The Continuum Companion to the Philosophy of Science

Edited by Steven French and Juha Saatsi
The Continuum Companion to the Philosophy of Science
The Continuum Companions series is a major series of single volume companions to key research fields in the humanities, aimed at postgraduate students, scholars and libraries. Each companion offers a comprehensive reference resource giving an overview of key topics, research areas, new directions and a manageable guide to beginning or developing research in the field. A distinctive feature of the series is that each companion provides practical guidance on advanced study and research in the field, including research methods and subject-specific resources.

The Continuum Companion to Continental Philosophy
Edited by John Mullarkey and Beth Lord

The Continuum Companion to Locke
Edited by S.-J. Savonious-Wroth, Paul Schuurman and Jonathan Walmsley

The Continuum Companion to Philosophy of Mind
Edited by James Garvey

Forthcoming in Philosophy:

The Continuum Companion to Aesthetics
Edited by Anna Christina Ribeiro

The Continuum Companion to Berkeley
Edited by Bertil Belfrage and Richard Brook

The Continuum Companion to Epistemology
Edited by Andrew Cullison

The Continuum Companion to Ethics
Edited by Christian Miller

The Continuum Companion to Existentialism
Edited by Jack Reynolds, Felicity Joseph and Ashley Woodward

The Continuum Companion to Hegel
Edited by Allegra de Laurentiis and Jeffrey Edwards
The Continuum Companion to Hobbes
Edited by S. A. Lloyd

The Continuum Companion to Hume
Edited by Alan Bailey and Dan O’Brien

The Continuum Companion to Kant
Edited by Gary Banham, Nigel Hems and Dennis Schulting

The Continuum Companion to Leibniz
Edited by Brendan Look

The Continuum Companion to Metaphysics
Edited by Robert Barnard and Neil A. Manson

The Continuum Companion to Political Philosophy
Edited by Andrew Fiala and Matt Matravers

The Continuum Companion to Plato
Edited by Gerald A. Press

The Continuum Companion to Pragmatism
Edited by Sami Pihlström

The Continuum Companion to Socrates
Edited by John Bussanich and Nicholas D. Smith

The Continuum Companion to Spinoza
Edited by Wiep van Bunge

The Continuum Companion to Philosophy of Language
Edited by Manuel García-Carpintero and Max Kolbel
This page intentionally left blank
The Continuum Companion to the Philosophy of Science

Edited by
Steven French
and
Juha Saatsi
Contents

Contributors ix

1 Introduction 1
Steven French and Juha Saatsi

Part I: Philosophy of Science in Context
2 Philosophy of Science and Epistemology 15
Alexander Bird
3 Philosophy of Science and Metaphysics 33
Craig Callender
4 Philosophy of Science and the History of Science 55
Don Howard

Part II: Current Research and Issues
A. General Issues in Philosophy of Science
5 Scientific Realism with a Humean Face 75
Stathis Psillos
6 Causation and the Sciences 96
Ned Hall
7 Scientific Models and Representation 120
Gabriele Contessa
8 Reduction, Multiple Realizability and Levels of Reality 138
Sven Walter and Markus Eronen
9 Explanation 157
Henk W. de Regt
10 Scientific Evidence 179
Malcolm R. Forster
11 Bayesian Confirmation Theory 197
James Hawthorne
Contents

B. Philosophy of Particular Sciences

12 Philosophy of Physics  
Nick Huggett 221

13 Philosophy of Biology  
Ingo Brigandt 246

14 Towards a Mechanistic Philosophy of Neuroscience  
Carl F. Craver and David M. Kaplan 268

15 Philosophy of Chemistry  
Robin Findlay Hendry 293

16 Philosophy of Mathematics  
Christopher Pincock 314

Part III: Past and Future

17 Travelling in New Directions  
Steven French and Juha Saatsi 337

18 A Brief Chronology of the Philosophy of Science  
Peter Vickers 359

Part IV: Resources

19 Annotated Bibliography 383

20 Research Resources 406

21 A–Z of Key Terms and Concepts 412

Index 439
Contributors

Alexander Bird
Professor
Department of Philosophy
University of Bristol
Bristol, UK

