writing system.

Great controversy resides over whether downward causation is not just upward causation recurring cyclically or recursively over time (e.g. Craver and Bechtel 2007; Bechtel 2011). It depends upon how one understands the phenomena tracked and represented by the focal level, either as identical and resulting from the same trajectory (stable genes that are faithfully transmitted over generations of, nonetheless different individuals), or as resulting from a different trajectory (because each individual is unique and thus has its own trajectory), or from a trajectory that perhaps crosses the focal level (through, for example, lateral gene transfer). The latter two examples imply *non-linear* and *multi-linear* dynamics and interactions *between different hierarchies* which requires *non-linear and multi-linear causation theories*. A sometimes causes B, B is sometimes caused by a combination of C or E, and at other times by D.

Although not stated explicitly, this view is adhered to by Eldredge (1985; Tëmkin and Eldredge 2015), who understands the genealogical and ecological hierarchy as different from one another yet interacting.

One of the things that symbioses demonstrates is that we need to go beyond. There is a reason why these events are being depicted by networks instead of hierarchies. Bacteria can instantly infect organisms at any “scale” or “level” of the hierarchy, and when they do, they bring forth a new reality in the form of a holobiont that immediately also functions as a new habitable zone for life. They *pop up* at any existing level of the hierarchy, and *jump* between hierarchies, without having to rewind or relive the previous genealogical chronology or grouping into an existing ecology. When trying to model that in traditional hierarchies (Fig. 30.5 **on the right**), it does not look clear, while networks or webs of life facilitate comprehension.

By investigating how the genealogical and the ecological hierarchy interact, Tëmkin and Eldredge (2015) open new research questions on how many hierarchies there are, and how they can become combined (Gontier 2010, 2017). In short, it necessitates pluralistic accounts on hierarchies that are better depicted into networks set in vector or Hilbert spaces, keeping in mind, of course, that networks remain hierarchical, and that any event has its own peculiar trajectory. And they require new causal explanations.

Much of these networks nowadays remain "unrooted" because we have no idea how to conceptualize *time* which today often is no more than a measure of distance in space. But we can go further than that. What processes of reticulate evolution and niche construction demonstrate, is how entities and processes, *distinct in space and time* from one another, are combined into a new *spacetime*.

Perhaps what I am saying can be made sense of by drawing analogies with Einstein-Rosenberg bridges that alternatively go by the name of "wormholes". But caution is required. For one, a wormhole, as traditionally conceptualized, is still too small (10^{-33} cm or -230000000 nm) for even the tiniest virus (i.e. the *Porcine circovirus*, 17 nm or 0.000002 cm) or prokaryote (i.e. the *Nanoarchaeum equitans* archaea, 400 nm or 0.00004 cm) to pass. Nonetheless, scholars are calculating how wormholes can be stretched. It is remarkable though, that it is viruses, archaea and bacteria, the smallest living entities on earth, that are so swift in their crisscross travels across niches and organisms in space and time or spacetime. Physicists theorize about parallel universes or the impact spacetime travel has on the traveler. On earth, one can safely say that symbiosis changes the identity of the traveler. Free-living cyanobacteria are quite different from the chloroplasts they evolved into when they entered eukaryotic cells; and every chloroplast inside a plant cell, is just like the nucleus of that cell, unique because of its specific genetic code as well as its life history. Viruses, such as the flu, attack in specific periods in time and space which leads to epidemics and pandemics. But where they go to in between, nobody really knows. If they are always around us, and everywhere, they should infect us all the time too. But some do not, and it is more likely to catch them in specific times of the year, around infected individuals. Viruses contain the most different genes, about 80% of them are exclusively found in these viruses. Several scholars (Villarreal and Witzany 2010) also consider viruses as preceding and perhaps underlying the origin of life on earth. If they would be space time travelers, then, and if you allow me the anthropomorphic expression, their attempts at infecting us makes one wonder what kind of (passed, distant or distant past) world they are trying to salvage by bringing it into the present hoping it will survive.

Another issue with wormholes is that they have this almost mystic air around them. But there is no reason to assume they only occur in galaxies far away. Theoretically, they can also take place right next to you, and perhaps even inside of you. Physicists do not know what happens once something goes inside, or what happens once it comes out, if that is at all possible. Biologists on the other hand, can not only observe bacterial or viral infections with their microscopes in "regular space", genetic engineering actually induces them all the time. Through acts of artificial symbiosis and artificial lateral gene transfer, genetic engineers alter genetic codes of organisms. By inserting foreign genes into viruses and letting them infect laboratory animals they investigate what anatomical, cognitive and behavioral changes the new genes induce. Whether this just happens in space and time or in spacetime and through wormholes is really something for physicists to calculate and have their say about. For now, it's a good metaphor by which we can think about these phenomena and investigate them further.

Because we can readily implement these ideas in our daily lives. Search your house for all the electronic equipment you have, and check the date and location it was manufactured. You have been bringing quite some different matter, made in different spaces or places with different time zones and manufactured in past years together during your lifetime. Yet it all forms part of your extended present. We are accustomed to understanding our houses as the result of labor and transportation of goods due to commerce and consumption, but perhaps that view is old-school and it is, instead, a form of spatiotemporal travel enabling you to create your niche. Your smartphone might be the same brand as mine, produced in the same year and the same factory, but it is different from mine because of its content. The same goes for the bionts we gather during our holobiont lifespans, and all can be captured by the notion of universal symbiosis (Gontier 2007).

Turning to *how we conceptualize the past*, we are accustomed to thinking about the past as something that lies behind us, in what is called a distant past. In our cosmographies, it resides somewhere far away on the lowest scale of the ladders, timelines or hierarchies we have built. But one of the things current physics is teaching us is that the past is, in fact (not just in poetry) all around us. We see the moon as it was 1.2 seconds ago, and the sun as it was 8 minutes ago. The more distant in space we look, the more back in time we go. The Hubble telescope, for example, enables comparisons of other galaxy formations it observes in space which enables conclusions on how our galaxy possibly formed (https://www.nasa.gov/press/2013/november/hubble-reveals-first-pictures-of-milky-ways-formative-years/#.WfStRWiPI2w).

In the opposite direction, gravitational waves or ripples in spacetime are teaching us is that some of that past is just reaching us now. Two years ago, observers detected a gravitational wave in spacetime that was presumably caused due to the collision of two black holes, far away from us and in a distant past (Abbott et al. 2016). When the gravitational wave passed by, it was rather swift at that, and wherever it is headed to, it concerns its own future, which might not necessarily be ours. Ever since, scholars have detected other such waves to pass by.

Coming to Terms with an Expanding and Evolving Multiverse

Symbiology demonstrates how *multiple* holobionts are formed from bionts that in turn construct new niches that additionally function as new and multiple ecologies. Holobionts, as niche constructors and as ecology providers, extend and significantly alter the world.

The "outer" world or environment where epistemology tries to get a grasp on has classically been interpreted as singular and purely physical. It either corresponds to the universe, earth, an abiotic environment, or a "more fundamental" physicalist level. (Holo)bionts alter that physical world and play an important role in "abiotic" processes such as the nitrogen, oxygen or carbon cycle, the earth's temperature, and the earth's atmosphere (Volk 2017). On earth, most organisms turn into dust, mud, soil or stone because if the conditions are not right, they will not be preserved. But no matter how deep one goes into water or digs inside the earth's mantle, so far life is found everywhere. Even in volcanos and acidic environments. Life thus significantly alter the spheres of the earth, extending well into space.

Dissecting any (holo)biont to its smallest particles, we find that they are made up of the same (sub)atomic particles that build matter. But those particles simply do not *explain* all there is to life. A reductionism to a purely physical stance is unwarranted. And downward or cyclic causation does not suffice either because life builds new genealogical and ecological hierarchies all the time, thereby introducing new spaces that all follow their own times and that combine different times or circadian rhythms together. Since its origin, life has incessantly created new realities from the subatomic particles onward and it is all real. It has created numerous new phenomena displaying all sorts of behavior.

Living organisms evolve this knowledge and transmit information on it to future generations, on to organismal neighbors, and they store it outside of them in their extended niches. This knowledge does not so much provide a theory about an outer physical environment, as information on how bio-realities can become construed and how one can survive within them.

Bio-realities alter the purely physical realm inside out, up to the point that such a realm has no independent existence anymore. That means that if earth once was a purely physical or physico-chemical object, today that object exists no more. It has traded place with the incessantly and newly evolving bio-realities. Life simply replaces the physical earth by recycling it.

On that view, knowledge no longer concerns a hypothetical relation between an organism and its external environment. *Knowledge is an evolving phenomenon that materializes into organisms and the overlapping biological realities they construe* (Fig. 30.6).

There are no doubts about adaptation, correspondence or truth values of the knowledge and information that life evolved, because there simply is no independent physicalist or physical ontological reality to compare it with. What is true for one organism, might not be true for another, but it does not make any of these organisms

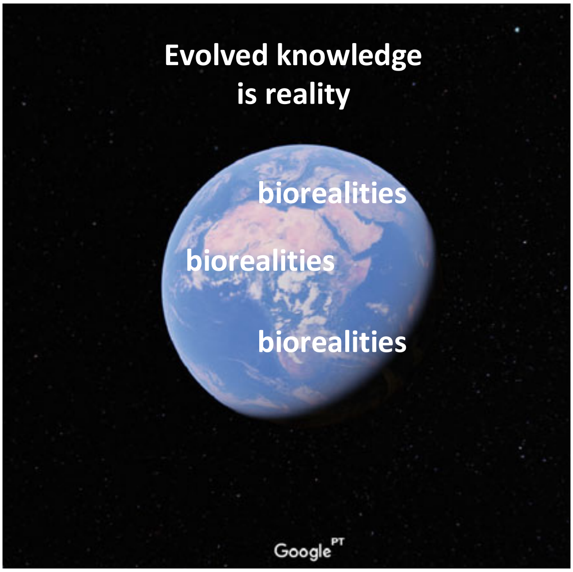


Fig. 30.6 Bio-realities and the equivalence between evolved epistemology and evolved ontology (Photo obtained from Google Earth that is under a creative commons, and adapted)

less real or existent. And what is true in one niche might not be true in another, but that does not make it less real locally. They *are* the currently existing *realities*.

Ontologically, the only comparison we can make is how the living earth relates to other planets and how it stands in the universe or multiverse. But on the one hand, that implies such a redefinition of ontology that one can wonder how useful that is. It would make more sense to give up on the ontology/epistemology distinction altogether.

On the other hand, we did not make the oceans of our world, but we use them for transport and we pollute them which alters their biotic and abiotic composition. We and other organisms such as the wolves that were reintroduced in Yellowstone Park change river banks and all organisms, even bacteria, change the composition of the soil and the atmosphere. We do not make the planets orbit around the sun, but we witness the events. There is a *past* universe out there. And in our entangled ways, we are the ones that see the material traces or the light it left, by bringing it into the present and into our biological realities through our evolved cognition and the extended instruments we make such as the Hubble telescope.

Some scholars wonder, for example, if the black hole that presumably resulted from the collision of two black holes, and that presumably caused the gravitational wave, is still there now, in its present. But one of the things our current knowledge on the speed of light and our measurement in light-years teaches us is that we are looking at structures belonging to a distant past. What the Hubble telescope sends back might be a picture of the "dead," comparable to tangible fossils we find in geological strata of species long extinct.

Our trips to the moon or one day soon mars furthermore demonstrate that we can bring the past into the present. And in so doing, our trips or technological missions such as the Mars Rover change those entities. If they were once lifeless (which is currently questionable for Mars), they are now planets where earthly life has extended toward. That is what evolutionary scholars call variation or even speciation through time, or what philosophers call a change in kinds. Life changes the ontological state of (parts of) the universe, not merely by thinking it with our minds, but by observing it happen or by actually doing it by going there and altering what once was, forever.

Finally, there is no real reason not to understand the physical cosmos or multiverse as a living "something", just like Lovelock and Margulis (1974) understood earth as a living planet. The cosmos can be understood as an individual that has a beginning, lifespan, and end (Ghiselin 1974). We already know that the universe metabolizes by expanding, and it is likely to reproduce by making more selves (Smolin 1997; Everett 1957). If true, then the multiverse, just like us organisms, evolves knowledge and constructs its own worlds. It makes symbiosis not only universal (Gontier 2007), but multiversal (but see Volk 2017, for example, on an abiotic view).

Summarizing, there simply does not exist one eternal physical or physicalist world out there, and there does not exist one truth. The universe or multiverse might be more durable in time, but it is not fixed. It also changes and evolves. What we are left with here on earth, are expanding and contracting biologically-informed realities or bio-realities.

For a detailed research program on how evolving knowledge and transmission thereof can be studied in all organisms from within these diverse evolutionary sciences, we refer the reader to Gontier and Bradie (2018; Gontier 2010, 2012). Here, we continue to focus on the implications of how we understand epistemology defined as evolved and extended knowledge and information.

Revising Traditional Evolutionary Epistemologies Considering the Newly Evolving Cosmology: Implications for Knowledge and Truth

Classic evolutionary epistemological insights include that:

(1) Organisms are embodied theories about the environment (Popper 1963; Campbell 1974; Wuketits 2006);
(2) Mechanisms are methodologies or heuristic search engines for acquiring theories about the environment (Campbell 1974; Riedl 1980);
(3) Human theories are disembodied organisms that evolve (Popper 1963, 1984).

We can adjust these views and say:

(1) (Holo)bionts are not just embodied theories, they are real and so is the knowledge they embody and evolve; and we can add that the niches they provide for other bionts, and the niches they build are not extended theories but spatiotemporal realities or bio-realities that often extend their makers in spacetime;

(2) mechanisms need to be replaced by process accounts, and what we find is that distinct processes have converging patterns in modes and tempos;
(3) the content of knowledge and the constructs (holo)bionts make indeed evolve, in congruence with their evolution; consequently, truth or reality is not one but varied; but in each variation, knowledge and reality (or the map and the territory) are equivalent.

(Holo)Bionts Are and Construct Bio-Realities

One of the major claims made by classic evolutionary epistemology is that it understands organisms as embodied *theories* (or conjectures in the Popperian sense of the word) about the world. Knowledge subsequently becomes redefined as a relation between the organism and its environment. Here, we examine and compare this claim to how socio-anthropological scholars define the historical, cognitive and sociocultural nature of epistemology and how philosophers of science evaluate the content of epistemology (knowledge). Afterwards, we examine how knowledge materializes in progeny, in other organisms, and in niches, making claims about an independent environment unwarranted.

Epistemologies as Methodologies and Theories, the Socio-Anthropological View

Epistemologies provide *knowledge* of the territory (understood as an independent or outer physicalist, biological or sociocultural world) through *theories* and *methodologies* (Fig. 30.7).

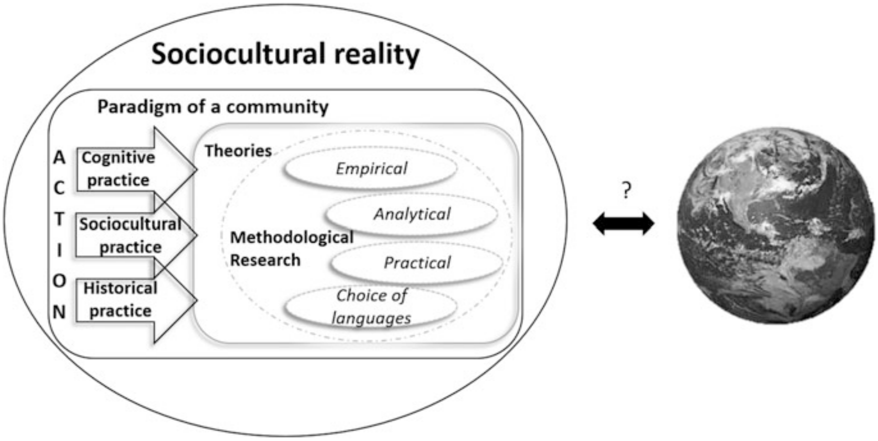


Fig. 30.7 Socio-anthropological view on epistemology. The figure also portrays what is known as the *reference problem*, the question of how human knowledge that is formulated in natural, formal, or mathematical languages relates to the world

Theories are obtained by *empirical* (observational), *analytical* (ideational, conceptual) and *practical* (experimental, instrumental, and technological) research that abides by *methodologies or* research *programs*. Research programs delineate a set of procedures or rules for how research is performed and how theories are formulated. Boundaries between theories and methodologies are indeed fuzzy, with some positing the primacy of methodologies over theories (Lakatos 1978), while others claim the opposite (Popper 1963; Kuhn 1962). Fact is that many of the current methodologies that scientists apply are informed by theories and vice versa.

One way to distinguish between them is by understanding methodologies as corresponding with *the act of doing epistemology which is acquiring knowledge* (through e.g. science). Theories refer more narrowly to the specific results obtained, i.e. *the content of epistemology which is knowledge*.

Theories are traditionally articulated in natural, formal, or mathematical *languages*, and applying methodologies often involves a choice of particular languages over others to formulate theories (Russel 1914; Stewart 2011).

Both theories and methodologies are dependent upon, and informed by human *cognition* as well as *historical and socioculturally-informed individual and group action* or *power* and *practice* (Bourdieu 1977) that is defined through concepts including field, habitus, capital, and doxa. Human *cognition* results from embodied, embedded, enacted, and extended minds (McLuhan 1964; Clark and Chalmers 1998; Rowlands 2010); and *action* results from *historically-informed, individual and sociocultural group behavior*. Together they form *mentifacts* that underlie *sociofacts* that often materialize into cultural *artifacts* that include scientific instruments (Huxley 1955). The result is knowledge that *extends* the individual knower, the sociocultural group it belongs to, and the time and place it first originated. Materialized, knowledge gives way to what Rousseau called artificial cultural societies, societies that extend and surpass our biological nature and natural habitats. This view grounds the classic nature/nurture divide, and the idea of a super-organic structure that is superimposed upon the biological and physical realm. Methodologies, the theories they propagate, and the cognitive, historical and sociocultural practices that underlie them are referred to as epistemological frameworks or paradigms (Kuhn 1962). Paradigms refer to the totality of knowledge of a scientific community. Summarizing, epistemology always has three sides to it:

(1) a methodological part that is itself informed by theory that refers to the act of doing epistemology which refers to acquiring knowledge;
(2) a content part, that refers to the actual knowledge that becomes formulated into theories;
(3) performing methodologies and formulating theories are cognitive, linguistic and sociocultural, individual and group endeavors that extend and materialize into sociocultural territories or realities.

(Holo)Bionts and the Niches They Build Are Knowledge

Here, we demonstrate that (holo)bionts meet all requirements imposed on *methodologies* that enable *an act of doing epistemology* which leads to the *acquisition of knowledge*.

All (holo)bionts *empirically* explore their extended present, mostly for food or shelter, and in most eukaryotes, for mates. They observe their niches. If not by making use of their evolved bodies that sometimes enable locomotion or complex senses than enable vision, touch or smell, then through complex biochemical processes. Slime molds (Reid et al. 2012), for example, are colonies of individual slime cells. These cells can live independently, but they often team up. They do not have a nervous system and thus nothing that resembles a memory. Nonetheless, when foraging for food, they will avoid places where they have foraged before. Not because they "remember" where they have been, but because they avoid the biochemical signals their slime trail left in the places they already foraged. The trails they leave function as an external memory map that enables successful navigation and exploration of their local niche. All (holo)bionts possess knowledge about their local niche, and they externalize it and leave trails of it, which is part of the process we call niche construction. Many (holo)bionts also perform *analytic research* of their niche. Animals do not always need to act to know or learn something. A moth flies to the light and burns its wings and dies. But we rarely see horses or lions walk into a fire or jump off a cliff. The neurocognitive sciences have demonstrated that thinking can be non-linguistic, and what we are used to call categories of the mind is present in other animals. Most eukaryotes "know" or recognize their children, and they know how many there are because they will look for them when lost. Spatiotemporal awareness, number sense, paternal and maternal relationships are traits currently studied and found to have evolved in quite a number of species. Many primates in addition have rudimentary theory of mind. They know that others know. Consequently, they will hide food or suppress food calls from others and only share with conspecifics of which they know shared food with them, or helped with grooming or fights.

All (holo)bionts evolve *practical methodologies*. Socio-anthropological schools of thought define practical methodologies as experimental, instrumental and technological research. Behavioral research has demonstrated that we are not the only ones that do so. Many species engage is social play which is often a way to experiment or practice hunting and fighting. Numerous (holo)bionts make use of their niche to build instruments or tools. A honeycomb is an extended complex instrument and technological complex that houses larvae and fabricates and stores honey. Termite mounds are equally complex factories that function as protecting nests for their inhabitants. Ant and bee colonies function as single individuals or superorganisms (Wilson 2005), and such requires complex forms of communication between e.g. the workers, the soldiers and the reproductives.

Turning to language, that might be uniquely human. But many (holo)bionts have evolved *complex communicative systems* for intra- and interspecies communication as well as for internal e.g. intracellular communication. Much of this can be studied from within the field of biosemiotics (Witzany 2014). Ants communicate through pheromones. RNA intermediates between DNA and proteins. Bacteria communicate through chemotaxis. Viruses possess the biochemical keys of our bodies locks, and they can fence of or immobilize our body's immune responses. Prey have often evolved forms of mimicry and either have the shape of predators, or they take on the colors of the niche to hide from them. As Darwin already noted, sentient organisms have evolved a series of expressions and emotions that inform their niche about their physical or mental state. Mice communicate through ultrasonic vocalizations, bats through echolocations, snakes understand their niches through heat-maps, and scorpions and butterflies not only see but respond to ultraviolet light. Primates have evolved complex multimodal forms of communication that make use of a combination of vocalizations, gestures, expressions and emotions, and, in humans, we add to that words or symbols. Words are by far the most deceptive way whereby we can communicate false or fantastic ideas that *dissociate* with the niche. Most animal communication systems are instead *associative*, they communicate about real-life events though they can lie about whether they are ongoing or not. In sum, (holo)bionts and the structures they are composed of have evolved methodologies that enable them to *acquire* and *build* knowledge that they *transmit* and *extend* onto their offspring and into their surroundings where it *materializes* and *alters reality*. In so doing, (holo)bionts and their extended niches are more than just theories about an external world. *They are knowledge, and that knowledge exists in the living earth that is made by it, they are reality.* Knowledge therefore is reality, or, stated otherwise, the map evolves the territory.

Evolved Knowledge Materializes into New Realities, Epistemology Understood as Knowledge Equals Ontology

In association with the evolutionary sciences, evolutionary epistemologists demonstrate that all (holo)bionts possess and evolve knowledge about their internal and external niches. In association with the socio-anthropological schools of thought, they have demonstrated that all (holo)bionts are actors in this world. They have evolved anatomical, cognitive, behavioral and sociocultural practices that extend into and modify existing niches and (holo)bionts pass on this knowledge, through the germline, horizontally and multi-directionally through learning.

We can add that this underlies the formation of new, biologically-informed or evolved realities which we call bio-realities (Gontier and Bradie 2018). Bio-realities are neither "purely" physico-chemical, nor exclusively biological or sociocultural. They are also not a new "realm" that "emerges," "infiltrates" or seats on top of older realms. They are new realities in spacetime that *replace* older realities, all the way down to its subatomic levels. They are, what ancients used to call a microcosmos that embeds within it a macrocosmos. And our living earth in turn is embedded within a multiverse.

Although this work focusses on bionts and holobionts, the evolved genetic codes can also be understood as evolved methodologies that provide information on material biont formation. Besides being in constant communication with our cells and our extended present, they are by far the most erudite on how the abiotic matter that surrounds us can, from the subatomic level onward, be recycled and brought into our world as living matter and energy. And the life it brings forth in turn incessantly alters the genetic codes through, for example, introgression of foreign genes into existing genomes.

In sum, the distinction between theory and knowledge becomes superfluous. (Holo)bionts and their niches are real entities in the world that underlie the formation of altered or new realities. Knowledge can no longer be understood as a homogenous entity that refers to a homogenous outer world of which some of its levels are more real, stable or permanent in time than others. Knowledge is particular and dependent upon the evolved bio-realities. What is true in one niche is not necessarily true in another, and when (holo)bionts die, their knowledge often dies with them, unless they were able to transport it into the niche, offspring or other (holo)bionts. Nonetheless, a purely solipsist view is impossible, because we are evolutionary related by common descent, and we all inhabit the living earth.

Process Accounts and Recurring Patterns

Traditional evolutionary epistemologists mainly worked from within Neodarwinian schools and understood evolution to happen by means of natural selection that was interpreted as a mechanism. Many also understood natural selection as a *methodology* that acquires knowledge about the world (Campbell 1974; Riedl 1980). On that account, natural selection is nature's way to build *theories* about an outer world.

Today, scholars recognize that evolution can occur by a myriad of "mechanisms" including drift, symbiosis, lateral gene transfer that all refer to distinct and ongoing processes. These theories are currently being "universalized" towards domains that extend the classic biological sciences, such as linguistics, sociology, and anthropology. As explained in the introduction, many social and cultural processes can be understood as selective. Interesting in that regard is that especially Campbell, and though not explicitly, understood "universal selection" as a recurring *cycle* of what he called *blind variation and selective retention* occurring over repeated periods of time. This cycle brings forth a pattern that recurs in the evolution of culture, of languages, and of anatomical form. "Descent with modification", is another pattern selection brings forth, but all known "mechanisms" bring forth this pattern.

Reticulate evolution, that brings forth horizontal patterns of information exchange and lineage crossings or blending, also characterizes processes of language mixing, or cultural hybridization. Drift theory that brings forth random patterns of evolution not only typifies how genes or (holo)bionts migrate and evolve, it is also found in how languages and material artifacts diffuse. And besides gradual patterns, also the pattern of punctuated equilibria (Eldredge 1985) has been found in the evolution of certain languages, species, and material cultural artifacts (Gontier 2015b).

A universalization of evolutionary mechanisms often implies a transition from mechanism to process accounts as well as an identification of recurring patterns (Gontier 2017, 2018). Process accounts demonstrate that mechanisms are not laws or forces, but conditional upon phenomena behaving in particular ways. It is the phenomena that demonstrate selective behavior or not, but there is nothing above or beyond the phenomena. Selection is not some force or law out there waiting to act. Only phenomena and processes exist. Mechanisms do not exist and can therefore not be methodological. Recurring patterns, these continue to provide *heuristic* information on how evolution occurs (Campbell 1974). But finding pattern similarities requires an observer that selects or directs attention to some but not other data. Though they provide knowledge on evolved processes, patterns do not provide methodologies for life to evolve. At best, they provide methodologies for a scientific observer.

In sum, the distinction between organisms as methodologies or theories becomes superfluous. Organisms are methodologies that underlie theory or knowledge formation, and to explain the evolution of real organisms, we can only refer to processes that in turn refer to real phenomena. Real phenomena often have pattern similarity, although that might result from our observing eye that chooses to focus on some but not other data.

Human Knowledge, Like All Knowledge, Evolves

Finally, by expanding epistemology to all domains of life, classic evolutionary epistemologists have demonstrated that knowledge evolves. It evolves in the form of embodied theories (which are identical to real (holo)bionts) and in the form of disembodied (holo)bionts (which refer to classic human theories).

Human knowledge remains particular. Our linguistic theories evolve like biological (holo)bionts and demonstrate "universal symbiogenesis" by stitching and patching old ideas together into new ones (Gontier 2007). But many ideas remain unrooted in niches, or they are dissociated with the multiple realities life's biodiversity builds and embodies. At most, they are part of our brain, or we extend them into books or into an extended or global mind such as the internet.

Because many of our ideas and theories are unrooted, they are prone not to be true or only partially true. But that does not take away from the fact that they are real for those who believe in them, which is why they are so dangerous sometimes. Ideas are very powerful. We are a species that kills over ideas.

Instead of holding them true, we should remember that our outlook is limited by our current and historically grown knowledge and it is biased toward our particular bio-reality that contains our particular cultures and languages. Progress therefore depends upon comparing different views to one another and to finding alternative instrumental ways to look at our surroundings. Other (holo)bionts and different worldviews can help with that, and we can find a moderate progress in how we are catching up with the old realities of the living world.

Much depends upon preservation because many ideas, not only the bad ones, are lost. Ideas can only survive when they are continuously transmitted. But while today so much funding goes toward conservation of biodiversity, little attention is given to the conservation of valuable ideas. Instead, projects get funded based upon innovation which carries within it the idea that everything said and done before is false. It turns scholars away from the past in search for a future, while the past has brought us here and is therefore more real than what has yet to come. In Nietzsche's wake, we can sit back and be jolly about how much of science involves a reinvention of the wheel because it does not care for history.

The ancients knew the importance of the past, and they used it to understand the present and to predict the future. They were not wrong when they found *cyclic patterns* in the return of the planets, the seasons, and the constellations. They just understood it from within their geocentric worldview, wherefrom we have since evolved. They were also not wrong in finding returning cycles of coming and becoming or generation and decay, they just did not know that outside perturbations could alter the chain of events trough, e.g. mutations. But altering the chain or not, all life and perhaps the entire multiverse continues to generate and decay. That is most certainly true and one of the biggest insights that comes from the ancient schools. Much of ontogenetic, phylogenetic and paleontological work nowadays involves a return to research on recurring cycles (Gontier 2016b). We find them in how the Darwinian principles repeat each generation anew, in how DNA translates into proteins through RNA, how organs develop in the body, how circadian rhythms evolve, how holobionts form, and perhaps even speciation and extinction events follow recurring periodicities. Many of these cycles now take on the form of *networks*.

Judeo-Christians were also not wrong by understanding that many events that characterize history are unique. It made them linearize time and attempt to develop chronologies. Natural history scholars were also not wrong when they continued these traditions and mapped the history of life as going from genes, to single cells to multicellular (holo)bionts, from fish to reptiles and mammals. Where they went wrong is that they assumed that this linear sequence of events is fixed, because today we know that unicellular bionts can penetrate multicellular ones and create holobionts. Symbioses jump between lineages in spacetime, and viruses and bacteria appear to travel through spacetime at the blink of an eye. None of it requires a rewinding of past events. Instead, it demands concepts of downward and horizontal inter-hierarchy causation, as well as non-linear dynamics; and it shows that chronologies are one-sided views that "merely" focus on the historical trajectory of one particular dataset.

Natural history scholars were also not wrong when they said that European culture evolved from hunter-gathering to agriculture to industrial and technological communities. Where they went wrong is that they assumed this was a natural order or prototype by which all cultures evolved. What they should have done instead, is realize that much of our current society continues to depend upon agriculture and industry, and they should have analyzed the particular histories of other cultures, and compare them to one another (Pinxten 1997).

Much of the phylogenetic ramifications that occurred within the trees of life, languages and cultures can now be proven by gene comparisons of living (holo) bionts and even, in some cases, ancient DNA retrieved from fossils. One of the things that came out of that is that we humans not only are a single species, cultural and linguistic phylogenetics proves that we have never been isolated into a homeland, with a home language and a home culture. Human populations have always crossed paths. They exchanged genes, microbes, animals, plants, humans, ideas, words, and material artifacts. That there once used to be isolate cultures is a false nationalist idea of the 19th and early 20th century that has no ground in reality.

The modern synthesis was also not wrong when it claimed that evolution occurs by means of natural selection. The problem is that they only provide one side of the evolutionary story, and they failed to see how drift, macroevolutionary theory, symbiology, ecology, and ontogeny or epigenetics identify key players in evolution.

While a moderate case can be made for progress in the sciences, we also need to come to terms with the fact that there is not one truth out there. Truth changes with time and space, and epistemologically, a pluralistic account should be favored. The question is not who is right or wrong, but how distinct insights from different human and organismal cosmologies together provide a deeper understanding of the complex and multiple realities that life has evolved up until today, and how we can move forward from there. Pluralistic schools go back to scholars such as James (1909), the American anthropologists including Kroeber and Boas, and the American pragmatists.

Concluding Remarks

The consequences of accepting that epistemology or knowledge comes and goes with the (holo)bionts that evolve it, is that it questions the existence of a single world or level within that world that is more real. Instead, it recognizes reality and truth as variable and evolving over time. At one point in time, we know that earth was a dead planet, but ever since life evolved, that planet has changed inside out by the (holo)bionts that inhabit it. Life recycles the once dead planet into a living one.

Knowledge comes and goes with the organisms that contract and expand in space and time. What is true for one (holo)biont, is not true for another. These (holo)bionts extend their knowledge into their progeny and onto their surroundings thereby altering it into the currently living earth. The ontological state of the world changes inside out.

In such a cosmology, there is no room for adaptationist accounts, for uniformitarianism or physicalism. If there once was an abiotic physical world, we evolved from it. There is also no room for reasons or causes that explain why things are as they are. There are only processes that involve phenomena bringing forth other phenomena. Such a view most certainly has room for free will, one that takes into account all the living, if we want to.

Finally, whether or not the above outlines for a new cosmology are true or false, or better useful, I leave to the reader. In our everyday lives, we do not need metaphysics. One might even say that it is heavy on the mind and perhaps unhealthy. We should not roam in our thoughts too much. So proceed, all is well and everything around you is real, even if you don't see the cells that make you up, or the underlying particles they are composed of. They appear to know what they are doing, sometimes better than your conscious self. We have evolved to live in this world we see around us, and we can, with a significant amount of moderation, and most of all by considering that others are just like us, trust our bodies.

But what then, do we do when in doubt? Descartes, for example, in his period of doubt about the truths of the world, compared it to being stuck in a forest. To find your way out, and quite consistent with his mathematics, he advised to keep a straight line. Later he went on to say that he could think his ideas and that because he thought them, they were real. He also added they were a gift from a benign God. His *cogito ergo sum* brought forth the phenomenological and cognitive sciences.

Truth is that when you keep a straight line in a forest, you keep bumping into the trees that form part of that reality right then and there. You might slip into a pond, get chased by some animals, get bitten by ticks, end up with some kind of lifelong disease transmitted in the bite, and get soaked by a tropical thunderstorm our kind induced due to global warming. When in doubt, give it a try, that's how real it gets.

The new cosmology also comes with an invitation, to open our minds to other and new ideas, to learn and to show respect for other views, because we cannot make things happen on our own. And we should realize that ideas, however beautiful, can also be destructive. People kill each other and themselves in the name of ideas daily. But in the end, they are just ideas, real for you but not necessarily for someone else. They are furthermore prone to change over time and with the generations that think them and that will remodel them anyway, and that should be encouraged. The availability of alternative frames of reference brings forth flexibility in deploying them which is virtuous because it gives freedom. Socioculturally and politically, we can step outside our local niches and learn from others, try to get along and build a better future for us all.

Acknowledgements Written with the support of the Portuguese Fund for Scientific Research (grant ID SFRH/BPD/89195/2012 and project number UID/FIL/00678/2013). Cordial thanks to Shyam for inviting me to contribute to this inspiring volume, and to Michael Bradie for much appreciated comments on a previous draft.

References

B.P. Abbott, LIGO Scientific Collaboration and Virgo Collaboration et al., Observation of gravitational waves from a binary black hole merger. Phys. Rev. Lett. **116**, 061102 (2016)

E. Anderson, *Introgressive Hybridization* (Wiley, New York, NY, 1949)

H. Arendt, *The Human Condition* (University of Chicago Press, Chicago, IL, 1958)

J. Barnes, *The Complete Works of Aristotle* (Princeton University Press, Princeton, 1984)

G. Barsanti, *La scala, la mappa, l'arbero: Immagini e classificazioni della natura fra Sei e Ottocento* (Sansoni, Firenze (Italy), 1992)

W. Bechtel, *Philosophy of Science* (Lawrence Erlbaum Association, Hillsdale, NJ, 1988)

W. Bechtel, Mechanism and biological explanation. Philos. Sci. **78**, 533–577 (2011)

P. Bourdieu, *Outline of a Theory of Practice* (Cambridge University Press, Cambridge, MA, 1977)

M. Bradie, Assessing evolutionary epistemology. Biol. Philos. **1**, 401–459 (1986)

M. Bradie, *Evolution and Culture*. (2017) https://doi.org/10.1002/9780470015902.a0027505

D. Campbell, Evolutionary Epistemology, in *The Philosophy of Karl Popper*, vol. 1, ed. by P. Schilpp (La Salle, Chicago, IL, 1974), pp. 413–459

F. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. Science **311**(5765), 128–1287 (2006)

A. Clark, D. Chalmers, The extended mind. Analysis **58**, 7–19 (1998)

C.F. Craver, W. Bechtel, Top-down causation without top-down causes. Biol. Philos. **22**, 547–563 (2007)

C. Darwin, *The Origin of Species* (Murray, London, 1859)

R. Dawkins, Universal Darwinism, in *The Philosophy of Biology*, ed. by D. Hull, M. Ruse (Oxford University Press, Oxford, 1983), pp. 15–35

R. Dawkins, *The Extended Phenotype* (Oxford University Press, Oxford, 1989)

N. Eldredge, *Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought* (Oxford University Press, New York, 1985)

H. Everett, Relative state formulation of quantum mechanics. Rev. Mod. Phys. **29**(3), 454–462 (1957)

V. Faria, E. Sucena, Novel endosymbioses as a catalyst for fast speciation. In N Gontier (ed) Reticulate Evolution (Dordrecht, Springer, 2015) pp. 107–120

J.F. Ferrier, *Institutes of Metaphysic the Theory of Knowing and Being* (William Blackwood and sons, Edinburgh, 1854)

M. Ghiselin, A radical solution to the species problem. Syst. Zool. **23**(4), 536–544 (1974)

A. Goldman, *Simulating Minds* (Oxford University Press, Oxford, 2006)

N Gontier, *Evolutionary Epistemology. Internet encyclopedia of Philosophy* http://www.iep.utm.edu/evo-epis/ (2006)

N. Gontier, Universal symbiogenesis: an alternative to universal selectionist accounts of evolution. Symbiosis **44**, 167–181 (2007)

N. Gontier, Evolutionary epistemology as a scientific method: A new look upon the units and levels of evolution debate. Theor. Biosci. **129**, 167–182 (2010)

N. Gontier, Depicting the tree of life: the philosophical and historical roots of evolutionary tree diagrams. Evol. Educat. Outreach **4**(3), 515–538 (2011)

N. Gontier, Applied evolutionary epistemology: a new methodology to enhance interdisciplinary research between the human and natural sciences. Kairos **4**, 7–49 (2012)

N. Gontier, Reticulate evolution everywhere, in *Gontier N*, ed. by Reticulate Evolution (Springer, Dordrecht, 2015a), pp. 1–40

N. Gontier, Uniting micro- with macroevolution into an extended synthesis: Reintegrating life's natural history into evolution studies, in *Macroevolution*, ed. by E. Serrelli, N. Gontier (Springer, Dordrecht, 2015b), pp. 227–275

N. Gontier, Guest-editorial introduction: converging evolutionary patterns in life and culture. Evol. Biol. **43**(4), 427–445 (2016a)

N. Gontier, Time: the biggest pattern in natural history research. Evol. Biol. **43**(4), 604–637 (2016b)

N. Gontier, What are the levels and mechanisms/processes of language evolution? Lang. Sci. **63,** 12-43 (2017)

N. Gontier, Pattern Similarity in Biological, Linguistic and Sociocultural Evolution. EVOLANG 12 Proceedings (World Scientific, 2018)

N. Gontier, M. Bradie, in *Acquiring Knowledge on Species-Specific Biorealities: The Applied Evolutionary Epistemological Approach*, ed. By R. Joyce. The Routledge Handbook of Evolution and Philosophy (Routledge, 2018)

S.J. Gould, R. Lewontin, The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. Proc. Royal Soc. London Ser B Biol Sci **205**, 581–598 (1979)

S.J. Gould, *Ontogeny and Phylogeny* (Belknap Press, Cambridge, MA, 1977)

S.J. Gould, *Wonderful Life* (W. W. Norton, New York, 1989)

P. Griffiths, R. Gray, Developmental systems and evolutionary explanation. J Philos. **91**, 277–304 (1994)

E. Haeckel, *Athropogenie oder Etnwickelungsgeschichte des Menschen* (Engelmann, Leipzig, 1874)

B. Hallgrimsson, B. Hall (eds.), *Epigenetics* (University of California Press, Berkeley, CA, 2011)

D. Hull, *Science as a Process* (University of Chicago Press, Chicago IL, 1988)

E. Jablonka, M. Lamb, *Evolution in Four Dimensions Cambridge* (MIT Press, MA, 2006)

W. James, *A pluralistic universe* (Longmans, London, 1909)

T. Kuhn, *The Structure of Scientific Revolutions Chicago* (Chicago University Press, IL, 1962)

I. Lakatos, *The Methodology of Scientific Research Programmes Cambridge* (Cambridge University Press, MA, 1978)

K. Laland, J. Kumm, M. Feldman, Gene-culture co-evolutionary theory. Curr. Anthropol. **36**, 131–146 (1995)

L. Laudan, Progress and its Problems: Towards a Theory of Scientific Growth. University of California Press (1977)

R. Levins, R.C. Lewontin, *The Dialectical Biologist* (Harvard University Press, Cambridge, MA, 1985)

R. Lewontin, The levels of selection. Annu. Rev. Ecol. Syst. **1**, 1–18 (1970)

R. Lewontin, Adaptation. Sci. Am. **239**, 157–169 (1978)

R. Lewontin, Organism and environment, in *Learning, Development and Culture*, ed. by H. Plotkin (Wiley, New York, NY, 1982), pp. 151–170

R. Lewontin, *The Triple Helix* (Harvard University Press, Cambridge, MA, 2000)

K. Lorenz, Kant's lehre vom apriorischen im lichte gegenwärtiger biologie. Blätter für Deutsche Philosophie **15**, 94–125 (1941)

A.O. Lovejoy, *The Great Chain of Being* (Harvard University Press, Cambridge, 1936)

J.E. Lovelock, L. Margulis, Atmospheric homeostasis by and for the biosphere: the Gaia hypothesis. Tellus, Series A, Stockholm: International Meteorological Institute **26**(1–2), 2–10 (1974)

L. Margulis, D. Sagan, *What is Life?* (University of California Press, Berkeley, CA, 2000)

L. Margulis, Symbiogenesis and symbionticism, in *Symbiosis as a Source of Evolutionary Innovation*, ed. by L. Margulis, R. Fester (MIT press, Cambridge, MA, 1991), pp. 1–14

J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Oxford University Press, Oxford, NY, 1995)

J. Maynard Smith, Group selection and kin selection. Nature **201**(4924), 1145–1147 (1964)

M. McLuhan, *Understanding Media* (McGraw Hill, New York, NY, 1964)

A. Mesoudi, Cultural evolution: A review of theory, findings and controversies. Evol. Biol. **43**(4), 481–497 (2016)

P. Munz, *Philosophical Darwinism* (Routledge, London, 1993)

F. Odling-Smee, Niche Constructing Phenotypes, in *The Role of Behavior in Evolution*, ed. by H. Plotkin (MIT, Cambridge, MA, 1988), pp. 73–132

S. Okasha, Multilevel selection and the major transitions in evolution. Philos. Biol. **72**, 1013–1025 (2005)

M. Pigliucci, An extended synthesis for evolutionary biology. The year in evolutionary biology 2009. Ann. N. Y. Acad. Sci. **1168**, 218–228 (2009)

R. Pinxten, *When the Day Breaks* (Peter Lang, Frankfurt am Mein, 1997)

K. Popper, *The Poverty of Historicism* (Routledge, London, 1957)

K. Popper, *Conjectures and Refutations* (Routledge & Keagan, London, 1963)

K. Popper, *Objective Knowledge* (Clarendon Press, Oxford, 1972)

K. Popper, Critical Remarks on the Knowledge of Lower and Higher Organisms, the So-Called Motor Systems, in *Sensory Motor Integration in the Nervous System*, ed. by O. Creutzfeldt, R. Schmidt, W. Willis (Verlag, Dordrecht, 1984), pp. 19–31

I. Prigogine, *From Being To Becoming* (Freeman, New York, 1980)

W.V.O. Quine, Two dogmas of empiricism. Philos. Rev. **60**, 20–43 (1951)

C.R. Reid, T. Latty, A. Dussutour, M. Beekman, Slime mold uses an externalized spatial memory to navigate in complex environments. PNAS **109**(43), 17490–17494 (2012)

R. Riedl, *Biologie der Erkenntnis: Die stammesgeschichtlichen Grundlagen der Vernunft* (Parey, Berlin, 1980)

A. Riegler, Like cats and dogs: radical constructivism and evolutionary epistemology, in *Evolutionary Epistemology, Language and Culture*, ed. by N. Gontier, et al. (Springer, Dordrecht, 2006), pp. 47–65

M. Rowlands, *The New Science of The Mind: From Extended Mind to Embodied Phenomenology* (MIT Press, Cambridge, MA, 2010)

M. Ruse, The View from Somewhere: A Critical Defense of Evolutionary Epistemology, in *Issues in Evolutionary epistemology*, ed. by K. Hahlweg, C.A. Hooker (State University of New York Press, Albany, NY, 1989), pp. 185–228

M. Sahlins, *Evolution and Culture* (University of Michigan Press, Ann arbor, MI, 1970)

E. Sapir, Do we need a ‘superorganic’? Am. Anthropol. **19**, 441–447 (1917)

B.F. Skinner, The evolution of verbal behavior. J. Experiment. Anal. Behav. **45**, 115–122 (1986)

L. Smolin, *The Life of the Cosmos* (Oxford University Press, New York, NY, 1997)

H. Spencer, *The Principles of Sociology* (Williams and Norgate, London, 1876)

I. Stewart, *The Mathematics of Life* (Basic books, New York, NY, 2011)

Stich From Folk Psychology to Cognitive Science (Cambridge, MA, MIT Press, 1983)

I. Tëmkin, N. Eldredge, Networks and hierarchies: approaching complexity in evolutionary theory, in *Macroevolution*, ed. by N. Gontier, E. Serrelli (Springer, Dordrecht, 2015), pp. 227–275

S.H. Toulmin, Understanding Princeton. NJ: Princeton University Press translated by J. Churchill. (The Hague: Martinus Nijhoff, 1972)

L. Van Valen, A new evolutionary law. Evol. Theor. **1**, 1–30 (1973)

L.P. Villarreal, G. Witzany, Viruses are essential agents within the roots and stem of the tree of life. J. Theor. Biol. **262**(4), 698–710 (2010)

T. Volk, *Quarks to culture: how we came to be* (Columbia University Press, New York, NY, 2017)

A.N. Whitehead, *Process and reality, Gifford lectures 1927–1928* (Cambridge University Press, Cambridge UK, 1929)

R.A. Wilson, Collective memory, group minds, and the extended mind thesis. Cogn. Process. **6**, 227–236 (2005)

G. Witzany, *Biocommunication of Animals* (Springer, Dordrecht, 2014)

L. Wolpert, *How We Live and Why We Die* (W. W. Norton & Company, New York, NY, 2009)

F. Wuketits, Evolutionary epistemology: The non-adaptationist approach. In N. Gontier, van Bendegem JP Aerts D (eds) Evolutionary Epistemology, Language and Culture pp. 33–46 (Dordrecht, Springer, 2006)

V.C. Wynne-Edwards, *Evolution Through Group Selection* (Blackwell, Odford, 1986)

O. Zhaxybayeva, W.F. Doolittle, Lateral gene transfer. Curr. Biol. **21**(7), R242–R246 (2011)

Chapter 31
Quantum Perspectives on Evolution

Diederik Aerts and Massimiliano Sassoli de Bianchi

Introduction

The key notion of *evolution* is employed with quite different meanings in the different fields of investigation. Let us observe first how it is usually understood in biology, and more particularly in Darwinian evolutionary theories. In those ambits, the term evolution refers to a very specific process of change of the transmissible characteristics of biological populations (unfolding over successive generations), produced by so-called *natural selection*.

Natural selection is based on deterministic criteria of reproductive advantages (the so-called survival of the fittest) but involves also elements of unpredictability in its functioning. Indeed, the mutations that can arise in the genome of the individual organisms, which can be inherited by the offspring and give rise to the variants on which selection operates, are assumed to happen in a random way.[1] So, evolution, according to the Darwinian view, is a selection process operating on randomly generated variations of the nucleotide sequence of the genomes, i.e., taking place on

[1] More precisely, in a Darwinian process, evolution occurs because selection affects the distribution of the randomly generated heritable variation across generations. This doesnt mean, however, that the variation in the biological world is always assumed by biologists to be perfectly random. We now know that there are many processes (assortative mating, biased mutation, etc.) that can cause variation to be non-random, but in many cases it is nevertheless considered to be “random enough” for a Darwinian model to apply.

D. Aerts (✉)
Center Leo Apostel for Interdisciplinary Studies and Department of Mathematics,
Brussels Free University, Brussels, Belgium
e-mail: diraerts@vub.ac.be

M. Sassoli de Bianchi
Center Leo Apostel for Interdisciplinary Studies, Brussels Free University,
Brussels, Belgium
e-mail: msassoli@vub.ac.be

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_31

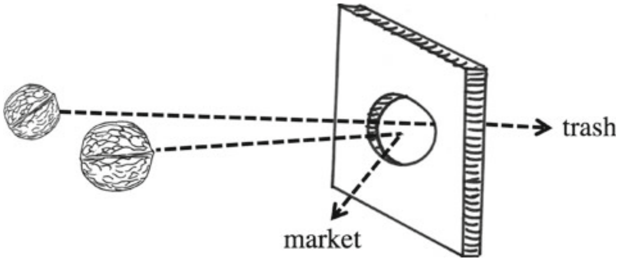


Fig. 31.1 Depending on its size, a walnut will either pass through the hole, ending up in the trash, or not pass through the hole, ending up on a market stall

forms of matter-energy having already actualized those properties that can affect the fitness of the organisms in question.