Ingo Brigandt
Assistant Professor
Department of Philosophy
University of Alberta
Edmonton, Canada

Craig Callender
Professor of Philosophy
Department of Philosophy
University of California
San Diego, CA, US

Gabriele Contessa
Assistant Professor
Department of Philosophy
Carleton University
Ottawa, Canada

Carl F. Craver
Associate Professor
Department of Philosophy
Washington University in St. Louis
St. Louis, MO, US

Henk W. de Regt
Associate Professor
Faculty of Philosophy
VU University of Amsterdam
Amsterdam, The Netherlands
Contributors

Markus Eronen
Junior Fellow
HWK Institute for Advanced Study
Delmenhorst, Germany

Malcolm R. Forster
Professor of Philosophy
University of Wisconsin–Madison
Madison, WI, US
and Professor,
Institute for Science, Technology, and Society,
Tsinghua University, China

Steven French
Professor of Philosophy of Science
Department of Philosophy
University of Leeds
Leeds, UK

Ned Hall
Professor
Department of Philosophy
Harvard University
Cambridge, MA, US

James Hawthorne
Associate Professor
Department of Philosophy
University of Oklahoma
Norman, OK, US

Robin Findlay Hendry
Reader
Department of Philosophy
Durham University
Durham, UK
Contributors

Don Howard
Professor of Philosophy
Department of Philosophy and Program in History and Philosophy of Science
University of Notre Dame
Notre Dame, IN, US

Nick Huggett
Professor
Department of Philosophy
University of Illinois at Chicago
Chicago, IL, US

David M. Kaplan
Adjunct Professor
Department of Philosophy
Washington University in St. Louis
St. Louis, MO, US

Christopher Pincock
Associate Professor
Department of Philosophy
Purdue University
West Lafayette, IN, US

Stathis Psillos
Professor of Philosophy of Science and Metaphysics
Department of Philosophy and History of Science
University of Athens
Athens, Greece

Juha Saatsi
Lecturer
Department of Philosophy
University of Leeds
Leeds, UK
Contributors

**Peter Vickers**  
Postdoctoral Fellow  
Center for Philosophy of Science  
University of Pittsburgh  
Pittsburgh, PA, US

**Sven Walter**  
Professor of Philosophy of Mind  
Institute of Cognitive Science  
University of Osnabrück  
Osnabrück, Germany
1 Introduction

Steven French and Juha Saatsi

This Companion provides an up-to-date overview of the philosophy of science, its central issues and its place in a wider disciplinary context. The essays in Part I, ‘Philosophy of Science in Context’, discuss the place of philosophy of science in relation to the rest of philosophy, to science itself and to the history of science. The essays in Part II, ‘Current Research and Issues’, cover many of the discipline’s key debates as conducted today, both in general philosophy of science as well as in regard to particular sciences. Then, Part III, ‘Past and Future’, gazes ahead by reflecting on trends and new issues, emerging both from the contributions included here and also more generally. The whole volume is forward-looking in flavour, with the aim of indicating exciting new research that is ‘hot’ today, and probably even more so tomorrow. This forward-looking emphasis distinguishes the present Companion from the other companions and encyclopaedias on the market. For the sake of completeness and for the volume to serve as a study aid, there is also ‘A Brief Chronology of the Philosophy of Science’, and Part IV, ‘Resources’, that includes ‘A-Z of Key Terms and Concepts’, ‘Annotated Bibliography’, and ‘Research Resources’.

Before we turn to an overview of the individual chapters and some of their interconnections, a few general words about philosophy of science and its disciplinary status are in order. As a discipline in its own right, it is a comparative newcomer, emerging with its own journals, conferences and so forth only in the past 80 years or so. However, its roots go back much further, in scientists’ own reflections on what they were doing and also in the fact that many philosophers, ever since Aristotle, have taken science to fall within their purview.¹ Broadly speaking, philosophy of science covers issues such as the methodology of science, including the role of evidence and observation; the nature of scientific theories and how they relate to the world; and the overall aims of science. It also embraces the philosophies of particular sciences, such as biology, chemistry, physics and neuroscience, and considers the implications of these for such issues as the nature of space-time, the mind-body problem, and the foundations of evolution.