To make an elementary example, consider a basket of walnuts. The Darwinian selection is then merely like a process where one would select those who are sufficiently big, according to marketing standards, and eliminate those who are too small, where sufficiently big (respectively, too small) means for instance to possess the property of not passing (respectively, passing) through a screen having a round hole of a given diameter (see Fig. 31.1). Since we said "merely", one may ask: On what else a selection could be based, if not on properties/variants that have already become manifest?

To answer this question, we move to another scientific domain, physics, where the term evolution is used with quite a different meaning. In physics, a physical entity (or physical system) is said to evolve when its state changes in a way that can be predicted (at least in principle) by solving the corresponding dynamical equations. Think for instance of Newton's second law of motion, for classical entities, or Schrödinger's wave equation, for (non-relativistic) quantum entities. Of course, for a change of the state to be predictable one needs to know exactly what are the force fields acting on the entity, as well as its initial state. Indeed, if for instance the entity's initial state is only known in probabilistic terms, then also future states will be generally known only in probabilistic terms.

However, with the advent of quantum mechanics, physicists were forced to recognize that there are other physical processes that can produce a change of the state of a physical entity, only describable in probabilistic terms, even when the initial state of the entity and the force fields acting on it are perfectly known. For historical reasons, the term evolution was not used to denote these additional processes; instead, they were called *measurements* and were generally understood as observational processes of physical properties associated with the entity, like those of being in a given spatial location, having a certain range of energies, or of velocities, a given spin value, etc.

When a quantum entity is in a so-called superposition state with respect to a measurable physical quantity (called *observable*, in the quantum jargon), there will be an irreducible unpredictability associated with the outcomes of the measurement/observational process. More precisely, if the pre-measurement state is a *superposition* of, say, *n* states having well-defined different values for the observable in question, then, prior to the measurement, there will be *n* distinct possibilities for the outcomes of the measurement. These possibilities refer to potential properties, i.e., properties that have not yet been actualized, but can and will be actualized (only one of them) during the measurement.

Therefore, in a quantum measurement we are also dealing with a selection process, but very different from a Darwinian one, as it does not operate at the level of the actual properties.[2] Instead, we have a selection operating directly at the level of the potential properties. Also, these potential properties do not refer to a situation of lack of knowledge about which property would be actual or not, prior to the measurement, as not only the (superposition) state of the physical entity is assumed to be perfectly known, but also to fully describe the entity's state of affair. Thus, a quantum measurement is a genuinely unpredictable process where the indeterminacy results from the fact that one must *break the symmetry of the potential*, forcing the physical system to acquire properties that were truly non-actual prior to the measurement; see Aerts and Sassoli de Bianchi (2014) and the references cited therein.

Coming back to Darwinian's natural selection, we now have a first possible element of understanding of why it describes a selection process which is of a very special kind, and therefore also of a very limited kind, so that one might wonder if it is sufficient to describe all biological processes of change (Ogryzko 1997; Gabora and Aerts 2005a, b; Aerts et al. 2006, 2011; Gabora et al. 2013; Aerts and Sozzo 2015).[3] Could it be that many of them would in fact be more general processes of actualization of potential properties? And what would be the consequences of that for our understanding of life in general?

Cracking Walnuts

Before digging further in these questions, let us consider a first possible objection, which is the following: for all practical purposes, quantum mechanics is known to only apply at the micro-level. When quantum effects are nevertheless observed at the macro-level (as in superconductivity, superfluidity, Bose-Einstein condensates, etc.), this is because the environmental conditions are of a very special kind, and more precisely such that they can offer an efficacious shielding from all influences that can provoke the decoherence of the system, like those resulting from the thermal photonic bombardment. As is well known, organisms living at the surface of our planet are constantly subjected to such intense thermal bombardment, so that for

[2] When using the term *selection* in relation to a process of actualization of an outcome, it should be understood in its more common sense of choosing one of several possibilities, and not in the specific sense of a natural selection, i.e., referring to processes where environmental factors can affect the distribution of randomly generated heritable variation across generations.

[3] Forms of change that are not explained by natural selection are also considered nowadays by biologists, like in evolutionary developmental biology or epigenetics. However, our focus in the present essay, also for the sake of simplicity, is on the central mechanism of Darwinian's natural selection.

them quantum mechanics would simply not apply (not at the macro-level at least). Thus, following this line of thought, processes of selection that are important for biological evolution should only be of the *selection of actual properties kind*, and not of the *selection through actualization of potential properties kind*.

To see that this objection only captures one aspect of the problem, one needs to consider that the notion of *quantumness*, when closely inspected, is more related to how systems are organized and behave in relation to the different contexts with which they interact than the fact that they would have a small or big spatial size, i.e., that they would be micro or macro systems, according to the usual understanding of these notions (Aerts and Sozzo 2015).[4]

Let us come back to the example of the basket of walnuts. We mentioned that we can measure their size. These measurements are of a purely discovery kind, as is clear that each walnut in the basket is either too small or sufficiently big (according to the previously adopted definition), i.e., either it possesses the too small property, or the opposite sufficiently big property, and this even before one tries to make it pass through the round screen hole (which symbolizes here the Darwinian natural selection of actual properties process). Variations that previously randomly occurred, produced the variety of walnuts sizes in the basket, on which the selection operates in a perfectly deterministic way: those having the *sufficiently big* property (not passing through the hole) can make their route to the market stalls, whereas the others (passing through the hole), considering our consumer society, will end up in the trash.

Other properties can however be considered, in association with the walnuts, which might be also relevant for their persistence (as a variety) on the market stalls. Indeed, if it is true that a selection can be operated by the merchants at the level of the market stalls, another selection can be operated by the consumers, based on different criteria. Indeed, if consumers prefer certain walnuts, based on these different criteria, they will continue to purchase them, favoring their maintenance or development on the market stalls. Let us consider as an example what we can call the *cracking well* property (Aerts 1999). By this we mean the property of a walnut to lend itself to be easily cracked, that is, cracked in a way that the shell breaks properly, without creating fragments, so that the walnut's kernel can be easily separated from its shell and eaten.

Can a consumer know in advance (i.e., before performing a test/measurement) if a given walnut possesses or not the cracking well property? Anyone having some experience in the process of cracking walnuts immediately understands that the situation is very different compared to the previous measurement of the size of the walnuts. Indeed, to know about a walnut's cracking well property the consumer must crack the walnut (i.e., perform in practice the measurement), and only by witnessing the result s/he will be able to tell if the cracking well property has been successfully tested (the shell and its kernel separate well) or not (the shell and its content get mixed in multiple fragments); see Fig. 31.2.

[4] Micro-systems are not necessarily spatial systems, as we will explain later in the article. In other words: micro is not necessarily small.

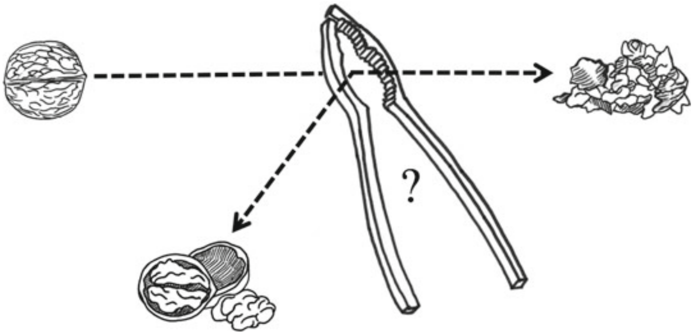


Fig. 31.2 When subjected to the nutcracker measurement, a walnut, independently of its size, can either crack well or crack bad, and the outcome cannot be predicted in advance, not even in principle

The reason why the outcome is now irreducibly unpredictable is that the cracking well and cracking bad properties are created by the very operations that are performed to test them. Indeed, when cracking a walnut, a human being cannot control all the fluctuations produced by their hands, the exact way in which the walnut is placed in the nutcracker, how their muscles' power is precisely applied and transferred to the walnut by the instrument, etc. Therefore, if we want to describe the reality of the walnut in relation to its crackability, we will have to describe its state as a *superposition of cracking well and cracking bad states*.

What is interesting to observe is that also in this case there is a selection process; however, it is a process that cannot be operated without disturbing the walnuts (to use Einsteins celebrated way of defining an element of reality), as there is no way to predict in advance the outcome of the crackability measurement. Thus, the selection process does not operate in this case by *discovering and then distinguishing* the crackable walnuts from the uncrackable ones, but by literally and directly *creating* either the crackable or the uncrackable ones. In other words, it is a selection of the actualization of potential properties kind.

Cracking walnuts is usually not considered to be a quantum measurement process. However, this is so only because we are used to conceive observations as resulting from processes of discovery, and not also from processes of creation, able to change the state of the observed entity (so that the entity needs not to possess the observed property prior to its observation). This prejudice is one of the reasons why quantum measurements remain so counterintuitive for many physicists to this day, as they are typical of observations containing not just aspects of discovery, but also, and especially, of creation, as evidenced by the very quantum *projection postulate*, describing how a state change following a measurement.

Quantum Machines

Another reason why the cracking walnut experiment is not commonly considered to be a quantum process, is that it does not possess all the symmetries that typically characterize experiments with micro-entities. However, it is possible to conceive a simplified version of it that is perfectly isomorphic to a pure (two-dimensional) quantum system. In other words, it is possible to conceive (and to some extent construct) a genuine macroscopic *quantum machine*, working at room temperature, as we are now going to illustrate (Aerts 1986; Aerts and Sassoli de Bianchi 2014).

In this simplified version of the experiment, we are only interested in the relative movement of the nut with respect to its shell, and instead of a whole nut, we only consider a tiny fragment of it, which we assume to be initially attached to some internal point of the shell, which in our idealization we otherwise consider to be empty. Also, we model the effect of the cracking by means of two different sequential movements of the nut fragment. First, a movement with which it penetrates the shell, in the direction of the force exerted by the nutcracker. Second, a movement in a direction orthogonal to the latter.

The first movement is assumed to be deterministic, and to somehow account for the direct compression effect produced by the nutcracker. The second movement is instead assumed to be indeterministic, and to account for the symmetry breaking effect associated with the way the shell breaks. To describe in a simplified way this second movement, we model the intricate dynamics of the breaking shell by the simpler dynamics of the breaking of an elastic band. More precisely, we consider that there is a breakable and uniform elastic band stretched inside the empty shell, along the direction of this second movement (see Fig. 31.3).

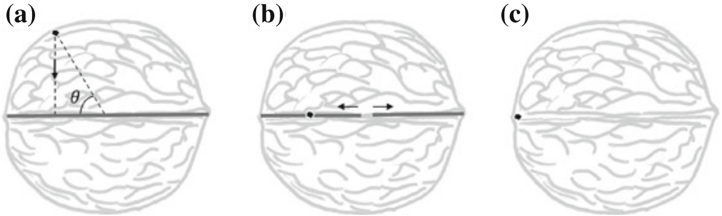


Fig. 31.3 **a** The nut fragment orthogonally moves towards the elastic band and remains stuck to it; **b** the elastic band breaks at some unpredictable point and collapses; **c** the collapse of the elastic brings the nut fragment towards one of the two end points, here the left one. If one calculates the probabilities for the fragment to be pulled towards the left end, assuming the elastic to be uniform, one finds the formula: $P_{\text{left}} = \frac{1}{2}(1 + \cos\theta)$, so that the probability to be pulled to the right is: $P_{\text{right}} = \frac{1}{2}(1 - \cos\theta)$. These are exactly the quantum probabilities associated with the measurement of a 2-dimensional quantum entity (like a spin-$\frac{1}{2}$ entity), with the nutshell's spherical shape representing the 3-dimensional *Bloch sphere of states*, and the nut fragment the point representative of the state of the entity within the latter

So, we have that the nut fragment first orthogonally moves towards the elastic band, remaining attached to it, then the elastic breaks, and by collapsing it brings the nut fragment towards one of its two end points (see Fig. 31.3). Adding the assumption that if the nut fragment is drawn to the left (resp., to the right), this corresponds in our idealization to the *crack well* situation (resp., *crack bad* situation), we obtain in this way a simplified description of the cracking walnut experiment.

Clearly, also in this idealized description the left (crack well) and right (crack bad) outcomes cannot be known in advance, as one cannot predict in advance in which point the elastic is going to break. However, it is possible to calculate the outcomes' probabilities. The calculation involves some simple trigonometry, taking into consideration the initial orientation of the fragment relative to the elastic, and surprisingly the obtained probabilities are perfectly isomorphic to those associated with an archetypal quantum experiment: the measurement of the value of a spin-$\frac{1}{2}$ entity relative to a given spatial direction (see Fig. 31.3). More precisely, the orientation of the nut fragment can be shown to be in a one-to-one correspondence with the state of the spin-$\frac{1}{2}$ entity, the orientation of the elastic band with that of the *Stern-Gerlach measuring apparatus*,[5] and the calculated probabilities to precisely correspond to those predicted by the quantum mechanical *Born rule* (Aerts 1986, 1999; Aerts and Sassoli de Bianchi 2014).

Let us also consider for a moment the possibility of combining different quantum machines. For instance, we can assume that two of them have their nut fragments located at the center of their respective spherical shells, and that they are connected by means of an extendable rigid rod, rotating around a pivot (see Fig. 31.4). Remarkably, one can show that this bipartite mechanistic entity behaves exactly like two spin-$\frac{1}{2}$ *entangled* entities in a so-called *singlet state*, and is thus able to violate *Bell's inequalities*[6] with the same $2\sqrt{2}$ numerical value as the latter do (Aerts et al. 2000; Aerts and Sassoli de Bianchi 2016a).

Clearly, the entanglement of the two nut fragments is made here manifest by the presence of the rigid rod, correlating their movements within the two empty shells, whereas for two microscopic spin entities one cannot generally represent their entanglement by means of a connection operating inside our 3-dimensional spatial theater. This *non-spatiality* (and consequently *non-locality*) of quantum entities is certainly a key aspect of their nature, which distinguishes them from the 3-dimensional quantum machines we humans can design and construct inside our Euclidean space. Precisely because of the restricted dimensionality of these mechanical machines, not all quantum systems can be simulated by them. However, these machines, despite all their limits, reveal an important (hidden) aspect of the quantum formalism: that quantum

[5] In a SternGerlach experiment the spin entities are sent through an inhomogeneous magnetic field, to observe their deflection, as revealed by observing the distribution of their impacts on a detector screen.

[6] Bell's inequalities express certain constraints that must hold when measurements are performed on composite systems, if one assumes that the components are *experimentally separated*, i.e., that they cannot influence each other, when measurements are performed on them.

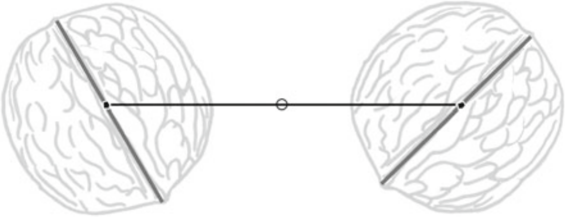


Fig. 31.4 Two nut fragments, initially placed at the center of their respective shells, are connected through an extendable rigid rod. Imagine that the elastic of the left fragment breaks first, in a way that the latter is pulled upwards. Then, because of the rod-connection, the right fragment will be forced to acquire the diametrically opposite position in its shell. If one assume that the rod-connection is then disabled, and that the right fragment orthogonally falls onto its elastic band, which then will subsequently break and draw the fragment to one of its end points, one can show that the resulting process is perfectly equivalent to a *product measurement* on a so-called (entangled) *singlet state*

probabilities, which violate *Kolmogorov's axioms*,[7] can be understood as resulting from processes where there are some uncontrollable fluctuations in the measurement context, which explain the indeterminacy of the measurements' outcomes.

Quantum machines also reveal that the statistical correlations obtained when measurements are performed on entangled entities can be understood as resulting from the existence of *actual connections* between them. All the mystery about entanglement should therefore be attributed to the fact that these connections, for some physical entities, cannot be represented (and therefore remain invisible) in our spatial theater, which is the reason why they appear to behave as interconnected composite entities, despite of the fact that there can be a spatial separation between the different components and nothing connecting them through space.

Non-universal Measurements

It is instructive to consider again the full, non-simplified cracking experiment. In what it differs from its quantum machine idealization? Clearly, the system is much more complicated and its geometrico-dynamical description will necessarily be much more elaborated. We still have a deterministic aspect involved, because of the constrained action of the nutcracker, which can only move on a specific plane, and of course, we also have an indeterministic aspect, associated with the unpredictable dynamics of the breaking shell, which replaces that of the idealized uniformly breakable elastic band. It is important to observe, however, that a person will each time operate the nutcracker in a different way. Also, not all *ways* to use the nutcracker are

[7] Kolmogorov's axioms, named after the Russian mathematician *Andrey Kolmogorov*, are three assumptions providing a precise mathematical formalization of (classical) probability theory.

available to be operated by a person, or group of persons. For instance, robots can operate a nutcracker in ways that are impossible for humans, and vice versa.

Considering once more the idealized elastic band representation, each way of operating the nutcracker on a walnut should be associated with a different non-uniform elastic band, i.e., an elastic for which not all points have the same probability to break. Moreover, as we cannot know in advance what will be the *way of cracking* selected by the human operator, to calculate the outcome probabilities we must now consider an average over all available ways. If these available ways correspond to all possible ways, then we obtain what is called a *universal average*, and the associated measurement is called a *universal measurement*.

There is an interesting result, which we will only mention here in passing without entering into details, according to which the Born rule of standard quantum mechanics can be understood as being precisely the expression of a universal average *over all possible ways of selecting an outcome* (Aerts and Sassoli de Bianchi 2014, 2015, 2016a, b, 2017a). However, when not all possible ways are available to be actualized, which is the typical situation in a real cracking nut experiment, we are in the more general (less symmetrical) situation of a non-universal measurement, and probabilities different than those predicted by the Born rule of quantum mechanics will be obtained. These are still non-classical probabilities, in the sense of not obeying Kolmogorov's axioms, but also non-quantum, in the sense of not being purely quantum.

In other words, the full cracking experiment describes a *quantum-like* process, that is, a process that while exhibiting many of the fundamental quantum features, cannot be handled by the standard quantum formalism. But the fact that cracking a walnut is typically a non-universal measurement is not the only reason why the standard quantum formalism does not apply. Another complication is introduced by the description of the *state space* of the measured entity, which is now the whole nut. When only considering an infinitesimal fragment of it, one can describe it as a point inside a 3-dimensional sphere, which in view of the so-called *Bloch representation of states* can be put in a one-to-one correspondence with the *Hilbert space geometry* of standard quantum mechanics. But the state of a whole nut, which can break into multiple pieces, cannot be described by a single point, so the geometry of its state space is more complex than that derived from a Hilbert space.

There are of course numerous other aspects in a full cracking experiment that makes it not describable by the standard (Hilbertian) quantum mathematics. But more general formalisms also exist. For instance, the present authors recently proposed a *general tension-reduction* (GTR) *model*, where non-universal measurements and non-Hilbertian state spaces can also be considered, thus generalizing in a natural way the more specific Hilbert model of quantum mechanics; see (Aerts and Sassoli de Bianchi 2016a) for a step by step construction of the model. But independently from the mathematical model one can use, what is important to understand, in pursuing our discussion, is that there is a fundamental element of quantumness in the cracking experiment that is the same as the one present in, say, a typical Stern-Gerlach experiment: the presence of a *weighted symmetry breaking process*, bringing the potentiality level into actuality.

Context Driven Actualization of Potential

After this quantum (and quantum-like) excursus, let us come back to our initial point about the possible shortcomings of the Darwinian natural selection account. Our hypothesis is that selection arises in nature both as selection of actual properties and selection through actualization of potential properties, and that in fact the former can be understood as a special case of the latter. More precisely, we can generally think of the evolution of an entity as resulting from its interaction with a *context*. Evolution, in other words, would be primarily a *process of actualization of potential properties under the influence of different contexts*, able to exert, in a sequential manner, their influence (Gabora and Aerts 2005a, b; Aerts et al. 2006, 2011).

Among them, we have the indeterministic ones, usually denoted (quantum) *measurements* in physics (although physicists, for historical reasons, do not think of them as contexts), which as we explained with the example of the quantum machines do not apply only to microscopic entities; and we have the deterministic ones, usually not called measurements, but simply *evolutions*, where the process of change is usually assumed to be continuous.

In quantum mechanics, the first possibility is described by the *projection postulate* and the associated *Born rule* of probabilistic assignment, and the second one by considering a so-called time-dependent *unitary evolution operator*, obeying the Schrödinger's (or Dirac's) equation, which acts on the entity's state to describe its change over time. Note however that, in ultimate analysis, a deterministic change of state can also be conceived as a measurement, and more precisely as a measurement having just one possible outcome, so that a continuous state change can also be described as a recursive application of these special one-outcome measurement processes. It is worth observing that this view of change, as a *context driven actualization of potential* (CAP) (Gabora and Aerts 2005a, b), describing the evolution of entities as processes of actualization of potential properties through reiterated interactions with multiple contexts (Aerts 2002), allow for the construction of more general models for the variations of forms than those permitted by the standard Darwinian view. As we mention already, processes of change produced by CAP can generate non-Kolmogorovian distribution of probabilities that can be very different from the classical distributions of chance of Darwinian evolution, and this because of the interplay between different mutually incompatible contexts. By mutually incompatible we mean that a state describing a condition of actuality for a given context (called *eigenstate*, in the quantum jargon) will become a state of potentiality (a superposition state) relative to another context, or evolve into a state of potentiality under the influence of another context.

Think of an entity like a hydrogen atom, which at a given moment would have been localized by means of that indeterministic context called *position measurement*. If the atom was a classical corpuscle, say at rest, it would remain in that position for all times thereafter. However, because of its quantum behavior, there will be an almost sudden spreading of its *wave-packet*[8] (consequence of the incompatibility between the position and momentum observables/contexts), with the result that all sorts of transitions to position states that would be forbidden according to classical evolution become available, in extremely short times.[9] In other words, assuming an alternation between different contexts, it becomes possible to describe evolutionary processes going from one eigenstate to another (classically forbidden) eigenstate of a same context, passing through the "backdoor" of the potentiality states; these are clearly processes that transcend those of the random variation and selection upon fitness kind, because in no way they can be described by them.

To better illustrate our point, let us come back once more to the walnuts example. We can consider the interplay between two contexts: that of the merchants, with their market stalls, and that of the consumers, with their dining tables. Consider a sufficiently big walnut. By definition, its probability of not passing through the screen with the round hole is equal to one. Hence, such walnut is in a state which is an eigenstate relative to the merchant context, which here symbolizes the selection upon fitness spatial context of Darwinian evolution. On the other hand, a sufficiently big walnut, when viewed from the perspective of the consumer context, can be described as a state of superposition between the cracking well and cracking bad eigenstates.

Now, a sufficiently big walnut, passing with certainty the deterministic size test might not necessarily pass the indeterministic crackability test, a process that only operates at the level of the consumers and therefore remains invisible (hidden) to the merchants. However, if a variety of sufficiently big walnuts end up having a too high probability of lending themselves to the cracking bad outcome, this will affect the consumers behaviors, who might change their habits and favor a different variety of walnuts.

Merchants that are unaware of the existence of the consumers contexts, will then witness a decrease, or disappearance, of certain varieties of walnuts from their stalls (as not purchased anymore by consumers) and a sudden increase, or appearance, of others, without any actual selection having being performed at the merchant level.[10] In other words, the walnuts entities can evolve in ways that depend not only on selections operating on their actual shapes (the spatial level of *phenotypes*, for biological organism), but also on the not yet actualized (potential) shapes (the cracked well and cracked bad shapes, in our example).

[8] "Wave-packet" is just another term for "quantum state," employed when the latter is expressed as a function of the position coordinates.

[9] Just to give an example, the wave-packet of (the center of mass of) a hydrogen atom, initially localized in a sphere of *Bohr radius* (approximately half of an angstrom), will typically spread to distances of tens of kilometers in a matter of one second.

[10] Note that our narrative remains consistent with the fact that a walnut, when in a cracked bad state, can pass through the round hole, whereas a walnut in a cracked well state will generally not.

Missing Fossils

It is instructive at this point to briefly put our discussion about CAP in the perspective of one of the problems of biological evolution: the observed general discontinuity (gaps) in the *fossils' records*. If we assume that the mechanism of selection of actual properties is just a special case of a more general mechanism of selection through actualization of potential properties, i.e., that biological evolution obeys also quantum-like laws, then we must distinguish the spatial (actual) life forms, which are entities in eigenstates relative to Earth's context, and are able to leave traces in the form of fossils, from the non-spatial (potential) life forms, which being in superposition states cannot leave traces, until they collapse toward a spatial state.

In the walnuts example, we dont see the cracked well walnuts on the merchants stalls (nor the cracked bad walnuts in the merchants trashes); however, the different varieties of walnuts in superposition states do compete together, via the consumers experiences, in a way that in the end will produce the disappearance of a variety and the appearance of another variety on the stalls. In other words, non-actualized lifeforms will also undergo evolutionary processes, but in ways that cannot leave visible traces at the level of Earth's context, and this could provide an explanation for the multiple gaps observed in the fossils' records (Aerts et al. 2006).

These gaps would be characteristics of evolutionary periods during which most of the evolutionary dynamics involve superposition (non-spatial) states, whereas the presence of a fossil would indicate a moment of sudden reduction (collapse) onto a spatial state, describing a new life form. Of course, gaps in fossils records need not to be explained only in this way. For instance, according to *punctuated equilibrium*, Earth's context will also experience possible sudden changes, exerting an environmental pressure that can explain the appearance of short-lived transitional species, having a smaller probability of yielding observable fossils (not all dying creatures become fossils, as special conditions are needed for fossilization to take place, which is a rare event). But punctuated equilibrium may not be sufficient as an explanation for all the observed gaps, and the hypothesis that quantum-like processes would also take place, however speculative, may turn out to be necessary to explain all the discontinuities in the traces left by organisms on the surface of our planet, during their long evolution.

There would be much to add regarding how unusual and limited Darwinian evolution really is, if viewed from the larger perspective of CAP processes. The neglecting of what happens at the level of the potential states has consequences from a mathematical/structural viewpoint, as non-Kolmogorovian (quantum-like) probability models allows for the modeling of a much larger number of experimental data. It has consequences also in the way the very notion of selection can be understood, at a fundamental level. If selection operates at the level of actual (spatiotemporal) properties, then variety needs to be manifested in the population and its descendants. On the other hand, if selection also operates at the level of potential (non-spatiotemporal) properties, through superposition states, then every individual in the population carries with it a variety of different (potential) traits, and selection mechanism also happen at the level of the single individuals, taking also into account their personal stories (which are also contexts that can change an individual state), so that a *Lamarkian-like* aspect would also be inevitably part of a quantum model of evolution.

In the absence of more specific models, the criticism to the Darwinian view that we have here presented certainly remains very abstract and speculative. But to work out realistic models with a predictive power is extremely difficult, because of the problem in enumerating in advance the different adaptive functions of interest, giving rise to the quantum observables (the contexts) influencing the evolution of the organisms of interest (Gabora et al. 2013). In other words, considering the complexity involved, the enunciation of statistical biological evolution laws might very well turn out to be an impossible task to achieve in practice.

Quantum Biology

Independently of the above difficulties, what are the reasons one can advocate in favor of a quantum view of biological evolution? Do we have signs of the presence of quantumness in the macro-word? We believe that the answer to this question should be affirmative.

We already described the possibility of the quantum machines, which exhibit remarkable quantum behaviors without suffering the typical environmental decoherence-like processes. These quantum entities are human artifacts, i.e., entities created by humans that certainly provide information about the particular culture of their creators. The culture in question is the *quantum culture* that originated from that epistemological evolution (not to say revolution) that has produced the discovery of quantum mechanics and, consequently, as one of us did in the eighties of last century, the possibility of replicating part of the behavior of non-spatial quantum entities by designing specific spatial entities and procedures to operate on them.

We want to stress that when considering evolution as a general process leading to *homo sapiens* as a species, one should not refrain considering these artifacts (which we have called quantum machines) as the result of this evolution. Indeed, human culture in a broad sense, and the knowledge we have gathered about the behavior of microscopic quantum entities, was necessary and has lead us first to the discovery of the nature of quantum entities in terms of ideas, hence at the level of the human minds, and then to the proposal of possible macroscopic physical realizations, such as the cracking walnut idealized one that we have described in some detail in the first part of this article.

We also mentioned that quantum machines are idealized versions of more general quantum-like entities, formed by systems that in given contexts can undergo processes of change of the CAP kind. Every time a system, be it biological or not, undergoes a symmetry breaking having some unpredictable (probabilistic) outcomes, we are in the situation where a selection is operated at the potentiality level. To give an example taken from human physiology, think of the spermatozoa reaching the far end of the uterus, which must face the dilemma of either turning left or turning right. It is a dramatic choice, as only rarely both ovaries of a woman release eggs simultaneously, so that one of the directions will end up being a dead-end. Different spermatozoa will select the left and right oviducts with different probabilities, which will depend on their state before being confronted with the left or right choice. The state of a sperm cell, relative to its possibility to turn left or right, can be assumed to depend on its previous short term evolution.[11] Thus, spermatozoa typically undergo quantum-like collapses, i.e., weighted symmetry breaking processes that, in their essence, are like the previously described cracking walnuts measurements.

Again, from a broad evolutionary perspective, this indicates that biological entities, like the cracking walnuts, can access part, but not all, of the quantum structure, and more precisely the part related to the CAP processes. This is the reason why only if we idealize the cracking walnuts' measurements, hence we also allow cultural evolution to work on top of biological evolution, we can reach the level of the pure quantum entities, with the quantum machines describing the idealized walnuts situations, which can also entangle by means of equally idealized rod mechanisms, which by the way can be further idealized and generalized by considering multidimensional versions of them, then able to replicate all possible quantum behaviors (Aerts and Sassoli de Bianchi 2014, 2015, 2016a).

The possibility of experimenting with physical entities under extremely well controlled conditions was also an epistemological evolutionary process that has made possible for us humans to meet the otherwise hidden quantum reality. An example is the discovery of *superfluidity* and *superconductivity*, thanks to the possibility of cooling down substances to temperatures close to the absolute zero, protecting them from the external (but also internal) thermal disturbances. Another significant example is the one of the *laser*, one of the rare quantum phenomena produced in a physics laboratory that cannot be decohered by the random thermal bombardment, as photons only interact extremely weakly with one another.

Other examples can be given of laboratory situations dealing with physical entities of a considerable size, where quantumness could be revealed when putting them under protected conditions, like *quantum Hall effects*, *Bose-Einstein condensates*, or even the observation of superposition states in small *mechanical resonators* (O'Connell et al. 2010). But all quantum measurements, also those dealing with elementary (micro) entities, are in fact experiments that can bring quantumness into contact with our human senses. Indeed, the very idea behind a quantum measurement with a micro-entity is to find a way not only of protecting it, to allow it to reveal its full quantum nature, but also to amplify such behavior, to make it consciously identifiable by a human observer using ordinary senses.

In measurement theory, this is the so-called *calibration condition*, which requires that, whenever the microscopic entity is in a given eigenstate, the macroscopic apparatus should indicate a corresponding pointer value in a non-ambiguous way. This is only possible if the macroscopic measuring apparatus and the microscopic

[11] It has for example been observed that after a forced right turn spermatozoa have an increased probability to turn left, when confronted with a left or right choice (Brugger et al. 2002).

measured entity maintain a connection, i.e., remain entangled, until the completion of the measurement. Thus, all quantum experiments, be they made on single elementary entities, or directly on large physical systems, they all end up displaying the investigated quantum effects at a scale that allow them to be recorded and observed by us humans. To put it in a catchy phrase: *quantum laboratories are quantum machines*. And from the evolutionary perspective we put forward in this article, they are all evolutions towards macroscopic quantum entities existing in the cultural realm of homo sapiens.

But human cultural evolution is not the only evolutionary process that has been able to harness quantumness to its advantage. Remarkably, it seems that living organisms are also able of harnessing some of the most unique quantum features to their biological advantage, and this on physiologically important timescales, managing to become less disturbed by the random bombardment of heat packets of energy. We can mention the example of *photosynthesis* (the process used by plants and other organisms to convert light energy into chemical energy), where evidence for the presence of quantum coherent energy transport over appreciable length scales was observed (Engel et al. 2007). Another interesting example is *avian magnetoreception*, the ability of some migrating species to navigate using Earth's magnetic field, where an explanation based on molecules that are created by a photochemical process, in spin-correlated states, has been proposed (Schulten et al. 1978; Gauger et al. 2011). Other quantum effects are also considered to possibly play important roles in the functioning of biological organisms, like long-range electron tunneling, and we refer to (Lambert 2013) for a review of current research on so-called *quantum biology*.

Most probably, in the years to come it will become more and more clear that our bodies, and life forms in general, are not just classical machines, but also quantum machines, that is, systems working with more internal coherence than initially expected. In fact, it is only very recently that biologists are starting to look to biology without the blinders of Newtonian physics, also considering all we have learned from quantum mechanics. Surprisingly enough, one of the research ambits where the quantum revolution starts to make its effects more felt is *cognitive sciences*. Indeed, in recent years an important discovery has progressively imposed itself among the scientists of the mind and human behavior: that we humans think, behave and take decisions decidedly in a quantum-like way. However, by this we are not affirming that our brains would necessarily be *quantum computers*, i.e., physical systems exploiting the existence of quantum effects (like interference and entanglement) at the micro-level, although this is certainly also a possibility, which has recently acquired more credibility in view of the growing evidence that life might indeed also exploit quantumness at the micro-level.

We can mention the well-known model by Hameroff and Penrose, where quantumness in the brain is hypothesized to be due to biomolecular processes taking place in *microtubular structures* (Hameroff and Penrose 1996). A more recent hypothesis is that the nuclear spins of phosphorus atoms could serve as rudimentary *qubits* (quantum bits) in the brain (Fisher 2015). These quantum models of the brain, as we mentioned already, rely on the existence of mechanisms that can protect the fragile quantum brain system from decohering too rapidly, considering our warm and wet brain environment. But independently of the role that might be played by quantum mechanics in describing the brain functioning at the micro-level, we can also reverse the logic of the investigation and simply look for quantumness at the level of the more abstract (but not for this less real) mind.

This of course will not necessarily provide a better understanding of the functioning of the brain, as is clear that the brain can be seen as a necessary condition for the spatiotemporal manifestation of the mind, but is not automatically a sufficient condition, particularly when considering the subtler aspects of our mental activity, like for instance our ability to have conscious experiences. Nonetheless, if we assume for a moment that most of our cognitive ability would emerge from the activity of the macroscopic brain machine, and that to model our cognitive/mental processes we need all the power of the quantum formalism (and possibly beyond), then we are somehow forced to conclude that such brain entity needs to work as a complex quantum machine. Not necessarily a quantum machine of the quantum computer kind, relying on the persistence of quantum effects at the micro-level, but certainly a quantum machine in the sense of a system exhibiting a *quantum organization*.

Quantum Cognition

Minds and brains need not to be viewed as identical entities, in the same way as they need not to be considered radically distinct aspects of our reality. We will come back shortly on this, but before that, let us briefly recall why cognitive processes require a quantum modeling. There are different ways to tell the story of the success of the quantum formalism in the modeling of cognitive and decision phenomena. To make a long story short, we can observe that human minds deal essentially with *concepts*, which are highly contextual entities, and that quantum mechanics is a theory that has precisely been designed to deal with *contextuality*. Also, in the same way a quantum entity can be in different states, a concept can also be understood as a *meaning entity* whose states depend on the (semantic) context in which it is immersed. Sometimes contexts will influence concepts (i.e., change their meanings, and therefore their states) in a deterministic way, other times in a perfectly indeterministic way, for instance when a mind is put in the situation of answering a question without having already a preprogrammed response, so that the latter must be created at the moment, similarly to how a potential outcome is actualized during a quantum measurement.

Furthermore, concepts can form *connections through meaning*, which in turn can produce significant correlations when these meaning-connections are tested/actualized in specific experimental situations. Thus, in the conceptual (human) realm, similarly to physics, one can design experimental situation where Bell's inequalities can be violated, showing that concepts can entangle in similar ways as quantum entities can do. Then, there is also the fact that when concepts are combined, new meanings can easily emerge in ways that cannot be described by considering the classical (Aristotelian) view that concepts would be just like containers of exemplars. These non-compositional emergence effects produced by conceptual combinations, when analyzed in statistical terms by performing experiments with large groups of subjects, can again be shown to be like the quantum mechanical interference effects, resulting from the superposition principle.

The above does not exhaust the list of quantum features that can be jointly observed at the micro-physical level and human conceptual level, and we refer the interested reader to the already vast literature about this new scientific field of investigation called quantum cognition; see (Gabora and Aerts 2002; Aerts and Gabora 2005a, b; Aerts 2009a; Khrennikov 2010; Busemeyer and Bruza 2012; Haven and Khrennikov 2013; Aerts and Sassoli de Bianchi 2015; Wendt 2015; Aerts et al. 2016) and the references cited therein. This remarkable correspondence has led one of us, in recent years, to ask a thought-provoking question: If the full quantum formalism has been applied to the modeling of human concepts with such an unexpected effectiveness, could this just indicate that the micro-physical entities would be themselves conceptual entities?

Conceptuality Interpretation

This kind of question is a typical "de Broglie move." Some readers might be aware that the French physicist *Luis de Broglie*, in the twenties of last century, following Einstein's successful introduction of photons in light waves to explain their interaction with matter, also asked a similar question, hypothesizing that if wave phenomena are to be associated with dual corpuscular properties, then also corpuscular phenomena are to be associated with dual undulatory properties. Likewise, if human conceptual entities are to be associated with a quantum behavior, could it be that micro-physical (quantum) entities are also to be associated with a conceptual behavior similar (although not identical) to that of human concepts? In other words, could it be that *quantumness* and *conceptuality* would be two terms referring to a same reality, or nature, which manifests at different organizational levels within our complex reality?

It is not the purpose of the present essay to make the case of what has been called the *conceptuality interpretation of quantum mechanics*, founded on the hypothesis that (Aerts 2010a): "the nature of a quantum entity is conceptual, i.e., it interacts with a measuring apparatus (or with an entity made of ordinary matter) in an analogous way as a concept interacts with a human mind (or with an arbitrary memory structure sensitive to concepts)." It is enough for us to observe that when this remarkable hypothesis is adopted, and its consequences explored (Aerts 2009b, 2010a, b, 2013, 2014; Sassoli de Bianchi 2015; Aerts and Sassoli de Bianchi 2017b), most of the quantum conundrums, such as entanglement and non-locality (usually considered to be not understood or even not understandable) suddenly become rather easy to explain.

Our interest here is to explore what are the consequences of the conceptuality interpretation concerning biological evolution, and evolution in general. If the above hypothesis is correct, then our understanding of inert matter must change. Indeed, though deprived of specific sensory organs, or of an apparent memory structure, inert matter would nevertheless be able to perceive (and create with) the surrounding material world in a way that is similar to how a human mind can perceive (and create with) its surrounding conceptual reality. The crucial point here is the observation of the similarity in behaviors: since physical entities behave analogously to cognitive entities, like human minds and concepts, by an argument along the lines of *Turing test* (here applied to cognition more than to intelligence, which is not necessarily implied by the former) the idea of a conceptual nature of the physical entities becomes likely, if not necessary, at least until proven to the contrary, i.e., until matter-energy, under closer inspection, would be seen to fail such Turing-like test.

Now, if it is true that a certain behavior presupposes a certain organization, this doesnt mean that a same organization would be needed to obtain a same behavior. Our example of the walnut-like quantum machine, behaving exactly as a spin-$\frac{1}{2}$ entity, is a perfect example of this. A spin-$\frac{1}{2}$ entity is certainly structured in a very different way than this quantum artifact, and it would be wrong to think that there is a sort of miniature of such object "within" a spin-$\frac{1}{2}$ entity, to explain its behavior during a measurement (like in the idea of the *homunculus*, in sixteenth-century alchemy). As an example, think of the different structures that living creatures can exhibit to digest nutrients. A creature with a stomach can certainly digest nutrients, but to conclude that to digest nutrients one needs a stomach would be a false syllogism. Similarly, it would be erroneous to conclude that cognitive activity needs brains and sensory organs to be carried out.

Pancognitivism

So, taking seriously the hypothesis that quantum entities are conceptual entities exchanging meaning (also called *coherence* in the quantum jargon) between entities made of ordinary matter (like the measuring apparatuses), the worldview that emerges is one that might be called *pancognitivism*, where everything within reality would be assumed to participate in cognition, with human cognition being just an example of it, expressed at a very specific organizational level. Thus, the pancognitivist view we put forward here is not a naïf one, where one would just assert that all sciences are mere theories of human mental content. Instead, our assertion/assumption is that the whole of our physical reality would be fundamentally conceptual, so that the way we scientists conceive the things out there would just be an aspect of how things in general exchange and internalize meaning about each others.

In other words, a mistake not to commit would be to misinterpret the assumption at the basis of the conceptuality interpretation as a tentative to promote a sort of radical anthropomorphic view of reality. It is just the notion that gives rise to the beingness (the *way of being*) of a quantum entity and of a human concept that are assumed to be the same, similarly to how, say, the notion of wave describes both the beingness of an electromagnetic wave and of a sound wave, which apart from that remain very different entities in their manifestation. Also, one could very well replace the notion of conceptual entity with that of *sign*, as introduced in *semiotics*, to equivalently formulate the conceptuality interpretation in a less human-centered way. Then, entities made of ordinary matter (like measuring apparatuses) would be interpreted as interfaces for these signs, instead of memories for the conceptual entities. But whatever interpretation one wants to adopt, the cognitive one or the semiotic one, the fact remains that a communication of some kind needs to be assumed to take place between the different physical entities, and, more importantly, such communication must have evolved symbiotically with the memories/interfaces processing the associated language/signs.

Quoting from (Aerts 2009b): "This introduces [⋯] a radically new way to look upon the evolution of the part of the universe we live in, namely the part of the universe consisting of entities of ordinary matter and quantum fields. Any mechanistic view, whether the mechanistic entities are conceived of as particles, as waves or as both, cannot work out well if the reality is one of co-evolving concepts and memories or signs and interfaces."

If the above is correct, then something similar to what happened in our human macro-world, with individuals using concepts and their combinations to communicate, may have already occurred, and continue to occur, *mutatis mutandis*, in the micro-realm, with the entities made of ordinary matter communicating and co-evolving thanks to a communication that uses a language made of concepts and combinations of concepts that are precisely the quantum entities and their combinations. No need to say, this remains for the time being a speculative view that needs to be further critically explored. However, it is also a fascinating view, having far-reaching consequences for our understanding of evolution in general.

If the right way to think of the evolution of matter-energy is as a change resulting from the interaction of conceptual entities with memory structures sensitive to their meaning, then the picture one needs to adopt for the overall description of our evolving physical reality would be that of *cultural evolution*. Thus, what is usually considered, on our planet, to be the secondary evolutionary process, which appeared following the evolution of the biological species, would in fact be a much more ancient process of change, and in a sense, the only fundamental process of change in force since the beginning of our universe.

Now, although there are no doubts that in some respects the evolution of cultures can be described similarly to Darwinian evolution, that is, in terms of random variations, competition on actual properties, and inheritance, it has also been recognized that, because of the genuine unpredictability of the human minds, which are entities that can invent strategies and conceptualize new situations by creating new meanings, it is also reasonable to assume that Darwinian mechanisms only play a minor role in cultural evolution (Gabora and Aerts 2005b). Indeed, the core issue in cultural evolution is to understand the *contextuality* and *compositionality* of conceptual entities (the human ones here), and the processes they subtend. And no surprise, the general processes that underlay conceptual evolution, and therefore cultural evolution, i.e., their basic modes of change, are those of CAP, i.e., of context-driven actualization of potential.

So, on one hand we have the evolution of our material universe, which according to the conceptuality interpretation should be viewed more as the evolution of a *cosmic culture*, possibly formed by multiple subcultures, and on the other hand we have the more recent episode, on our planet, of the emergence of the specific *human culture*, with its ability to investigate the nature of the material universe from which it emerged (which however should not be reduced to what can be represented in our limited spatiotemporal theater). And, somehow in between these two evolutionary processes, we must place biological evolution, also likely to be in part produced by CAP quantum-like processes.

It is important to stress once more, to avoid misunderstandings, that we are not negating the validity of the Darwinian evolutionary mechanisms. We are just emphasizing that they need to be reframed within a larger culture-like and CAP-like evolutionary picture, in the same way as, for example, classical mechanics needs to be reframed in the larger conceptuality-like picture of quantum mechanics, or even in larger pictures that encompass both classical and quantum theories (Aerts and Sassoli de Bianchi 2016b). Hence, Darwinian evolution should not be taken as the model for cultural evolution (i.e., epistemological and conceptual change). Instead, it is cultural evolution that should be considered the right model (the right metaphor) for biological evolution. In other words, the processes of change happening at the conceptual, psychological and social levels around us, are the more general ones, with the Darwinian-like evolution only constituting a very special case. And, as we suggested already, quantum laws would also contribute to biological evolution, as we have recently begun to discover in the emerging field of quantum biology. But our understanding of these quantum laws, as we also explained, should not be limited to their appearance at the micro-level, and should also include their appearance at a more general *structural level*, i.e., as specific forms of organization.

The Reach of Evolution

According to the view we are here considering, evolution could have started long before the advent of the biological realm, possibly even at the primordial stages of formation of our universe. More specifically, we can ask: Why should we assume that the observed growth of complexity in biological entities only started from single-cell organisms, like bacteria, and not already from the basic elements of matter themselves, like quarks, electrons, neutrons, protons, atoms, molecules, etc.? And, more importantly: Is it plausible that the increase of complexity only goes bottom-up, from bacteria to multicellular organisms, from plants to animals, then the human beings and their cultures?

To estimate the plausibility of this dominant view amongst today evolutionists, one should take into consideration the actual complexity of the pre-biological entities, which was revealed to us by quantum theory and the CAP-driven processes of change it subtends. If the latter is taken seriously, in the sense that one really tries to understand and explain why and how quantum entities can behave the way they behave, then, in our view, a conceptuality interpretation becomes very plausible. Let us mention, however, that the difficulty one might experience in endorsing a non-naïf pancognitivist worldview like the one we are presenting here can be found in our necessarily parochial view about what a cognitive entity should look like and behave like. Quantum mechanics, and more recently quantum cognition, has brought us possibly closer to the mystery, allowing us to reach a more universal (less human-centered) understanding of cognitive processes, making more evident what are those elements of reality characterizing them, such as interference, emergence, entanglement and contextuality, which require the sophisticated mathematical language of quantum theory (and its possible generalizations) to be properly formalized.

This also means that when we deal with a piece of inert matter, one should view it from a double perspective. First, it plays a very active communicative role, as a type of proto-memory in the realm of the micro-world, with the quantum "particles" (including also the so-called *quasiparticles*, like phonons) playing the role of proto-concepts. Second, it plays a passive role (hence the designation of "inert") in the realm where human life exists, e.g., on the surface of our planet, and this is the reason why, in this realm, it can fairly well be modeled by classical physics and the typical interactions it describes. So, in the realm of our planet's surface, inert matter, contrary to living matter, would describe a sort of evolutionary cul-de-sac, in the sense that lifeless entities would be typically those that were unable (or haven't yet been able) to protect themselves against the random bombardment of heat photons, or find a way to transfer quantumness at a new organizational level, i.e., find a viable road towards *coherent macroscopicity*.

To quote from Aerts and Sozzo (2015): "One could state that a nervous system is an amplifier for quantum from the micro-level to the macro-level, because it allows the entity with the nervous system to develop complicated strategies of defense against random perturbations with changes that are destructive for the evolved organization. In the case of human beings, this capacity of defense has evolved to a very sophisticated level, fully exploring the amplifying effect of the nervous system, and giving rise to cultural cognition, with languages and other cultural items as manifestations of it".

Concerning more specifically human culture, it can be understood as a further way to make quantumness manifest at the macro-level, as we explained already. Quoting again from Aerts and Sozzo (2015): "Human culture is also an evolutionary process, albeit not Darwinian. It has not only managed resistance against the random bombardment of heat energy packets, but also evolved to use this heat energy and make it into non-random energy. Humans energy-harvesting from heat started with the first steam engine, which literally is the transformation of random energy into structured energy. Does this give rise to quantum structure? Not always, and not automatically, but this is certainly the case for the energy used in those laboratories that have produced quantum effect at room temperature".

It should also be emphasized that biological (pre-cultural) life, an unlikely episode that happened at the local level of Earth's crust, is a hinge between two global evolutionary levels: the quantum micro-level, from which it emerged, finding a way to protect itself and construct nervous systems, and the equally quantum cultural level, which emerged from the latter through the creation of an abstract form of communication. In other words, evolution, seen from our human perspective, is a progression from the global conceptual-like micro-level to the equally global human cultural level, passing through the "needle eye" of biological evolution, which is probably also the process that is better approximated by classical Darwinism. Note that human culture is global because it is potentially boundless. Indeed, the ability to create knowledge and using it to provide support, independently of the environmental conditions, allows for an unlimited reach, and the power to alter (with time) the entire universe.

Closing Thoughts

After our excursus across different sciences, to draw a bigger (and possibly also more truthful) picture of the nature of our complex world, and of the mechanisms responsible for its evolution, it is time to conclude with some final thoughts.

We have argued that quantum effects are more ubiquitous on the surface of our planet than what is usually considered, and that they manifest at different degrees in the different organizational levels, like the micro and macro, the inert and living, the biological and cultural, despite the chaotic heath bath in which the planet is immersed. What we have presented also touches one of the crucial debates in evolution: the distinction between so-called *bottom-up* and *top-down* processes. The former are those processes where complexity is assumed to progress from simplicity, typically by Darwinian natural selection (plus some other favorable circumstances), whereas the latter are the processes where an already formed complexity, like the one expressed by us homo sapiens, design additional complex entities, like artifacts, by combining simpler elements in interesting and useful ways. This distinction is of course important to make, but one should not conclude from it that human design would be a unique and once in a time result of biological and cultural evolution. Indeed, if it is true that our best understanding of the interaction of quantum entities with matter is to describe them as meaning entities evolving in meaning-sensitive environments, it then follows that top-down design-like processes might have happened multiple times, in parallel or alternation with the bottom-up ones, during the entire evolution from primordial matter to biological and cultural life.

In that respect, we observe that the existence of sophisticated languages, as communication tools, is an uncontroversial sign of the presence of top-down-design entities. Indeed, this is also one of the criteria usually adopted to identify the presence of extra-terrestrial intelligent life. Hence, the question of whether there would be some plausibility in considering the existence of multiple episodes of top-down-design in our evolutionary history, from matter to human culture, could be answered by precisely investigating the research field of quantum cognition. The key question we should ask is the following: Is the observed unreasonable success of quantum theory, in the modeling of human cognition, sufficient to validate the conceptuality interpretation of quantum mechanics, where quantum entities are described as interacting by means of a sophisticated (proto) language? The present authors, based on their understanding derived from previous studies on operational-realistic approaches to quantum physics and quantum modeling of human cognition, are likely to answer in the affirmative to the above question, but of course can only leave it to the readers to form their opinion and provide their own answers, after having deepened their understanding of these fascinating and truly interdisciplinary approaches to reality.

As a last thought, we would like to mention the so-called *Fermi paradox*, namely the apparent contradiction between the lack of evidence and, at the same time, the rather high probability estimates (e.g., those given by *Drake's equation*) for the existence of extraterrestrial civilizations. If the collection of macroscopic objects populating our material universe are to be classified amongst the dead end roads of macroscopicity, then a good earthly analogy for them would be the huge landfills to be found close to the megacities. The cultural artifacts found on such places are entities that have lost their cultural coherence and meaning-connections with those ambits in which they prospered before landing in the trash.

To give an example, if pieces of paper with traces of printed words can still be found in such landfills, they can no longer function as *carriers of meaning*, as their initial states (for instance the states associated with the whole books containing these words, telling meaningful stories) would have decohered, and the same of course happens with other kinds of cultural artifacts. It is of course not on such landfills that one should look for culture. In a similar way, the spatiotemporal universe is perhaps not the best place to look for finding life and culture within our reality. Life and culture might indeed more abundantly be found not so much by exploring our universe *in width*, i.e., its spatial vastness, but *in depth*, i.e., exploring those regions that, from our spatiotemporal perspective, appear to be non-spatial and non-temporal, and in that sense more conceptual than objectual.

It is perhaps this *in-depth-direction* that has been traveled (at least in part) by those individuals that have tried, in the ambit of so-called *inner (re)search*, to access more universal forms of quantumness, at the price of learning how to silence all possible forms of decohering disturbances. Think for instance of the practice of sensorial isolation known as *pratyahara*, the fifth element among the eight stages of *Patanjalis Yoga* (Ravindra 2009). We can think of it as a gateway, created by a specific inner technology, to pass from the experience of (spatial) "external" states to (non-spatial) "internal" states, with the latter to be further stabilized and deepened by the successive practice of concentration (*dharana*) and abstract meditation (*dhyana*). It is maybe no surprise then to observe that inner researchers frequently report of the encounter with rich and abundant life forms and cultures, in the course of their inner (in-depth-direction) journeys, some of which are also described as being more advanced than ours.

Acknowledgements We are grateful to Liane Gabora for her constructive comments that helped improving the manuscript. We are also thankful to Nelson Abreu, for providing the stimulus and motivation to write this article.

References

D. Aerts, A possible explanation for the probabilities of quantum mechanics. J. Math. Phys. **27**, 202–210 (1986)

D. Aerts, The stuff the world is made of: physics and reality, in *The White Book of 'Einstein Meets Magritte'*, ed. by D. Aerts, et al. (Kluwer Academic Publishers, Dordrecht, 1999), pp. 129–183

D. Aerts, Being and change: foundations of a realistic operational formalism, in *Probing the Structure of Quantum Mechanics: Nonlinearity, Nonlocality, Probability and Axiomatics*, ed. by D. Aerts, M. Czachor, T. Durt (World Scientific, Singapore, 2002), pp. 71–110

D. Aerts, Quantum structure in cognition. J. Math. Psychol. **53**, 314–348 (2009a)

D. Aerts, Quantum particles as conceptual entities: a possible explanatory framework for quantum theory. Found. Sci. **14**, 361–411 (2009b)

D. Aerts, Interpreting quantum particles as conceptual entities. Int. J. Theor. Phys **49**, 2950–2970 (2010a)

D. Aerts, A potentiality and conceptuality interpretation of quantum physics. Philosophica **83**, 15–52 (2010b)

D. Aerts, La mecánica cuántica y la conceptualidad: Sobre materia, historias, semántica y espacio-tiempo. Scientiae Studia 11, pp. 75–100 (2013). Translated from: D. Aerts, Quantum theory and conceptuality: matter, stories, semantics and space-time, arXiv:1110.4766 [quant-ph], Oct 2011

D. Aerts, Quantum theory and human perception of the macro-world. Front. Psychol. **5**(554) (2014). https://doi.org/10.3389/fpsyg.2014.00554

D. Aerts, S. Aerts, J. Broekaert, L. Gabora, The violation of Bell inequalities in the macroworld. Found. Phys. **30**, 138–1414 (2000)

D. Aerts, S. Bundervoet, M. Czachor, B. DHooghe, L. Gabora, P. Polk, S. Sozzo, On the foundations of the theory of evolution, in *Worldviews, Science and Us: Bridging Knowledge and Its Implications for our Perspectives of the World*, ed. by D. Aerts, et al. (World Scientific, Singapore, 2011)

D. Aerts, M. Czachor, B. DHooghe, Towards a quantum evolutionary scheme: Violating Bell's inequalities in language, in *Evolutionary Epistemology, Language and Culture. A Non-Adaptationist, Systems Theoretical Approach*, eds. by N. Gontier, et al. Theory and Decision Library. Series A: Philosophy and Methodology of the Social Sciences, (Springer, 2006), pp. 453–478

D. Aerts, L. Gabora, A theory of concepts and their combinations I: the structure of the sets of contexts and properties. Kybernetes **34**, 167–191 (2005a)

D. Aerts, L. Gabora, A theory of concepts and their combinations II: a hilbert space representation. Kybernetes **34**, 192–221 (2005b)

D. Aerts, M. Sassoli de Bianchi, The extended bloch representation of quantum mechanics and the hidden-measurement solution to the measurement problem. Ann. Phys. **351**, 975–1025 (2014). See also: Ann. Phys. **366**, 197–198 (2016)

D. Aerts, M. Sassoli de Bianchi, The unreasonable success of quantum probability I: quantum measurements as uniform measurements. Journal Mathematical Psychology **67**, 51–75 (2015)

D. Aerts, M. Sassoli de Bianchi, The extended bloch representation of quantum mechanics. explaining superposition, interference and entanglement. J. Math. Phys. **57**, 122110 (2016a)

D. Aerts, M. Sassoli de Bianchi, The GTR-model: a universal framework for quantum-like measurements, in *Probing the Meaning of Quantum Mechanics. Superpositions, Dynamics, Semantics* *and Identity*, eds. by D. Aerts et al. (World Scientific Publishing Company, Singapore, 2016b), pp. 91–140

D. Aerts, M. Sassoli de Bianchi, *Universal Measurements* (World Scientific, Singapore, 2017a)

D. Aerts, M. Sassoli de Bianchi, S. Sozzo, T. Veloz, On the conceptuality interpretation of quantum and relativity theories. To be published in Foundations of Science, in *Proceedings of the International Symposium Worlds of Entanglement*, held at the Free University of Brussels (VUB) on 29–30 September 2017b. arXiv:1711.09668 [physics.hist-ph], Nov 2017

D. Aerts, S. Sozzo, What is Quantum? unifying its micro-physical and structural appearance, in *Quantum Interaction. QI 2014*, eds. by H. Atmanspacher, et al. Lecture Notes in Computer Science, vol. 8951 (Springer, Cham, 2015), pp. 12–23

D. Aerts, M. Sassoli de Bianchi, S. Sozzo, On the foundations of the brussels operational-realistic approach to cognition. Front. Phys. **4**, 17 (2016). https://doi.org/10.3389/fphy.2016.00017

P. Brugger, E. Macas, J. Ihlemann, Do sperm cells remember? Behav. Brain Res. **136**, 325–328 (2002)

J.R. Busemeyer, P.D. Bruza, *Quantum Models of Cognition and Decision* (Cambridge University Press, Cambridge, 2012)

G.S. Engel et al., Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. Nature **446**, 782–786 (2007)

M.P.A. Fisher, Quantum cognition: the possibility of processing with nuclear spins in the brain. Ann. Phys. **362**, 593–602 (2015)

L. Gabora, D. Aerts, Contextualizing concepts. in *Proceedings of the 15th International FLAIRS Conference (Special Track 'Categorization and Concept Representation: Models and Implications')*, American Association for Artificial Intelligence, Pensacola, Florida, 14–17 May 2002

L. Gabora, D. Aerts, Evolution as context-driven actualisation of potential: toward an interdisciplinary theory of change of state. Interdis. Sci. Rev. **30**, 69–88 (2005a)

L. Gabora, D. Aerts, Distilling the essence of an evolutionary process and implications for a formal description of culture, in *Proceedings of Center for Human Evolution Workshop 4: Cultural Evolution, 18–19 May 2000*, ed. by W. Kistler, Foundation for the Future, (Bellevue, WA, 2005b)

L. Gabora, E.O. Scott, S. Kauffman, A quantum model of exaptation: incorporating potentiality into evolutionary theory. Prog. Biophys. Mol. Biology **113**, 108–116 (2013)

E.M. Gauger, E. Rieper, J.J.L. Morton, S.C. Benjamin, V. Vedral, Sustained quantum coherence and entanglement in the avian compass. Phys. Rev. Lett. **106**, 040503 (2011)

E. Haven, A.Y. Khrennikov, *Quantum Social Science* (Cambridge University Press, Cambridge, 2013)

S.R. Hameroff, R. Penrose, Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness, in *Toward a science of consciousness; the first Tucson discussions and debates* eds. by S.R. Hameroff, A.W. Kaszniak, A.C. Scott, Also published in Math. Comput. Simul. **40**, 453–480 (1996) (MIT Press, Cambridge, MA, 1996), pp. 507–540

A.Y. Khrennikov, *Ubiquitous Quantum Structure* (Springer, Berlin, 2010)

N. Lambert et al., Quantum biology. Nat. Phys. **9**, 10–18 (2013)

A.D. O'Connell et al., Quantum ground state and single-phonon control of a mechanical resonator. Nature **464**, 697–703 (2010)

V.V. Ogryzko, A quantum-theoretical approach to the phenomenon of directed mutations in bacteria (hypothesis). Biosystems **43**, 83–95 (1997)

R. Ravindra, *The Wisdom of Patanjali's Yoga Sutras: A New Translation and Guide by Ravi Ravindra*, (Morning Light Press, 2009)

M. Sassoli de Bianchi, Taking quantum physics and consciousness seriously: What does it mean and what are the consequences? To appear, in *the proceedings of the 1st ICC, held at the Research Campus of the IAC*, in Portugal, from 22–24 May 2015

K. Schulten, C.E. Swenberg, A. Weller, A biomagnetic sensory mechanism based on magnetic field modulated coherent electron spin motion. Z. Phys. Chem. **111**, 1–5 (1978)

A. Wendt, *Quantum Mind And Social Science* (Cambridge University Press, Cambridge, 2015)

Part V
MISC

Chapter 32
In the Deserts of Cartography: Building, Dwelling, Mapping

Robert T. Tally Jr.

Any discussion of the map and the territory, at least insofar as it touches on literary or cultural studies, will almost inevitably turn to the evocative little story by Jorge Luis Borges, tantalizing titled "On Exactitude in Science." It is certainly one of the most recognizable, even most canonical, texts in spatiality studies, broadly conceived, and it always helps to set a properly philosophical tone when thinking about the problem of representation.

At once elegiac and absurd, the fragment—that is, a text presented as if it were a fragment from a larger narrative, but it fact complete unto itself—tells of an imaginary empire in which the passion for mimetic accuracy in mapmaking had reached its zenith with the creation of the ultimate chart, drawn up according to a one-to-one scale, such that the map was coextensive with the territory it was supposed to represent. Citing a fictional source (namely, Suárez Miranda, *Viajes de varones prudentes*, Libro IV, Cap. XLV, Lérida, 1658), which already serves to distance the narrative from the presentation of it and add an element of archival authority to the history, Borges writes:

> In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.[1]

[1] Jorge Luis Borges, "On Exactitude in Science," *Collected Fictions*, trans. Andrew Hurley (New York: Penguin, 1999), p. 325. Borges cites a fictional source: Suárez Miranda, *Viajes de varones prudentes*, Libro IV, Cap. XLV, Lérida, 1658.

R. T. Tally Jr. (✉)
Department of English, Texas State University, 601 University Drive, San Marcos, TX 78666, USA
e-mail: robert.tally@txstate.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_32

In Borges's vision, a narrative of the absurd "exactitude" in the geographic science of the earlier cartographers concludes with a bleak scene of a desert wasteland, a veritable non-place occupied by animals, beggars, and the odd scraps of the imperial map.

Borges's story of a map coextensive with its territory has become a haunting reminder of the absurdity of the quest for perfectly mimetic representations in cartography and, by extension, in other arts and sciences. An earlier dramatization of this idea, from Lewis Carroll's *Sylvie and Bruno Concluded*, is much more humorous in making a similar point, as I will discuss below, but here the air of melancholy or the sense of loss pervades Borges's brief narrative in such a way to preclude its being seen as a joke (or, at least, not merely as a joke).[2] Famously, Jean Baudrillard used the Borges fable to illustrate his conception of late-twentieth-century hyperreality, in which the simulacrum precedes the genuine article it was supposed to mimic. For Baudrillard, the map precedes and, in a way, produces the territory. Baudrillard actually inverts the order depicted in the fable. Whereas Borges wished to highlight the surreal vision of a representation that attempted, as it were, not only to replicate but to replace the original, Baudrillard suggests that, in our time, the simulacrum precedes the referent entirely. There is no original to be copied. For Baudrillard, the tattered remains of the territory might be found in the margins of the map, not vice versa, and thus the deserts are not those of the old Empire, but of our own "real" world. As he notoriously puts it, in a manner that found favor with the producers of *The Matrix* films and other science fiction enthusiasts (Slavoj Žižek among them), we occupy "*the desert of the real itself.*"[3]

From the perspective of the geographical sciences, these speculations over the perfect, one-to-one scalar depiction of territorial space in a map are, quite rightly, amusing absurdities, thought experiments that remind us that all representation is figurative, metaphorical, or allegorical.[4] The conceptual dilemma posed by a consideration of the relationship between the map and the territory is rather simpler than the hyperreality thesis of Baudrillard, who finds that there are no originals to be copied, no referent to which the sign refers, and not territories to be mapped; there are only copies, signs, and maps. However, most critics are as yet unwilling to

[2] Lewis Carroll, *Sylvie and Bruno Concluded* (London: Macmillan, 1893), 169.

[3] Jean Baudrillard, *Simulacra and Simulation*, trans. Sheila Faria Glaser (Ann Arbor: University of Michigan Press, 1994), 1–2, italics in original. In *The Matrix* (1999), directed by the Wachowskis, a character introduces another to the fact that what is taken for human reality and lived experience is in fact only a great computer simulation, punctuating this surprising news with the line, "Welcome to the desert of the real." This phrase was used as the title of 2002 book by Slavoj Žižek, in which the author employed a Lacanian and Marxist analysis of the terrorist attacks of September 11, 2001, and the media responses to them. See Žižek, *Welcome to the Desert of the Real: Five Essays on September 11* (London: Verso, 2002).

[4] See, e.g., J. B. Harley, *The New Nature of Maps: Essays in the History of Cartography*, ed. Paul Laxton (Baltimore: Johns Hopkins University Press, 2001).

give up on referentiality in toto, even if they are willing to question reality as it appears, perhaps by interrogating the conditions for the possibility of apprehending what we think of as reality as such. (This is a legacy of Kant, among others.) At a more practical level, any users of the map recognize the degree to which the map cannot be "true" to the territory it purports to represent. But one of the first consequences of the realization that a perfectly mimetic image of the respective space on a chart is impossible is that we come to realize that we can always imagine *better* —not necessarily more accurate, but more useful—maps. Or, as Fredric Jameson has put it in his well known "digression on cartography" in his *Postmodernism* book, when "it becomes clear that there can be no true maps," then, "at the same time it also becomes clear that there can be scientific progress, or better still, a dialectical advance, in the various historical moments of mapmaking."[5] Along these lines, we might say that the failure of the cartographers to create the ultimate, perfect map is actually a boon to map-users, which is to say, everyone. Without a perfect map, we are free to make maps that suit our needs and desires.

Returning to Borges's "On Exactitude in Science," then, we can focus our attention, not so much on the neat idea of a surreal map that is a point-for-point graphic replication of the territory, but on the aftermath of this would-be triumph of geography. In other words, leaving aside the mapmakers with their ambition, ingenuity, and ultimate failures, we can look at the post-geographic age in which the great map was deemed useless and pitilessly "delivered […] up to the Inclemencies of Sun and Winters." In this epoch, according to Borges's tale, the tattered remnants of the map that can be found here and there in "the Deserts of the West" are all that remains of the "Disciplines of Geography" in that land, which might be taken as a damning indictment of the era and of the people living in it. These are a people who have become uninterested in geographical science, who have lost respect for their ancestors' accomplishments, and abandoned the past treasures to the realm of wind and dust. Although Borges does not necessarily report it this way, this is our age and our land. We are living in the deserts of cartography.

The vision is elegiac, if not indeed apocalyptic. The deserts of cartography conjure up an image of undeniable loss, but it is also that sign of progress, as the epistemic triumphs of a great theory-oriented generation become impractical encumbrances to a later, more pragmatic generation. Proper mapmaking, at least as an adjunct to a formal disciplinary field of geography, ceases. The old Map deteriorates. This era is typified by the open spaces in which those remnants of the map, the scattered and tattered fragments of the great systematic representation of the world which now blow in the wind, forming temporary shelters to stray animals and vagabonds. Remnants, remains, residue … that which is left behind. Perhaps ours is

[5] Fredric Jameson, *Postmodernism, or, the Cultural Logic of Late Capitalism* (Durham: Duke University Press, 1991), 52.

the age of the remainder? An epoch of the residual, where the cultural dominants are intolerable and the emergent forms are almost too horrible to imagine. Now seems a perfect time to take note of the traces, those mementoes of former valiant efforts, as the present seems all too dystopian for so many, while the future cannot be imagined apart from a sort of end-of-the-world scenario, an apocalypse without recovery, Armageddon without hope. These fragments of the map, currently littering the deserts and offering the barest shelter to vagrants, might provide clues to an alternative cartography, vistas into another world.

The image of the desert, bestrewn with the ragged remnants of the grand imperial map, evokes bleak austerity. The desert is a kind of non-place, a space of homelessness or estrangement in which the individual or collective subject is forever displaced, without necessarily being able to become reoriented. For instance, the great cultural geographer Yi-Fu Tuan, in *Space and Place: The Perspective of Experience*, has defined *place* as a sort of pause, a resting of the eyes, or an instant of awareness when one isolates, if only momentarily, a portion of otherwise undifferentiated space, and in noticing it as such, imbues it with meaning.[6] At this point, it becomes familiar, like a home, again if only for a moment, whereas the still inchoate spaces surrounding it remains alien, uncanny, menacing, and dangerous. The desert, sometimes literally and often figuratively, conjures up an uncanny sense of a vast, uninhabitable, and unhomely space.

The desert is not a home, though it may be a space through which one must pass, a zone of transgression or of liminality. It might be likened to the "non-places" identified by Marc Augé in his influential study, *Non-Places: Introduction to the Anthropology of Supermodernity*. Augé examines transitory sites, such as airports, train stations, hotels, highways, and supermarkets, which in a sense are not so much *places*—that is, locations instilled with meaning, dense with historical and social reference, the result of creative human endeavor, and so forth—as *non-places*, uniform, homogeneous zones of transit in which modern humans increasingly spend their lives. Occupying these entirely, perhaps all-too-social spaces, we experience another sense of homelessness, a desert of another kind.[7] But more likely, the desert could be characterized as an *atopia*, which Siobhan Carroll has analyzed as spaces "antithetical to habitable place"; she adds to the list of manmade atopias such as those mentioned by Augé a number of "natural atopias," such as the North Pole, the middle of the ocean, the desert, or outer space, although she also notes how cyberspace is frequently imagined as a somewhat positive, manmade atopia. Carroll concludes that, whether these atopias are viewed as spaces that either liberate or threaten the individual subject, they have become increasingly useful in "orientating ourselves to the sublime space of the planet and the human networks

[6] Yi-Fu Tuan, *Space and Place: The Perspective of Experience* (Minneapolis: University of Minnesota Press, 1977), 161–162.

[7] See Marc Augé, *Non-Place: Introduction to an Anthropology of Supermodernity*, trans. John Howe (London: Verso, 1995).

that span its surface."[8] In the unhomeliness (or uncanniness) of such *atopian* sites we may also come to make sense of the places in which we might feel at home.

In *Being and Time*, Heidegger postulated that our experience of anxiety was intimately tied to the uncanny (*unheimlich*) and thus reflected a profound sense of being "not-at-home."[9] This unease or estrangement is in a way similar to that "homelessness" which Heidegger later identified as the "destiny of the world," a pervasive and troubling condition. In his "Letter on Humanism," Heidegger asserts that a certain homelessness is the condition of contemporary man. The "homeland" that is lacking is understood "in an essential sense, not patriotically or nationalistically but in terms of the history of Being." Ontologically speaking, human beings require a *heimlich* place. "Homelessness," he continues, "is the symptom of the oblivion of Being."[10]

There is a vaguely romantic appeal to this sense of homelessness. It carries something of the flavor of Georg Lukács's "transcendental homelessness" in *The Theory of the Novel*, in which it is used to characterize the condition of man in "a world abandoned by God."[11] In such a world, which lacks the sense of totality given in an earlier epoch (the age of the epic), the novel becomes the form-giving form by which humans can make sense of their world. In my reading of Lukács's work, I have suggested that this might also be imagined as a kind of cognitive mapping, to use Jameson's well known term.[12] That is, the novel is a form that can be used to give form to the world of limited human perspective and experience by coordinating that experience with a sense of the broader social totality. In this way, it might function in a manner similar to that of a map, which provides a figurative representation of space, often complete with a bird's-eye-view perspective, that can thus enable the individual subject to locate him- or herself in relations both to other places and to a projected, more global space. As Jameson had described a somewhat simplified version of cognitive mapping, drawing on Kevin Lynch's discussion of "wayfinding" and "imageability" in his *The Image of the City*, "Disalienation in the traditional city, then, involves the practical reconquest of a sense of place and the construction or reconstruction of an articulated ensemble which can be retained in memory and which the individual subject can map and remap along the moments of mobile, alternative trajectories."[13] And, as Miroslav Holeb has intimated in his

[8] Siobhan Carroll, "Atopia/ Non-Place," in *The Routledge Handbook of Literature and Space*, ed. Robert T. Tally Jr. (London: Routledge, 2017), 159, 164–165.

[9] See Martin Heidegger, *Being and Time*, trans. John Macquarrie and Edward Robinson (New York: Harper and Row, 1962), 233.

[10] Martin Heidegger, "Letter on Humanism," trans. Frank A. Capuzzi, in Heidegger, *Basic Writings*, ed. David Farrell Krell (New York: Harper and Row, 1977), 217–219.

[11] Georg Lukács, *The Theory of the Novel*, trans. Anna Bostock (Cambridge: MIT Press, 1971), 88.

[12] See my "Lukács's Literary Cartography: Spatiality, Cognitive Mapping, and *The Theory of the Novel*." *Mediations* 29.2 (Spring 2016): 113–124.

[13] Jameson, *Postmodernism*, 51.

poem, “Brief Thoughts on Maps,” a map, even the wrong map, may help one find one’s way home.[14]

Transferring this idea to the sense of homelessness referred to above, we might suggest that, for those occupying the alien space of the desert, there is an urgent need for a form of mapping that will make possible a sense of place or a “homeliness.” It may be ironic to think that, for the “Animals and Beggars” inhabiting them, the “Tattered Ruins of the Map,” in fact, are like a home. Can one make oneself “at home” in a map? Aside from the scant shelter from the “Inclemencies of Sun and Winters” that sheer parchment can provide, the fragments of a map may well offer solace, even comfort, to the errant wanderer and his shadow. Dwelling in the deserts of cartography, one necessarily discovers places and projects relations among them, constellating the assorted points into a meaningful ensemble, and thus, perhaps, if not making oneself at home exactly, then making sense of one’s own place in the world.

While the desert seems to be a particularly inhospitable place, that does not mean one cannot possibly feel at home there. Not only are there the cultures and populations that have managed to survive, even thrive, in the desert environment, but many have been immediately struck by the beauty of the desert or have developed an affinity for it over time, such that the desert landscape represents, for some, an altogether “homely” territory.

For example, Tuan, in his 1990 Preface to the Morningside Edition of *Topophilia* (which had originally been published in 1974), recounts the narrative of a camping trip he took with several of his fellow graduate students from Berkeley to Death Valley in the early 1950s. Awaking to a sunrise over a landscape utterly foreign to him in his previous personal experience, Tuan reports witnessing “a scene […] of such unearthly beauty that I felt transported to a supernal realm and yet, paradoxically, also at home, as though I had returned after a long absence” (xi). Tuan, who is interested in the phenomenological apprehension or experience of space and place, quite rightly observes that the favored environs for some people might be thoroughly uninhabitable or distasteful to others. The site of one person’s topophilia might well engender feelings of topophobia in another. As Tuan continues his meditation on his own affective geography with respect to the ostensibly

[14] See Miroslav Holub, “Brief Thoughts on Maps,” trans. Jarmila and Ian Milner, *Times Literary Supplement* (February 4, 1977): 118. This poem relates the story, itself a retelling of a tale formerly told by the Nobel Prize winner Albert Szent-Györgi, of a Hungarian reconnaissance unit, hopelessly lost in a snowstorm in the Alps during World War I. At the brink of despair and resigning themselves to death, they find a map that one soldier had kept in his pocket. Using it to locate their bearings, the soldiers manage to make it back safely to camp. There the commanding officer, who had been wracked with anguish and guilt over the loss of his troops, asked to see this miraculous map that had saved their lives. A soldier handed it over, and it was revealed to be a map, not of the Alps, but of the Pyrenees. The moral of the story appears to differ among its tellers. Szent-Györgi’s point in originally recounting the anecdote was to show that, in science, even errors or false starts can lead to success. Holub’s broader intention in retelling the tale, however, may have been to show how, in the words of his poem, “life is on its way somewhere or another,” regardless of one’s sense of orientation.

bleak terrain of a place like the Death Valley National Monument, "[t]he desert, including its barren parts and (I would even say) especially those, appeals to me. I see in it purity, timelessness, a generosity of mind and spirit" (xi). The geographer admits that his preference for the desert over, say, the rain forest is a prejudice, but such personal or cultural feelings about a space are entirely consistent with the human understanding of and engagement with the environment. Undoubtedly, Tuan says,

> peoples of the desert (nomads as well as sedentary farmers in oases) love their homeland: without exception humans grow attached to their native places, even if these should seem derelict of quality to outsiders. But the desert, despite its barrenness, has had its nonnative admirers. Englishmen, in particular, have loved the desert. In the eighteenth and nineteenth centuries, they roamed adventurously in North Africa and the Middle East, and wrote accounts with enthusiasm and literary flair which have given the desert a glamor that endures into our time […] Why this attraction for Englishmen? The answers are no doubt complex, by I wish to suggest a psychogeographical factor—the appeal of the opposite. The mist and overpowering greenness of England seems to have created a thirst in some individuals to seek their opposite in desert climate and landscape. (xii)

For Tuan, as for Heidegger, the love of place involves a sense of being "at home" there, but Tuan also insists upon the ways that many, including non-natives and absolute strangers, can feel at home in any place, depending on the person and the place.

Tuan's generally positive disposition and his admiration for T. E. Lawrence's *Seven Pillars of Wisdom* may have led him to overlook the brazen Orientalism, colonial designs, and frequently racist ideas that accompanied the Englishmen's affinity for North African or Middle Eastern terrain. For example, in *Orientalism*, Edward Said shows how Lawrence's consideration of "the Arab" was in many ways much like the psycho-geography of the desert in Tuan, for this race, like the space it inhabits, is primitive, pure, and timeless (229–231). Indeed, there is something vaguely ominous in Tuan's otherwise cheery sense of "the appeal of the opposite" when one considers the *mission civilisatrice* that functioned as the ideological foundation of direct imperialist conquest.[15] The otherwise innocent preference for the exotic environment of a foreign land may be revealed to entail, in the fullness of time, the colonization of territories and the extension of empire into new spaces on the map. Borges's imperial geographers, as we well can surmise, were not merely mapping an Empire out of intellectual curiosity or scientific scruples, but at least in part as a means of extending power over this territory and its inhabitants.

The map is remarkable thing. It is among the most useful and flexible tools available to mankind, offering a strictly figurative representation of a given territory while at the same time serving as the most practical guide. As Gilles Deleuze and Félix Guattari have asserted, "The map is open and connectable in all its dimensions; it is detachable, reversible, susceptible to constant modification. It can be torn, reversed, adapted to any kind of mounting, reworked by an individual, group,

[15] See, e.g., Edward Said, *Orientalism* (New York: Vintage, 1978), 54–57.

or social formation. It can be drawn on a wall, conceived as a work of art, constructed as a political action or as a mediation."[16] One does not normally associate mere works of art, whose realism is at best a measure of the artist's own choices of metaphor or simile, with the everyday, nuts-and-bolts business of going from point A to point B in the "real world." And yet all recognize the degree to which a map, even the fantastic maps of Borges's fabled cartographers, is an allegorical device. It is a fiction, not unlike a story, that employs any number of figural means to imaginatively depict, not the real territory, but an alternative version of it. Significantly, perhaps, the usefulness of a map is directly related to its being a work of fiction, or in other words a non-mimetic representation of the territory it is supposed to depict.

"What a useful thing a pocket-map is!" remarks the narrator during a memorable scene in Lewis Carroll's *Sylvie and Bruno Concluded*, a scene often thought to be the inspiration for Borges's account in "On Exactitude in Science." In *Sylvie and Bruno Concluded*, Carroll includes as part of a conversion between the titular heroes and one Mein Herr a brief discussion of maps. Mein Herr confesses that he had just lost his way, so that he needed to consult his pocket-map. This then leads to the comment about how useful this item can be, leading Mein Herr to discourse upon the relative value of maps drawn up on different scales:

> "That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"
>
> "About six inches to the mile."
>
> "Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"
>
> "Have you used it much?" I enquired.
>
> "It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."[17]

As I mentioned above, this version is much more cheerful and humorous. The grand map is drafted, but never unfurled, and the territory is allowed to serve as its own map. In his own variation on the theme of the map coextensive with the territory it purports to represent, Neil Gaiman has extracted a more distinctively literary lesson from these parables, asserting that "One describes a tale best by telling the tale. [...] The tale is the map which is the territory."[18]

[16] Gilles Deleuze and Félix Guattari, *A Thousand Plateaus*, trans. Brian Massumi (Minneapolis: University of Minnesota Press, 1987), 12.

[17] Carroll, *Sylvie and Bruno Concluded*, 168–169.

[18] Neil Gaiman, *Fragile Things: Short Fictions and Wonders* (New York: HarperCollins, 2006), xix–xx.

Northrop Frye, in his broader discussion of the ways that literary criticism can be likened to mapmaking with respect to the territory of literature, astutely highlights the word "nearly" in Carroll's story. Much as the farmers and others in Mein Herr's country may feel that they can simply inhabit the map, the urgency of the cartographic imperative cannot easily be suppressed.[19] Furthermore, as Frye puts it, "Surely there must be a middle ground between a map that tells us nothing about the territory and a map that attempts to replace it."[20]

The deserts of cartography, those wastelands of representation, speak to the sense of the world system in the present moment, which is not only unable to represent it adequately, but which can only imagine it as something terrible, impersonal, and ultimately fatal. How does one construct a working map under these circumstances? How does one dwell in the remnants of the great maps? Can we find ways to map anew, to produce cartographies of the future worthy of living beings, as opposed to ghosts, the undead, and others who do not truly live.

The work of art itself offers a clue. In his meditation on the origin of the work of art, Heidegger distinguishes between the world and the earth, which may provisionally be understood as the social and historical project of our own existence, on the one hand, and the natural or material conditions of our environment on the other (even if Heidegger would not necessarily put it that way). In some respects, I believe, these might be reimagined as the map and the territory as well. These two spatial dimensions inform not only our being, but also our projects, the means by which we give our lives and works meaning. In Heidegger's words, "The setting up of a world and the setting forth of earth are the two essential features in the work-being of the work."[21] Jameson has discussed this Heideggerian distinction, underscoring the rift between the terms. As Jameson explains,

> The force of Heidegger's account lies in the way in which a constitutive gap between these two dimensions is maintain and even systematically enlarged: the implication that we all live in both dimensions at once, in some irreconcilable simultaneity, at all moments both in History and in Matter, at one and the same time historical beings and "natural" ones, living simultaneously in the meaning-endowment of the historical project and in the meaninglessness of organic life. But this in turn implies no only that no philosophical or aesthetic synthesis between these dimensions is attainable, but also that "idealism" or "metaphysics" can be defined by this impossible project, whose logical alternatives are marked out by the obliteration of history and its assimilation to Nature, or by the transformation of all forms of natural resistance into human, historical terms.[22]

[19] See my forthcoming *Topophrenia: Place, Narrative, and the Spatial Imagination* (Bloomington: Indiana University Press, 2018).

[20] Northrop Frye, "Maps and Territories," *The Secular Scripture and Other Writings on Critical Theory, 1976–1991*, eds. Joseph Adamson and Jean Wilson (Toronto: University of Toronto Press, 2006), 439.

[21] Heidegger, "The Origin of the Work of Art," trans. Albert Hofstadter, in Heidegger, *Basic Writings*, 172.

[22] Jameson, *Raymond Chandler: The Detections of Totality* (London: Verso, 2016), 77.

If, for Heidegger, any symbolic means of overcoming this rift or attempts to unify these dimensions of world and earth invariably lead to error, then one might suggest that the alternative lies with inhabiting the rift, learning to live with ghosts (as Derrida has put it).[23] Mapping, along with other forms of aesthetic production, is a key means by which we makes this space inhabitable for ourselves. In Jameson's words, "The function of the work of art is then to open a space in which we are ourselves called upon to live within this tension and to affirm its reality."[24]

The work of art, in this case, may well be the map itself, which in a perverse turn of events—the ruse of history or the dialectical reversal—turns out to be the territory after all, but only insofar as the artist-cartographer is prudent. Indeed, if the attribution is to be believed, Borges's "On Exactitude in Science" comes from a work titled *Travels of Prudent Men* by Suárez Miranda, and it makes sense that a prudent traveler in the empire of lost cartography would make note of the remnants of the map scattered across the territory it purported to depict. Prudence dictates caution, after all, particularly with respect to speculation, and the wisdom associated with prudence is always both pragmatic and principled. The prudent artist does not confuse the representation for its referent, and the artist cannot dwell within the work of art. However, the artist gives shape to the world though the work of art, just as the cartographer figures forth the world in attempting to figuratively represent a given territory. In this way, the map and the territory maintain themselves in a somewhat uneasy, yet lasting equipoise in our minds and in our experience. Building a place for ourselves in the deserts of cartography, we dwell in the place that is meaningful only insofar as it may be mapped.

[23] See Jacques Derrida, *Specters of Marx*, translated by Peggy Kamuf (London: Routledge, 1994), xviii: "If it—learning to live—remains to be done, it can happen only between life and death. Neither in life nor in death *alone*. What happens between the two, and between all the "two's" one likes, such as life and death, can only *maintain itself* with some ghost, can *only talk with or about* some ghost. So it would be necessary to learn spirits [...] to learn to live *with* ghosts, in the upkeep, the conversation, the company, or the companionship, the commerce without commerce of ghosts. To live otherwise, and better."

[24] Jameson, *Raymond Chandler*, 78.

Chapter 33
Territory, Geographic Information, and the Map

Donald G. Janelle and Michael F. Goodchild

Introduction

From ancient times, geographers and cartographers have recognized that the map is not the territory. Nonetheless, maps have always played an important role for cataloging, displaying, exploring, and analyzing reality at geographic scales. Maps are designed for many purposes—to enable discovery of patterns and relationships at local through global scales, to facilitate navigation/wayfinding, and to document the content and uses of space, among many other applications. For centuries, these multiple uses of maps depended on traditions of arduous field surveys, manual drafting, and laborious printing processes, all of them labor-intensive and expensive. Nevertheless, the multiple uses that a map may support were recognized by government and private administrators, military commanders, explorers, travelers, and academic researchers as contributing benefits that justified such expenditures. Over the past few decades, the shift from paper maps to maps that are derivative of spatial databases through applications of mapping software, geographic information systems (GIS), and other visualization and display technologies, has contributed to a profusion of new map uses (many unanticipated). In addition, this integration of new technologies has led to wide-spread public reliance on spatial-data displays that are different in legacy and in fundamental properties from the traditions of paper maps, and to enhanced flexibility and economies of scale from being able to

D. G. Janelle (✉)
Department of Geography, Center for Spatial Studies, University of California,
Santa Barbara, CA 93106-4060, USA
e-mail: janelle@geog.ucsb.edu

M. F. Goodchild (✉)
Department of Geography, University of California,
Santa Barbara, USA
e-mail: good@geog.ucsb.edu

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2_33

readily discover and make use of diverse (but compatible) geo-referenced data from multiple sources.

A stunning aspect of this transition over the past few decades has been the rapid automation of reciprocal information flows among maps, among territories, and between maps and territories for a broad range of purposes—e.g., weather forecasting, environmental monitoring, disaster response, facility management of geographically distributed real estate, performance assessments and energy audits for engineering infrastructure, merchandising through location-based services and delivery systems, and the list goes on. In addition, the move to digital has allowed a rapid expansion of the possible range of geographic information types, which are no longer constrained by the two-dimensional, largely static nature of map-derived information. The general public is today familiar with the kinds of three-dimensional, rapidly changing information that are employed in current online mapping and wayfinding apps.

All realms of science and the humanities, and businesses, governments, and institutions, have experienced the intensified volumes and speeds of information flows and communication linkages that have altered their professional practices. In many instances, these developments have changed the nature of the materials, environments, and subjects that they service. Although the verdict is out, the profound global integration of information access and communication capabilities, and the emergence of data-driven multi-media immersive environments, harken to outcomes espoused in the writings of Marshall McLuhan (see Cavell 2002) and the images of global consciousness embraced in Pierre Teilhard de Chardin's noosphere (de Chardin 1959).

When the models/maps and the territories/reality are embedded in automated continuous communication with each other, questions about the distinction between map and territory arise. With this concern in mind, this chapter seeks to broaden the scope of our understanding of map-territory relationships in the geographic realm, documenting their mediation through the digital linkage and communication of geographic information. The discussion begins with:

1. identification of some of the key issues regarding the map's association with reality;
2. the development of a schematic representation of alternative sources of maps and territories;
3. a brief introduction to the foundations and research frontiers of geographic information science;
4. examples of integrating space-time geographic information with the needs of science and society; and
5. an argument in support of embedding a culture of critical spatial thinking through education and public discourse.

The Map's Association with Reality

The intentions of map makers and the perceptions and applications of map users are not always congruent, but it is useful to be aware of the knowledge base and thought processes that underlie each. The Dutch artist Johannes Vermeer shows his classic *The Geographer* (Fig. 33.1) in a contemplative pose, surrounded by a globe, maps, and books, and dividers in hand, as if in the act of a thoughtful pause in drawing a map. Referencing this painting, Downs (1997) calls attention to the geographer's gaze beyond the map (out the window), highlighting the importance of observation, connection to the world, and selectivity in what is to appear on the map. In this case, the mediation of the territory-to-map transfer draws on the skill and knowledge of the geographer who makes decisions in the face of incomplete information and with tools of finite and often limited accuracy.

The map maker is assumed to be knowledgeable in the technicalities of map projections and scale, and their distortional effects on the properties of real-world

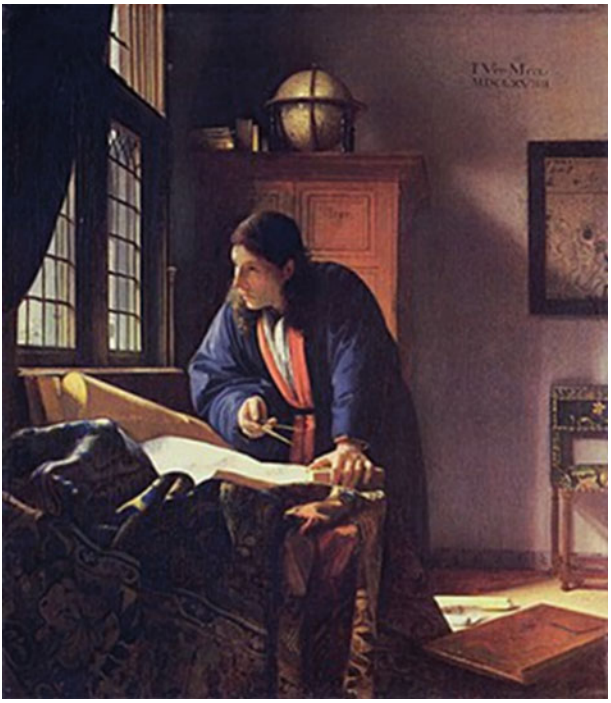


Fig. 33.1 Johannes Vermeer's *The Geographer*, painted in 1668–1669 and currently in the Städelsches Kunstinstitut in Frankfurt, Germany. Source of image: http://en.wikipedia.org/wiki/The_Geographer

spaces, and is cognizant of the limitations in the uses of such maps. However, it is hazardous to assume that all of the myriad users of maps possess such understanding.

Vagueness with respect to map scale and how space and time are visualized, interpreted, or defined has been the basis of delightful literary expositions, references, and imaginative creations of alternative worlds. For instance, in Lewis Carroll's *Alice's Adventures in Wonderland* or *Through the Looking-Glass, and what Alice Found There*, Alice is empowered to vary her size to scale with the behavioral habitats of rabbits, mice, and other creatures, and to assume the role of a pawn in the alternative world and behavioral rules of chess and the constraints of moving about strategically on a chess board. However, what may be appreciated as literature takes on substantive consequences when seeking answers to basic geographic questions for which one expects definitive answers.

Some Issues Regarding Map Generalization

Scale dominates as a basic concept in the literature of the cartographic sciences and geography with the recognition that the density and precision of information displayed on a map or computer screen are constrained, and that the mapping of phenomena is often scale-specific (Montello 1993). Uncertainty in the interpretations based on maps also may result from a lack of understanding about the methods and map projections used to transfer information from globular to planar surfaces, leading to variations of map users' perceptions of shape, size, direction, distance, and network connectivity with regard to local or global scales of observation.

For a specific example, consider the responses to queries about the lengths of coastlines through Web search engines (Table 33.1). For India, presumably authoritative sources list results ranging from 7,000 to 17,181 km but provide no background information about the scale of the map or the precision of the instruments used for the measurements. As Mandelbrot pointed out in his early work on fractals (Mandelbrot 1977), the measured length of a coastline depends directly on

Table 33.1 Measurements of the length of the coastline from Web search

Length (Km)		Authority	URL
India	New Zealand		
7,000	15,134	U.S. Central Intelligence Agency	https://www.cia.gov/library/publications/the-world-factbook/fields/2060.html
17,181	17,209	World Resources Institute	https://en.wikipedia.org/wiki/List_of_countries_by_length_of_coastline

Source Based on Web search by authors

the resolution of the measuring device and the scale of the map; measurement to a finer resolution and finer scale will almost certainly result in a longer result.[1]

Such variations in response may suggest naïveté, even in the case of authorities, in the understanding of basic geographic concepts regarding scale and its measurement. In addition, questions about the attributes of geographic spaces, such as “what is urban?” may yield entirely different responses that reflect variations in historical and geographic context, national standards, or the interpretive license assumed by authors. Such misunderstandings of basic geographic concepts and lack of precision over the meaning of words are among the sources of ambiguities that pose challenges to map makers, confuse map users, and obfuscate the nature of map-territory relationships in the geographic context. In short, they leave the user of a map uncertain about the real nature of the territory being represented.

The transition from sheet maps to digital maps has created additional ambiguities for assessing correspondence between maps and reality. Take, for instance the concept of representative fraction (RF, shown, e.g., as 1:100,000 or 1:1,000,000), used for decades to express the ratio of a distance unit on the map to distance in the same units on the earth’s surface, and widely used for national topographic map series and for atlas maps. In an exhaustive consideration of scale for digital maps as computer screen displays, Goodchild and Proctor (1997) illustrate how RF is meaningless when map users can zoom-in to observe smaller areas in greater detail or zoom-out to expose larger areas of the earth’s surface, or project what is shown on a small computer to a much larger display screen.

Mapping agencies (e.g., the U.S. Geological Survey or the Ordnance Survey of Great Britain) traditionally have adopted standard rules to guide cartographers in the amount and categories of information that could be displayed on maps of different scales. Adapting map information (geo-referenced content) for zooming in and out, digital geo-browsers (e.g., Google Earth) and geographic information systems (GIS) identify hierarchies of information content (such as names of places according to categories of size, streets according to their capacities for traffic, and hydrographic and geomorphic features according to their geographic extent) that appear automatically as a user zooms in or out at different scales. In general, the creators of sheet maps and digital display systems draw upon rules and standards to guide these transitions in information content, but users are left on their own to understand the rationale behind what is displayed and how this might thwart or enhance their information needs.

Vermeer’s *The Geographer* invites speculation about the map maker in selecting information to map, the provenance of the information, and the relationship of mapped observations to a broader geographic context. However, our brief exposure to the implications of map generalization raise questions about the map users’ understandings of basic spatial concepts that could influence the validity of their interpretations and their applications of maps. As the flow of digital data intensifies

[1] Length will also be affected by the way coastline is defined in practice: should estuaries be measured, and what about offshore islands and enclosed lagoons?

in quantity and speed of movement through space and time, a more explicit centrality to geographic information has emerged as it becomes a primary intermediary in facilitating communication between territories and maps.

Geographic Information

Geographic information refers to any information about any subject that has a defined location, be it a street address, latitude-longitude coordinate, telephone area code, postal code, or compartmentalization of Earth space (e.g., city block, county, nation, and continent). Such geo-referencing of information is, in all likelihood, also temporally coded as points or intervals on a time scale of any duration, measured from micro-seconds to the outer margins of expanding geological and cosmic time scales. However, this discussion is limited to the realm of geographic interest—the scale of Earth's surface and near surface and her inhabitants. "Any information about any subject" means precisely that—as long as it is coded with locational and temporal attributes. The subject could be an individual human being, a microscopic plant, music, human beliefs and opinions, or more traditionally accepted notions of geographic phenomena, such as rivers, lakes, boundaries, ridges, and place names (see Skupin and Fabrikant 2003). It could also refer to events which have defined locations, be they ephemeral events, episodes of variable duration, or repetitive periodic events. With this very broad interpretation of geographic information in mind, it is possible to focus on the directionality of information flows to identify sources of maps and territories.

The Source of Maps and Territories

Increasingly, the map-to-territory transition is monitored and controlled through the automated integration of contemporary information, communication, spatial, and sensor technologies. This integration of technologies and their uses alters the traditional understandings about the correspondence between reality (territory) and the model (map). In Fig. 33.2, the dashed boundaries around the oval representations for territory and map reflect their often-indeterminate extent, permeability, and potentials for contraction and expansion. From linked multi-purpose geo-coded and time-stamped digital data sources, researchers can include or screen out various forms of information to create a seemingly infinite array of possible mappings from the same databases (see Monmonier 1996).

Some common examples of the linkages shown in Fig. 33.2 are suggested to add clarity to the illustration. In keeping with the theme of this chapter, the examples pertain to geographic spaces.

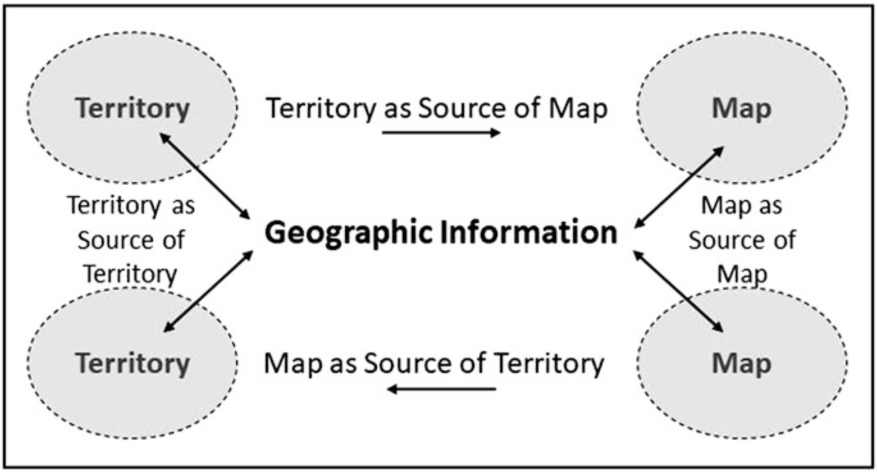

Territory
Territory as Source of Map
Map
Territory as Source of Territory
Geographic Information
Map as Source of Map
Territory
Map as Source of Territory
Map

Fig. 33.2 The centrality of geographic information in understanding territory-map, map-territory, map-map, and territory-territory relationships. Graphic by authors

Territory as Source of Map

This may be the most obvious type of example in reference to geographic space. In this instance, observations of the territory, as recorded in the logs of explorers, in the detailed field notes of land surveyors, in the national censuses of population and other rigorous compilations of attributes and contents of spatial units (e.g., city blocks, planning districts, towns and cities, counties, states and provinces, and nation), or imagery captured via satellites, are transformed by the cartographer into maps of various kinds.

Map as Source of Territory

There are many examples where maps become the source of territories. The design professions (e.g., architecture, urban and regional planning, and engineering) convert ideas into maps and plans that guide the development of territories (buildings, cities, bridges, tunnels, canals, pipelines, transportation networks, electrical grids and other forms of infrastructure that service human-built environments). Such mappings are frequently linked to the temporal staging and organization of interdependent events so that prerequisite resources are in place to support subsequent levels of project completion as well as ongoing operations.

Similarly, but in an entirely different domain of human activity, strategic plans are often represented as flow diagrams and maps to help guide the mobilization of players on sports teams and the marshalling of resources for military campaigns. In these instances, mappings are likely to include contingency plans and alternative actions based on plausible expectations of counter actions by opposition forces. As we know from historical hindsight and personal experiences, the outcomes in games and warfare are often highly uncertain and mapped-out plans, though of significant assistance, are not always prescriptive of what might be preferred.

Maps are also key tools in administration, and in the dominance of one class over another (Harley 2002). Mapping was always a priority activity of colonial powers, as evidenced by the early efforts of the Survey of India (Keay 2000) and remains so in continuing debates over such issues as the partitioning of former Yugoslavia and the Occupied Territories of the West Bank. Eventually such mapping activities lead to changes in the cultural dimensions of territory.

Map(s) as Source of Map(s)

A typical example reflects changes over time in the representation of reality and can be conceived as a projection or prediction to a point in the future based on current and past mappings. The weather forecast or storm-tracking models seek to represent likely events and paths of movement based, in-part, on the hindsight of past such representations of the same event or prior similar events.

Another example stems from the concept of spatial autocorrelation, whereby near neighbors in space are expected to exhibit greater similarity with each other than more distant neighbors. This is frequently referred to as Tobler's First Law of Geography, "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). This simple and common-sense insight from Waldo Tobler (a computational cartographer and spatial theoretician of geography) has become a foundation concept that (often in conjunction with measures of temporal autocorrelation) has helped foster research progress in spatial demography, spatial econometrics, spatial epidemiology, and other fields that rely on rigorous understanding of spatial patterns and processes (Sui 2004).

Territories as Source of Territories

Replication and imitation occur frequently (though not necessarily with perfection) across geographic space. Take the example of place names and the replication of the name "London" or "Moscow" or "Paris" far beyond their points of origin. At neighborhood scales throughout the larger North American cities, there are instances of "Little Tokyo" "Little Italy", "Little Saigon", and "China Town" and other similar designations, often accompanied with street names, public squares, architectural features, and monuments that attempt to capture historical ties, diet preferences, lifestyles, memories, and similar links to otherwise distant territories.

The concept of a “geographic analog” appears frequently in planning disciplines and in scientific research. A geographic analog equates the attributes of another (often distant) place or region with the attributes of a local place or region. For instance, a planner might identify an analog location with similar population and housing characteristics as a strategy to evaluate the likely outcomes from alternative decisions regarding the implementation of a new housing development for the local area. Similarly, an environmental scientist may seek out research sites that match the conditions that prevailed in similar research elsewhere in order to investigate a process of environmental change, and to establish a baseline for comparison in reporting research findings.

The four general types of territory-map relationships described above provide a basis for reviewing in more detail their association with inventories and investigations of the natural and social processes that occur in geographic space at the surface and near-surface of Earth. The review that follows draws on contemporary approaches in geographic information systems and science as the primary source of mapped representations, applications, and analysis of spatial data in diverse fields of scientific inquiry.

Geographic Information Systems and Science

Although geographic information science (GIScience) has its origins in the digital age, it draws on a rich legacy of geographic concepts and interest in spatial analysis that pre-date the emergence of wide-spread computer applications in mapping and spatial analysis. Approaches to gathering, representing, and analyzing geographic data have benefited significantly from computerized information and communication technologies.

Computerized geographic information systems (GIS) appeared initially in the mid-1960s (Thompson and Petchenik 1988) and have developed rapidly into increasingly sophisticated software systems for acquiring, processing, managing, displaying, and analyzing spatial data that are stored as raster and/or vector database layers (Longley et al. 2015). Layers of raster data capture information in a grid structure where each grid cell represents a specific value, while vector layers focus on the specification of geo-referenced points, lines, and polygons. The phenomena represented by rasters or vectors may be distributed continuously across space (e.g., annual precipitation or altitude), or may be conceptualized as discrete spatial entities (e.g., a church, railroad line, city, or country). Raster cells or vector entities may be associated with any number of attributes, which are stored in associated data tables. For example, the layer pertaining to lakes may contain a number of descriptive variables for each lake (area, altitude, shoreline land uses, measures of water acidity, and so on). By overlaying layers of spatial data digitally, researchers, commercial firms, government scientists and bureaucrats, and others can search for relationships, monitor changes over time through visual displays or through analytical tools and modeling systems that are included within standard GIS software.

Unlike the traditional notion of a map on a sheet of paper, digital representations from databases are potentially much broader in scope and significantly more flexible, allowing for a range of exploratory representations and analyses, and contributing efficiencies for handling much larger quantities of data than had been possible in the past. New uses of GIS for management of land and other resources, and for scientific studies in the natural and social sciences, have exploited novel data sources and have augmented technical and theoretical insights from cooperative studies with researchers from closely related areas of investigation, most prominently from the cognitive, computer, and information sciences. Goodchild (1992) referred to these developments as geographic information science. GIScience is now supported by dozens of scientific journals, national and international academic organizations and conferences, and hundreds of academic degree-granting programs and research centers distributed broadly among colleges and universities around the world.

Geographic Information Science (GIScience)

GIScience seeks to identify the fundamental issues of acquiring and using spatial data for representation of patterns and processes that occur on the surface and near surface of the Earth. These issues are broad in scope, including how to assess spatial data accuracy, uncertainty in geographic data, cognition of uncertainty in the design and in the interpretation or use of representations, and the integration of GIScience into science generally and, also, into society.

GIScience is also an empirical discipline concerned with the identification of principles and theories regarding the form of phenomena on or near the Earth's surface; Tobler's First Law is often identified as the most important and practically useful of such principles (Anselin 1989). Readers seeking more detailed discussion of these issues and approaches for handling them may wish to consult widely accepted textbooks about GIS and GIScience (e.g., see Longley, et al. 2015). It is useful at this stage to itemize some of the many new sources of spatial data that GIScience and supportive disciplines draw on to improve descriptive and explanatory understanding of Earth's territories. Many of these new data sources have resulted from the integration of information and communication systems with geographically distributed sensor technologies and remote imagery from satellite-based sensors. But, in addition, historical records and traditional data sources have been digitized and made available for public access and analyses that expand details about human occupancy and use of geographic space over a broader temporal span. Examples are numerous and some are notable. For instance, census data reaching back in time to the late 1700s in the United States document decadal shifts in regional development, demographics, and family histories[2]; church records of births, deaths, and marriages in Europe since the 15th century expose shifting cultural influences from migrations and territorial conflicts; and extensive biographical records for China facilitate spatial analysis of social networks and marriage linkages across administrative regions from the 7th through 19th centuries.[3]

To suggest that the world is or might someday become fully instrumented to document, archive, and process data about its natural and human characteristics and processes would be a severe over-statement. Nonetheless, selective and piecemeal applications of an emergent geospatial cyberinfrastructure (Yang et al. 2010) are possible to service critical needs in support of human activities and general welfare, corporate efficiency, and scientific research. This infrastructure, conceived broadly, harnesses the networked integration of geographically distributed sensor technologies with computational, information, and communication technologies, plus expertise. Sensors that can transmit data from distant locations are widely used for remote environmental monitoring for pollution and air quality, tracking water resources and flood risks, controlling chemical and diverse fluid flows in industrial processes and pipeline systems, assessing energy use for buildings and other infrastructure, measuring structural stress and performance indicators associated with bridges, dams, and other major engineering projects, and for tracking vehicles and deliveries in transport.

To the listing of Earth-bound fixed and mobile sensors, satellite remote-sensing imagery adds a global perspective, capturing information at periodic intervals ranging from near-continuous in seconds and minutes to a few days apart. Aside from the value of having broad global coverage, the information retrieved from such imagery is of special significance for building a knowledge base about the environments of regions on Earth that remain otherwise inaccessible.

The integration of these and many other different data streams with computational resources and online communication networks constitute an expanding data universe of time-stamped and geo-referenced data, akin to what Colwell (2004) referred to as a geographic portal for science. Through Web portals and cloud-storage sites, these data can be drawn upon for use in creating and disseminating information as maps, graphs, animations, immersive environments, and other forms of output in support of research, applications, and education.

[2] The National Historical Geographic Information System (NHGIS) provides population, housing, agricultural, and economic data, for geographic units in the United States from 1790 to the present—see https://www.nhgis.org/.

[3] For information about the China Biographical Database, CBDB, see https://projects.iq.harvard.edu/cbdb; a related database is the China Historical Geographic Information System, CHGIS, featuring place names and historical administrative units of Chinese Dynasties—see https://sites.fas.harvard.edu/~chgis/.

Integrating Space-Time Geographic Information with the Needs of Science and Society

In this section, we present examples of how space-time geographic information enhances scientific understanding of natural and human behavioral processes. In many cases the tools that make mapping and modeling and their dissemination easier and more effective (e.g., the Global Positioning System (GPS),[4] cell phones, and the Internet) are also used in ways that re-shape geographic territories and their associated human activity patterns. For instance, it is now widely accepted that the advent and growth of online shopping via Internet sites, such as Amazon.com, have altered the land use composition of cities. Thus, many shopping activities have transferred from purchases at local stores to direct home delivery from very large product-aggregation-and-distribution centers that service national, continental, and global markets. As a consequence, some shopping opportunities (e.g., book and music stores) have disappeared from local markets. In addition, other kinds of human-interactive behaviors have been facilitated that were not extant prior to the 21st Century—for instance, the use of a smart-phone app to identify whether or not friends are within walking distance from where one is located. We are learning about the impacts of such new behaviors as they become adopted for widespread use, and we can expect that some of these new capabilities for human interactions may have transformative implications for territory-map relationships going forward.

The breadth and diversity of implementation of new services made possible by the digital integration of information and communication technologies (ICT) with the technologies for spatial data analysis are illustrated through a number of anecdotal and documented cases, as follows:

a. As an example of how GIScience and ICT are shaping scientific practices, geomorphologists from England have documented how the study of geomorphology has entered a new era of capabilities for mapping and modeling processes of physical landform development (Smith and Griffiths 2017). These include the ability to transition from two-dimensional maps to three-dimensional renderings of surface and subsurface depictions by linking geological and geographic information sources, and to augment these with temporal data to better understand the forces that shape the Earth's surface. Remote sensing has expanded the scope for investigating larger areas of the Earth's surface and the increasing use of methods such as Light Detection and Ranging (LiDAR) and digital photogrammetry have greatly enhanced the acquisition, processing, and representation of geomorphic landscapes in three dimensions, with the goal of transitioning to multi-modal immersive displays in the near future.
b. Synoptic climatology illustrates the importance of building long-term temporal and spatial data archives. Exploratory data and graphic information systems help

[4] GPS was developed by the U.S, Department of Defense; emulations, such as Russia's GLONASS and China's BeiDou, are collectively referred to as GNSS (global navigation satellite systems).

to decode and represent salient discoveries from archives that date back for more than a century. This allows for data aggregation at a variety of temporal and spatial scales. While this has been an approach pioneered through collections of meteorological data to improve weather forecasts, model storm tracks, and advance climatological research, the building of comparable data archiving systems for other areas of research has also been important, as for example in glaciology, hydrology, oceanography, and seismology.

c. Sensors are increasingly used for the monitoring and management of human activities, and the data from such sensors are being communicated, integrated, and processed in support of decision-making. Buildings, bridges, dams and other structures are being monitored to provide early warning of problems. Traffic is being monitored through the use of overhead cameras and loop detectors. In the rapidly expanding field of Precision Agriculture, sensors installed on agricultural equipment are being used to adjust the application of fertilizers and herbicides, and to provide more accurate, spatially detailed, and timely geographic information on production.

d. Citizens are increasingly engaged not only in the use but also in the acquisition of geographic information (Goodchild 2007; Sui et al. 2012). Programs such as Waze[5] enlist drivers in capturing information on traffic conditions to improve navigation; OpenStreetMap[6] works with volunteers to create accurate base maps, often from fine-resolution imagery; back-yard weather stations feed real-time observations to a central Web interface for dissemination of fine-resolution information on evolving weather patterns and processes (e.g., Weather Underground[7]); and companies such as Google rely heavily on citizens to correct and update their map databases.

e. Automated public alert systems have proliferated in recent years, providing early warning for earthquakes, tsunamis, forest fires, and hazardous weather events. These systems may rely in part on citizens, as in the real-time scanning of Twitter messages or other Internet content for references to events (Li and Goodchild 2010; Zhong et al. 2016). In addition, integrated infrastructure networks of sensors, computational modeling, emergency alert systems, and general news dissemination now provide projected estimates of the paths, severity, and timing of impending impacts at specific geographic locations and across geographic territories.

f. Human beings are on the move as they engage in daily life—traveling to work and school, shopping, enjoying recreational activities, and carrying out a host of obligatory and discretionary activities. Social science researchers, urban planners, and emergency response managers have been interested in documenting

[5] https://www.waze.com/, provides a community-based GPS traffic and navigation app that allows users to contribute information on traffic jams, accidents, and road changes, and to setup carpooling arrangements.

[6] https://www.openstreetmap.org/.

[7] https://www.wunderground.com/.

such activities for many decades. Until recently, they made use of time diaries from volunteers and conducted travel surveys to estimate road usage and population densities at different times of the day and days of the week. However, today, geo-coded data streams from Internet-based social networks and from the widely distributed use of GPS-equipped smart phones provide the means for continuously mapping near real-time human presence in space. These data streams are captured for vehicle navigation systems and for online displays of traffic volumes on road networks. In addition, location-based services (LBS) have commercialized and socialized geographic information to promote business and to facilitate human interactions in real-time (Goodchild 2009). Examples include targeted advertising based on cell-phone users' locations, e.g., Foursquare,[8] and for finding where one's friends are at any given time, e.g., Swarm.[9]

g. As noted earlier, Downs (1997) interprets Vermeer's *The Geographer* as needing to combine the content of maps with a personal ability to know and sense the real world (the territory). In today's terms, we would say that virtual reality (the map, and by extension the GIS) is being compared with augmented reality (a combination of the map and the territory). Both augmented and virtual realities are implemented in today's technology in the form of head-mounted devices that replace the entire field of view (virtual), devices that replace part of the field of view (augmented, e.g., Google Glass), and apps that attempt to match what is seen on the screen to the user's field of view. Clearly, digitalization has been instrumental in breaking the stereotype that maps are limited to two-dimensional views of reality. Virtual reality and augmented reality technologies offer significant opportunities to step beyond two-dimensional views of mapped geographic space.

Building a Culture of Spatial Literacy in the Digital Age

The importance of building a culture of spatial thinkers is a theme that the authors and many other scholars have taken up in recent years (e.g., Goodchild and Janelle 2004; National Research Council 2006; Sinton 2013). It stems, first and foremost, from a conviction that spatial is indeed special, that working with spatial data cannot be approached through a minor modification of standard practices, but must be addressed ab initio. For example, the techniques of inferential statistics were developed for controlled experiments in which a defined population was sampled by giving every member of the population an independent and equal chance of being selected; and the sample was then analyzed to make inferences about the population. These assumptions are rarely if ever tenable for spatial data that have

[8] https://www.foursquare.com/.

[9] https://www.swarmapp.com/.

been obtained through uncontrolled or *natural* experiments, from phenomena that according to Tobler's First Law are almost always autocorrelated. Data on the census tracts of Los Angeles, for example, cannot reasonably be regarded as independent or as a random sample of any larger population. Spatial data also exhibit spatial heterogeneity (Anselin 1989), implying that it is almost never possible to generalize from a limited study area to a larger slice of geography or to the entire Earth.

Good practice to a spatial thinker involves an awareness of the importance of scale and resolution. For example, the natural terrain surface exhibits discontinuities of slope and thus is not everywhere differentiable, implying that any measure of slope must be specific to resolution. An awareness of the importance of uncertainty is also essential, along with an understanding of how uncertainty can be modeled (Zhang and Goodchild 2002) and visualized, and propagated through the stages of analysis and modeling (Heuvelink 1998).

Although reference has already been made to the extent to which we are now able to step beyond the constraints of the traditional paper map, the metaphor of the map retains a strong hold on our thinking. The layer is still the primary means for organizing geographic information in today's technology, and a GIS is still frequently described to a lay audience as a digital container of maps. Traditional practices always favored the flat paper map over the globe, given the difficulties of producing globes in large numbers, and the complications of shipping and storing them. But in a digital world these advantages of flat paper maps disappear. Yet projection, the process of flattening the Earth so that it can be portrayed in a map, remains a large part of teaching about GIS even today, despite the popular success of digital globes such as Google Earth. The conceptual difficulties of dealing with the distortions that result from projection continue to confound our use of geographic information at global scales (e.g., National Research Council 2006, p.146).

Spatial thinking is an amalgam of concepts (knowledge), tools (for spatial analysis and representation), and reasoning (spatial cognition and ways of thinking). This chapter has focused on those spatial concepts and tools that are most germane in the geographic context, such as distance, direction, spatial heterogeneity, analogs, and measures of relationships of phenomena within and across geographic spaces. These may not be the most central concepts of spatial thinking in all disciplines, but we argue that there is currently insufficient attention to developing a general approach to building a culture of spatial literacy in education at all levels and in public media presentations.

The rapid growth of GIS in the global economy and its adoption as a research tool by scholars from dozens of disciplines (see Brunn and Dodge 2017) demonstrate a need for inclusion of fundamental geographic principles and skills as an important component of general education. Early adopters of GIS and GIScience have been the environmental sciences, various areas of resource management and planning, and, especially in the past two decades, the social and health sciences. More recently, the terms geo-humanities and digital humanities have attracted the attention of historians, religious scholars, linguists, and others interested in the geographic dimensions of human culture and the arts. Nonetheless, many scholars in the humanities see the concept of *place* as more central to their needs than that of *space* and would prefer that GIS have more seamless integration with the temporal dimension of human history than is currently the case (See Bodenhamer et al. 2010).

Many other disciplines probe the frontiers of science at fundamentally different scales than geography and GIScience (e.g., astronomy, biology, chemistry, physics, and several branches of engineering) and with different core spatial concerns (Grossner and Janelle 2014). Nonetheless, geographic metaphors such as *map* are used widely across disciplines for which spatial descriptions and representations of the Earth's surface layer are not central. This may be the case, in part, because human beings, regardless of training, all share human experiences at geographic scales. Design-oriented disciplines give more attention to spatial capabilities, such as mental rotation and spatial perspective. Other disciplines find greater utility in mathematical models and related graphs and diagrams than in geographic maps and GIS representations. However, animations, interactive visualizations, virtual environments, and forms of augmented reality are currently areas of significant development in computer science, in the arts, and in media-oriented disciplines and professions. Many of these areas of development share value across most, if not all, disciplines and enhance the case for more general inclusion of spatial literacy as a core investment in educating future generations.

Conclusions

In the context of geographic space, "territory-map" relationships have transitioned in recent decades to "territory-geographic information-map" and "map-geographic information-territory" relationships that invite new interpretations of reality in the digital age. GIScience provides examples of how geographic information is enabling expansive capabilities to break down knowledge silos by using location as a basis for linking data from diverse fields (Scholten et al. 2009).

The two-way communication between the map and the territory is increasingly the norm in scientific endeavor. Through successive feedback loops, one imagines a convergence between the territory and the map that melds their individual characteristics into a near unitary system, where the map (the model) and the territory (reality) become progressively less easily distinguishable from one another, though never the same. New data sources and new methods of data acquisition are steadily improving spatial resolution, allowing positions to be measured to decimeters, and providing social data for smaller and smaller aggregations. We know, for example, that the 10 m positional accuracy that is sufficient for many current wayfinding and navigation applications will have to improve to 5 mm or better to support driverless vehicles. But there can never be full convergence, since the territory is infinitely complex, and capable of revealing more and more detail ad infinitum, and exact measurement of position will always be impossible even with the most sophisticated instruments.

Science, through exposure and experience, is increasingly proficient with time-sensitive digital map execution, vector-raster data models, and object-field transformations; and, the general public is increasingly comfortable with the in-out global perspectives provided by geo- or map-browsers with the zooming capabilities of, e.g., Google Earth, and with the use of online weather maps, traffic maps, and general GPS navigation tools. At this stage, we do not fully understand how developments in immersive technologies and augmented reality will alter the trajectories of scientific understanding about geographical environments. Although we do not know now how they might be embedded in the territories of everyday lives of people, work environments, and education, it is important to maintain a critical but experimental and supportive frame of mind to any opportunities they provide to enhance modeling capabilities, solve real-world problems, and create substantive contributions to education.

In the world of data-driven science many have drawn attention to the fact that the data are not the territory, yet much Fourth Paradigm rhetoric forgets this (Hey et al. 2009). Map users need to be aware that no geographic information is ever the truth, to be cognizant of what is missing and/or distorted, and to understand the impact of omissions and distortions on presumed discoveries. We have touched on these themes throughout this chapter. However, the instigation of this message will be most cogent when general education begins to supplement the tools of geographic awareness with the powers of core concepts in spatial thinking or, in the context of this chapter, with the core concepts of geospatial thinking.

References

L Anselin, What is special about spatial data? Alternative perspectives on spatial data analysis. Technical Report 89–4. Santa Barbara, (National Center for Geographic Information and Analysis, CA, 1989)

D.J. Bodenhamer, J. Corrigan, T.M. Harris (eds.), *The Spatial Humanities. GIS and the Future of Humanities Scholarship* (Indiana University Press, Bloomington, IN, 2010)

S.D. Brunn, M. Dodge (eds.), *Mapping Across Academia* (Springer, Dordrecht, NL, 2017). https://doi.org/10.1007/978-94-024-1011-2

R. Cavell, *McLuhan in Space. A Cultural Geography* (University of Toronto Press, Toronto, 2002)

R. Colwell, The new landscape of science: a geographic portal. Ann. Assoc. Am. Geogr. **94**(4), 703–708 (2004)

P.T. de Chardin, *The Phenomenon of Man. With an Introduction by Julian Huxley*; Translated from French by Bernard Wall. New York, Harper (1959)

R.M. Downs, The geographic eye: Seeing through GIS? Trans. GIS **2**(2), 111–121 (1997). https://doi.org/10.1111/j.1467-9671.1997.tb00019.x

M.F. Goodchild, Geographical information science. Int. J. Geogr. Inf. Syst. **6**(1), 31–45 (1992)

M.F. Goodchild, Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. Int. J. Spat. Data Infrastruct. Res. **2**, 24–32 (2007)

M.F. Goodchild, Location-based services, in *Manual of Geographic Information Systems*, ed. by M. Madden (American Society for Photogrammetry and Remote Sensing, Bethesda, MD, 2009), pp. 779–786

M.F. Goodchild, D.G. Janelle, Thinking spatially in the social sciences, in *Spatially Integrated Social Science*, ed. by M.F. Goodchild, D.G. Janelle (Oxford University Press, New York, 2004), pp. 3–22

M.F. Goodchild, J. Proctor, Scale in a digital geographic world. Geogr. Environ. Model. **1**(1), 5–23 (1997)

K. Grossner, D.G. Janelle, Concepts and principles of spatial literacy, in *Space in Mind: Concepts for Spatial Learning and Education*, ed. by D.R. Montello, K. Grossner, D.G. Janelle (The MIT Press, Cambridge, MA, 2014), pp. 239–261

J.B. Harley, *The New Nature of Maps: Essays in the History of Cartography*, ed by P. Laxton (Johns Hopkins University Press, Baltimore, MD, 2002)

G.B.M. Heuvelink, *Error Propagation in Environmental Modelling with GIS* (Taylor and Francis, London, 1998)

A.J.G. Hey, S. Tansley, K.M. Toole, *The Fourth Paradigm: Data-intensive Scientific Discovery* (Microsoft Research, Redmond WA, 2009)

J. Keay, *The Great Arc* (HarperCollins, London, 2000)

L. Li, M.F. Goodchild, The role of social networks in emergency management: A research agenda. Int. J. Inf. Syst. Crisis Response Manag. (IJISCRAM) **2**(4), 49–59 (2010)

P.A. Longley, M.F. Goodchild, D.J. Maguire, D.W. Rhind, *Geographical Information Systems and Science*, 4th edn. (Wiley, Hoboken, NJ, 2015)

B.B. Mandelbrot, *The Fractal Geometry of Nature* (D.W.H. Freeman and Company, New York, 1977)

M. Monmonier, *How to Lie with Maps*, 2nd edn. (University of Chicago Press, Chicago, 1996)

D.R. Montello, Scale and multiple psychologies of space. In *Spatial Information Theory: A Theoretical Basis for GIS,* ed. by A.U. Frank, I. Campari (Springer, 1993) pp. 312–321

National Research Council, *Learning to Think Spatially: GIS as a Support System in the K-12 Curriculum* (The National Academies Press, Washington, DC, 2006), http://www.nap.edu/openbook.php?record_id=11019

H.J. Scholten, R. van de Velde, N. van Manen (eds.), *Geospatial Technology and the Role of Location in Science* (Springer, Dordrecht, The Netherlands, 2009)

D.S. Sinton, *The People's Guide to Spatial Thinking* (National Council for Geographic Education, Washington DC, 2013)

A. Skupin, S.I. Fabrikant, Spatialization methods: A cartographic research agenda for non-geographic information visualization. Cartogr. Geogr. Inf. Sci. **30**(2), 95–115 (2003)

M.J. Smith, J.S. Griffiths, Physical landscapes. In S.D. Brunn, M. Dodge, eds. *Mapping Across Academia* (Springer, Dordrecht, NL, 2017) pp. 23–44. https://doi.org/10.1007/978-94-024-1011-2_2

D.Z. Sui, Tobler's First Law of Geography: a big idea for a small world? Ann. Assoc. Am. Geogr. **94**(2), 269–277 (2004)

D.Z. Sui, S. Elwood, M.F. Goodchild (eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (Springer, New York, 2012)

W. Tobler, A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**(2), 234–240 (1970)

R. F. Tomlinson, B.B. Petchenik (eds.), Reflections on the revolution: the transition from analogue to digital representations of space, 1958–1988. Am. Cartogr **15**(3), 243–334 (1988)

C. Yang, R. Raskin, M. Goodchild, M. Gahegan, Geospatial cyberinfrastructure: past, present and future. Comput. Environ. Urban Syst. **34**, 264–277 (2010). https://doi.org/10.1016/j.compenvurbsys.2010.04.001

J.-X. Zhang, M.F. Goodchild, *Uncertainty in Geographical Information* (Taylor and Francis, New York, 2002)

X. Zhong, M. Duckham, D. Chong, K. Tolhurst, Real-time estimation of wildfire perimeters from curated crowdsourcing. Scientific Reports 6, Article number: 24206 (2016). https://doi.org/10.1038/srep24206

Afterword

Early Maps

Maps have been with us longer than written language. The need to help people find their way arose early, and sketching a map was a simple form of giving and preserving information. Maps facilitated communication between people with different cultures and different spoken languages. They also enabled one to convey and preserve information and did it in such a simple and natural way that they were used thousands of years before written languages developed.

Maps have a much simpler relation to their objects than spoken or written language. They can be understood and used by anyone who has some grasp of structure and can judge whether two structures are similar. Almost everybody can do this, across cultures and languages. One must be able to recognize the nodes, but these are often indicated by pictures: a tree, a river, a well etc. In 1962, a mammoth tusk 36.5 cm long with an engraving that was created approximately 25,000 years ago was discovered in Pavlov in the Czech Republic. It contains a map-like representation of a mountain, a river and valleys and trails around Pavlov. This is generally accepted to be the oldest preserved "orientation map" in the world, a depiction that enables one to find one's way to the depicted items. This was usually the purpose of the oldest maps one has found.

However, not only the neighborhood, but also the stars on the heaven engaged the early humans. On September 12, 1940, a group of caves were discovered in Lascaux in Dordogne in southwestern France. They contained more than 600 wall paintings which are estimated to be around 17,000 years old. Mostly they are of animals and simple maps, but there is also a painting with dots mapping out part of the night sky, including probably the three bright stars Vega, Deneb and Altair.

While most early maps seem designed to guide us through the neighborhood, what then about maps of the heaven? Knowing where the various stars are may help us find our way on the earth and especially on the ocean, but these early star maps were not found close to any ocean or major lake.

Many maps, especially later maps, give us an overview of an area without primarily being designed to guide travelers. Often they gave an overview of the known parts of the world or of a particular country. The Lascaux star dots

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2

represented a group of the most dominant stars and can be regarded as a map that gives their relative location. However, they and many other cave paintings, in Lascaux and other places, are located in pitch dark and often cramped places, many of them hardly accessible for humans.

Were they made to be seen not by humans, but by gods? Or to express feelings? Are they sacred art? Perhaps the star engravings were all three: maps, works of art, and sacred objects?

Orienteering Maps

With the advent of printing, maps were produced abundantly. One might think that the advent of GPS put an end to this. GPS can produce a temporary map on the screen of a mobile phone. However, the temporary map is not necessary, all one needs, is instructions about the directions one should take: right, left or straight ahead.

Nevertheless, nowadays more maps are printed than ever before, to be used in orienteering races all over the world. Orienteerers are no doubt the main users of maps in the world today. This sport, which started late in the 19th century, now attracts up to 25,000 participants in some races. These races have different courses for different age groups and different levels of difficulty, and they often go on for a week, sometimes more.

These maps make use of a lot of different features, special signs for different kinds of details: fences, houses, forest, water, etc. Importantly, they show contours: lines connecting points of equal elevation. On old orienteering maps, up to around 1950, the scale might be 50,000 and 10,000 and can be up to 1:2000 in sprint races, and the vertical distance between contours (somewhat misleadingly called 'equidistance') is never more than 5 meters, down to one meter in flat areas.

Another important feature of maps are the meridians, lines going south-north on the map. Originally these pointed to the geographic poles. So they did also on the early orienteering maps. However, orienteers usually use a compass to find their way, and the compass needle points towards the magnetic poles. In the early days of orienteering, one had to adjust for this deviation from true north or south, by twisting the compass house a few degrees after one had adjusted it to the geographic meridians on the map. It was a great time-saver when in the sixties they started to draw magnetic meridians on the orienteering maps. This is always done nowadays, and one might wonder why one did not get this great idea earlier. The reason is that the location of the magnetic poles changes over time. The angle between geographic north and magnetic north varies correspondingly. This means that maps with meridians directed to the magnetic poles get outdated after a few years. In the early days of orienteering maps were expensive to produce and one wanted them to be useful for a few years, and they had meridians directed towards the geographic poles. Nowadays, they are printed for each race, and beginning orienteers never hear about magnetic deviation.

This change is illustrated when one compares introductions to orienteering written in the early days of orienteering with orienteering today. Early introductions to orienteering contained not only explanation of contours and signs, they also explained the difference between magnetic north and geographic north. (I wrote such an introduction in 1954, it was reprinted and used for some years. However, the section on magnetic deviation is no longer needed, such subtleties have now gone out of fashion and also largely out of memory.)

Maps and Culture

Maps of all kinds are often good guides to culture. They illustrate the activities and the concerns of the users: travel, vegetation, natural resources, temperatures, etc., and they give historical and political information. They show the progression of wars, and they also sometimes tell us about preparations for war. Orienteering maps provide an example. In 1945, when orienteering again became permitted after the German occupation of Norway, one used maps left over after the German army, printed in Germany before the war. Norway did not have the printing facilities needed for maps of this size and quality, so the information was sent to Germany and the maps were printed there. What one did not know, was that in Germany a second set of maps was produced, ready to be used during the German invasion of Norway in 1940. These were left in Norway after the German capitulation and used as orienteering maps, and they had German text printed on them, for example “Als Schiesskarte nicht geeignet.”

This is only a tiny taste of all the information one can read out of maps, in addition to the wealth of information that maps are designed to provide. Maps are typical cultural objects. They are used to convey and store information, and the information they contain tells much about the culture. However, here as in the case of all cultural objects, what is not explicitly stated, but can in various ways can be read out of the objects, is often the most important source of insight in a culture. Cultural objects are sedimentations of practices, to use Husserl’s words, and these sedimentations are often a key to understanding a culture.

Maps and Thinking

Maps are used for lots of other purposes than finding one’s way and showing the terrain. Anything can be represented on a map, and maps have turned out to be valuable tools for thinking. Maps, drawings, graphs, and diagrams combine features of language and image and play a crucial role in our cognition and reasoning. They are used again and again through the history of mathematics. Good notations incorporate the advantages of diagrams. A simple example, familiar to everyone, is the use of Arabic versus Roman numerals. Already in elementary school one learns how to exploit of the Arabic notation, with its utilization of location, of zero etc.

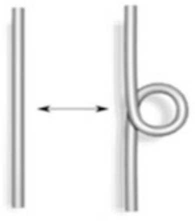
Modern examples are Euler diagrams and their generalizations by Venn and later by Peirce and many others. A concise survey is given in the Stanford Encyclopedia article "Diagrams." Peirce was particularly important in this tradition. He discussed the optimal way of diagramming mathematical and, especially, logical structure, and his many interesting contributions to this field have been developed further in many directions (see the bibliography at the end of this article).

The study of diagrams is now pursued actively not only by mathematicians and logicians, but also in computer science, philosophy, linguistics, architecture, art, music and many other fields. Typical are the many cross-disciplinary groups and conferences where representatives of these various fields get together to learn from one another. There are too many contributors to mention, but should I name one, Kenneth *Manders*, University of Pittsburgh, would be a strong candidate. In 1995 he wrote a paper "The Euclidean diagram," which was not published until 2008, as Chapter 4 in Paolo Mancosu, ed., *The Philosophy of Mathematical Practice*. pp. 80–133. Manders started his work on diagrams already in the 1970s, and he has inspired colleagues in Pittsburgh and several other places, including Stanford, where he stayed at the Center for Advanced Study in the Behavioral Sciences in 1987–88 and has come back for lectures often.

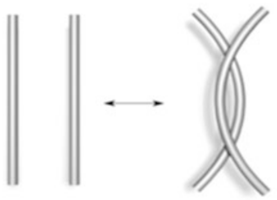
A branch of mathematics that is closely connected with maps, drawings, graphs, and diagrams, is topology. In particular, in the branch of topology called "knot theory" this kind of study has become very fruitful. The study of knots started several centuries BC in Chinese artwork and in Tibetan Buddhism. A mathematical theory of knots was first developed in 1771 by Alexandre-Théophile Vandermonde. In 1926–27 J. W. Alexander and G. W. Briggs, and, independently, Kurt Reidemeister, presented three local moves (known as Reidemeister moves) and proved that any two planar knot diagrams belonging to roughly the same knot can be related by a finite sequence of Reidemeister moves from one of them to the other[1]:

[1] J. W. Alexander and G. W. Briggs, "On types of knotted curves." Annals of Mathematics 28 (1926–27), pp. 562–586.

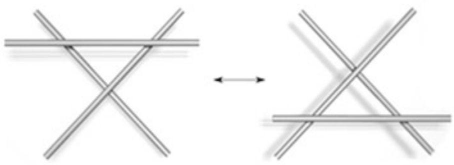
Reidemeister, Kurt, "Elementare Begründung der Knotentheorie", Abh. Math. Sem. Univ. Hamburg, 5 (1) 1927, pp. 24–32. See also Reidemeister's book Knotentheorie (Ergebnisse der Mathematik und ihrer Grenzgebiete, Alte Folge, Bd. 1, Heft 1) Berlin: Springer, 1932, reprinted 1974. English translation: Knot Theory. BCS Associates, Moscow, Idaho, 1983.



Type I Twist and untwist in in either direction



Type II Move one loop completely over another



Type III. Move a string completely over or under a crossing

Reidemeister was a master of using visualizing and diagrams in reasoning. His topology courses in Göttingen, which I followed in the mid-fifties, were pedagogical masterpieces, excellent examples of using diagrams in the teaching of mathematics.

Twenty years before that, also in Göttingen, Herman Weyl wrote an interesting article on the relevance of topology for comprehension, translated into English as “Topology and abstract algebra as two roads of mathematical comprehension.” The American Mathematical Monthly, 1932, 102(5), 453–460. (1995).

As one should expect, there are very many interesting connections between maps and mathematics. I will only mention one, also from topology: the four color problem. It dates back to 1852 when the law student Francis Guthrie, trying to color a map of England’s counties, found he needed four different colors if two regions sharing a border could not share a color. He then conjectured, and attempted to prove, that four colors sufficed to color any map in this way. He asked his brother, who studied with the mathematician/logician Augustus de Morgan, to convey the problem to him. De Morgan became very interested in the problem, but found it very difficult and already the same day wrote to Sir William R. Hamilton in Dublin about it. Later de Morgan also wrote to the Harvard mathematician and astronomer Benjamin Peirce, and through him it reached his son Charles Saunders Peirce, who was then in his twenties and got quite engaged by the problem. He extended the problem to other kinds of surfaces, and constructed a map on a torus that requires 6 colors. Later it was shown that 7 colors are required on a torus.

Several times, mathematicians claimed to have proved the four color conjecture, but all the alleged proofs turned out to be fallacious. In 1975, Martin Gardner in his Scientific American column reported that a map with 110 countries had been devised that required five colors. However, the date of that column was April 1, it was an early case of what is now called "alternative truths."

The next year, finally, the problem was solved, using a technique that for fifty years had been developed to deal with this kind of a problem: one reduced complicated maps to a minimal set of map configurations that could be tested by mere calculation. In 1976, at the University of Illinois, Kenneth Appel and Wolfgang Haken reduced the testing problem to a check of 1936 configurations. These configurations were checked by computer. None of them required more than four colors, and a complete solution to the Four Color Conjecture was thereby achieved. This problem of checking these maps one by one was doubly controlled with different programs and different computers. Their proof showed that there exists no map requiring more than four colors.

What makes the four color problem particularly interesting is that it is the first mathematical problem which has not been solved by thinking alone, but only by thinking assisted by a computer. No human being has gone through the proof. Simpler proofs have been found, but they all require a computer.

In 1979 Thomas Tymoczko wrote an article, "The Four-Color Problem and its Mathematical Significance", The Journal of Philosophy 76 (2): 57–83, where he discusses the philosophical significance of the use of computers in mathematical theorem proving. There and in later articles he argued that proofs should give insight, they shall tell us not only that a conclusion is true, but why it is true. This question spreads to more and more fields as computer-assisted proofs too large to be directly verifiable by humans have become commonplace. Last year a new record was set, a two-hundred-terabyte proof was given that cracked the "Boolean Pythagorean triples problem." 200 terabytes is roughly equivalent to all the digitized text held by the US Library of Congress. But what insight does it give us in the theorem that is proved, except that it is true—but is that an empirical truth about a platonic world, or what is it?

Dagfinn Føllesdal

Bibliography

G. Allwein, J. Barwise (eds.), *Logical Reasoning with Diagrams* (Oxford University Press, New York, 1996)

C. Ambrosio, Iconic representations and representative practices. Int. Stud. Philos. Sci. **28**(3), 255–275 (2014)

M. Anderson, B. Meyer, P. Olivier (eds.), *Diagrammatic Representation and Reasoning* (Springer, London, 2002)

A.P. Atkin, *The Routledge Philosophers* (Taylor & Francis, 2016)

M. Bergman, S. Paavola, A.V. Pietarinen, H. Rydenfelt (eds.), *Ideas in Action: Proceedings of the Applying Peirce Conference* (Nordic Pragmatism Network, Helsinki, 2010)

S. Krämer, C. Ljungberg (eds.), Thinking with diagrams: the semiotic basis of human cognition, in *Semiotics, Communication and Cognition*, 17 (De Gruyter, 2016)

P. Mancosu (ed.), *The Philosophy of Mathematical Practice* (Oxford University Press, 2008)

J. Mumma, M. Panza (eds.), Diagrams in mathematics; history and philosophy. Special issue of Synthese **186** (1) (2012). Includes papers from two workshops, at Stanford 2007 and in Paris 2008, among them papers on two central philosophers who are too seldom discussed in this context: Hume and Kant

R. Netz, Greek mathematical diagrams: Their use and their meaning. Learn. Math. **18**(3), 33–39 (1998)

R. Netz, *The Shaping of Deduction in Greek Mathematics. A Study in Cognitive History* (Cambridge University Press, 1999)

A.-V. Pietarinen, *Logic of the Future* (an edition of Peirce's unpublished writings on his theory of existential graphs) (Forthcoming)

J.F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations* (Brooks Cole Publishing Co., Pacific Grove, CA, ©2000). Actual publication date, 16 Aug 1999. See also his "Existential Graphs" MS 514 by Charles Sanders Peirce, with commentary, forthcoming

F. Stjernfelt, *Diagrammatology: An Investigation on the Borderlines of Phenomenology, Ontology, and Semiotics* (Springer, Dordrecht, 2007)

F. Stjernfelt, *Natural Propositions: The Actuality of Peirce* (Docent Press, Boston, 2014)

S. de Toffoli, 'Chasing' the diagram—the use of visualizations in algebraic reasoning. Rev. Symb. Log. **10**(1), 158–186 (2017)

S. de Toffoli, Forthcoming: (with P. Findlen and G. Priest) Tools of Reason: The Practice of Scientific Diagrammatic from Antiquity to the Present. Special Issue of Endeavour (2018)

Titles in This Series

Quantum Mechanics and Gravity
By Mendel Sachs
Quantum-Classical Correspondence
Dynamical Quantization and the Classical Limit
By A. O. Bolivar
Knowledge and the World: Challenges Beyond the Science Wars
Ed. by M. Carrier, J. Roggenhofer, G. Küppers and P. Blanchard
Quantum-Classical Analogies
By Daniela Dragoman and Mircea Dragoman
Quo Vadis Quantum Mechanics?
Ed. by Avshalom C. Elitzur, Shahar Dolev and Nancy Kolenda
Information and Its Role in Nature
By Juan G. Roederer
Extreme Events in Nature and Society
Ed. by Sergio Albeverio, Volker Jentsch and Holger Kantz
The Thermodynamic Machinery of Life
By Michal Kurzynski
Weak Links
The Universal Key to the Stability of Networks and Complex Systems
By Csermely Peter
The Emerging Physics of Consciousness
Ed. by Jack A. Tuszynski
Quantum Mechanics at the Crossroads
New Perspectives from History, Philosophy and Physics
Ed. by James Evans and Alan S. Thorndike
Mind, Matter and the Implicate Order
By Paavo T. I. Pylkkanen
Particle Metaphysics
A Critical Account of Subatomic Reality
By Brigitte Falkenburg
The Physical Basis of the Direction of Time
By H. Dieter Zeh

S. Wuppuluri and F. A. Doria (eds.), *The Map and the Territory*,
The Frontiers Collection, https://doi.org/10.1007/978-3-319-72478-2

Asymmetry: The Foundation of Information
By Scott J. Muller
Decoherence
and the Quantum-To-Classical Transition
By Maximilian A. Schlosshauer
The Nonlinear Universe
Chaos, Emergence, Life
By Alwyn C. Scott
Quantum Superposition
Counterintuitive Consequences of Coherence, Entanglement, and Interference
By Mark P. Silverman
Symmetry Rules
How Science and Nature are Founded on Symmetry
By Joseph Rosen
Mind, Matter and Quantum Mechanics
By Henry P. Stapp
Entanglement, Information, and the Interpretation of Quantum Mechanics
By Gregg Jaeger
Relativity and the Nature of Spacetime
By Vesselin Petkov
The Biological Evolution of Religious Mind and Behavior
Ed. by Eckart Voland and Wulf Schiefenhövel
Homo Novus-A Human without Illusions
Ed. by Ulrich J. Frey, Charlotte Störmer and Kai P. Willführ
Brain-Computer Interfaces
Revolutionizing Human-Computer Interaction
Ed. by Bernhard Graimann, Brendan Allison and Gert Pfurtscheller
Extreme States of Matter
On Earth and in the Cosmos
By Vladimir E. Fortov
Searching for Extraterrestrial Intelligence
SETI Past, Present, and Future
Ed. by H. Paul Shuch
Essential Building Blocks of Human Nature
Ed. by Ulrich J. Frey, Charlotte Störmer and Kai P. Willführ
Mindful Universe
Quantum Mechanics and the Participating Observer
By Henry P. Stapp
Principles of Evolution
From the Planck Epoch to Complex Multicellular Life
Ed. by Hildegard Meyer-Ortmanns and Stefan Thurner
The Second Law of Economics
Energy, Entropy, and the Origins of Wealth
By Reiner Kümmel
States of Consciousness

Experimental Insights into Meditation, Waking, Sleep and Dreams
Ed. by Dean Cvetkovic and Irena Cosic

Elegance and Enigma
The Quantum Interviews
Ed. by Maximilian Schlosshauer

Humans on Earth
From Origins to Possible Futures
By Filipe Duarte Santos

Evolution 2.0
Implications of Darwinism in Philosophy and the Social and Natural Sciences
Ed. by Martin Brinkworth and Friedel Weinert

Probability in Physics
Ed. by Yemima Ben-Menahem and Meir Hemmo

Chips 2020
A Guide to the Future of Nanoelectronics
Ed. by Bernd Hoefflinger

From the Web to the Grid and Beyond
Computing Paradigms Driven by High-Energy Physics
Ed. by René Brun, Frederico Carminati and Giuliana Galli-Carminati

The Language Phenomenon
Human Communication from Milliseconds to Millennia
Ed. by P.-M. Binder and K. Smith

The Dual Nature of Life
Interplay of the Individual and the Genome
By Gennadiy Zhegunov

Natural Fabrications
Science, Emergence and Consciousness
By William Seager

Ultimate Horizons
Probing the Limits of the Universe
By Helmut Satz

Physics, Nature and Society
A Guide to Order and Complexity in Our World
By Joaquín Marro

Extraterrestrial Altruism
Evolution and Ethics in the Cosmos
Ed. by Douglas A. Vakoch

The Beginning and the End
The Meaning of Life in a Cosmological Perspective
By Clément Vidal

A Brief History of String Theory
From Dual Models to M-Theory
By Dean Rickles

Singularity Hypotheses
A Scientific and Philosophical Assessment

Ed. by Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart

Why More Is Different
Philosophical Issues in Condensed Matter Physics and Complex Systems
Ed. by Brigitte Falkenburg and Margaret Morrison

Questioning the Foundations of Physics
Which of Our Fundamental Assumptions Are Wrong?
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

It From Bit or Bit From It?
On Physics and Information
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

How Should Humanity Steer the Future?
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

Trick or Truth?
The Mysterious Connection Between Physics and Mathematics
Ed. by Anthony Aguirre, Brendan Foster and Zeeya Merali

The Challenge of Chance
A Multidisciplinary Approach from Science and the Humanities
Ed. by Klaas Landsman, Ellen van Wolde

Quantum [Un]Speakables II
Half a Century of Bell's Theorem
Ed. by Reinhold Bertlmann, Anton Zeilinger

Energy, Complexity and Wealth Maximization
Ed. by Robert Ayres

Ancestors, Territoriality and Gods
A Natural History of Religion
By Ina Wunn, Davina Grojnowski

Space, Time and the Limits of Human Understanding
Ed. by Shyam Wuppuluri, Giancarlo Ghirardi

Information and Interaction
Eddington, Wheeler, and the Limits of Knowledge
Ed. by Ian T. Durham, Dean Rickles

The Technological Singularity
Managing the Journey
Ed. by V. Callaghan, J. Miller, R. Yampolskiy, S. Armstrong

How Can Physics Underlie the Mind?
Top-Down Causation in the Human Context
By George Ellis

The Unknown as an Engine for Science
An Essay on the Definite and the Indefinite
Hans J. Pirner

CHIPS 2020 Vol. 2
New Vistas in Nanoelectronics
Ed. by Bernd Hoefflinger

Life – As a Matter of Fat
Lipids in a Membrane Biophysics Perspective

Ole G. Mouritsen, Luis A. Bagatolli
The Quantum World
Philosophical Debates on Quantum Physics
Ed. by Bernard D'Espagnat, Hervé Zwirn
The Seneca Effect
Why Growth is Slow but Collapse is Rapid
By Ugo Bardi
Chemical Complexity
Self-Organization Processes in Molecular Systems
By Alexander S. Mikhailov, Gerhard Ertl
The Essential Tension
Competition, Cooperation and Multilevel Selection in Evolution
By Sonya Bahar
The Computability of the World
How Far Can Science Take Us?
By Bernd-Olaf Küppers
The Map and the Territory
Exploring the Foundations of Science, Thought and Reality
By Shyam Wuppuluri, Francisco Antonio Doria