Throughout its development, its relationships with other disciplines and research programmes, both in philosophy and in the history of science, have
been under scrutiny. For example, in the mid-1980s, the philosophy of science was described to one of us as being in a state of degeneration and as being fractured into inward looking cliques and sub-disciplines. It was even strongly suggested that anyone seeking to do productive research in philosophy should look elsewhere. The grounds for such an uncompromising view lay with the belief that the ‘new directions’ for philosophy of science – indicated by Suppe in his magisterial overview from 1977 (Suppe 1977) – had effectively run into the sand. In that work, Suppe set out both the shift away from those perspectives on science associated with the logical positivists, and also the rise and fall of the views of Kuhn and others that emphasized both the importance of the history of science and the purported role of social factors. In their place, Suppe envisioned that there would emerge a philosophy of science based on a robust form of realism that asserts that we can have knowledge of how the world ‘really is’ and that observation, experiment and the methodology of science play a fundamental role in obtaining that knowledge. It was this realism – shaping it, developing it, exploring its metaphysical and epistemological underpinnings – that became the focus of much work over the subsequent decade or so. No longer were philosophers of science attempting to demarcate science from pseudoscience, or presenting grand, large-scale methodological frameworks which they could use to tell scientists how ‘good’ science should be done. Instead, they became preoccupied with what appeared to those outside the field to be minor issues and problems, or were locked into particular and sometimes highly technical subfields, such as the philosophy of physics.

At the same time, the rejection of views that tried to place science in its historical as well as sociopolitical contexts brought in its train a turning away from the history of science. Historians of science, for their part, contributed to what came to be seen as the divorce of the two disciplines by focusing increasingly on the above contexts, to the exclusion of the ‘internal’ forces that philosophers were concerned with, having to do with evidence and its relationship to theories. By the early 1990s this division had become so wide and so entrenched that concerns were raised on both sides and meetings were held to analyse the causes and consider whether bridges could still be built between the two sides (see Steinle and Burian 2002).

Furthermore, despite hopes that the examination of the practices of science would have profound implications for metaphysics and epistemology, the philosophy of science has also seemed to be ploughing its own distinct furrow with only minimal contact with other, hugely significant areas of philosophy. Metaphysics, for example, has experienced a major resurgence in recent years, with exciting new research undertaken on a wide range of topics, such as modality, mereology, the nature of time and so on. It has even generated a new sub-discipline, meta-metaphysics, which examines, for example, whether
metaphysical questions have determinate answers, and if they do, what methodology one would use to choose one metaphysical system over another (see Chalmers et al. 2009). But while all this activity has been going on, the concerns of metaphysicians and the metaphysical views they espouse have come to seem increasingly distant from those of philosophers of science and, indeed, from any grounding in science itself. Metaphysical discussions of the ontological nature of things, for example, have proceeded with little or no consideration of what current science tells us about those things in their most fundamental form. Even where there are obvious points of contact, the two fields seem to be moving along different trajectories: recent metaphysical accounts of scientific laws, for example, appear to rely on some naïve conceptions based on simplistic ‘toy’ examples and make little, if any, mention of conservation laws and symmetry principles which, it can be argued, are just as fundamental for modern physics. This situation has led some to decry contemporary metaphysics as either divorced from science altogether or, at best, paying it only lip service (Ladyman and Ross 2007).

Although less pronounced, one can detect a similar tension between philosophy of science and epistemology. The latter is concerned with the study of knowledge in general and justified belief in particular. In contemporary discussions, a division has appeared between those that regard the justification of belief, and hence knowledge, as depending on internal factors, such as reflection on one’s own cognitive processes, and those that hold it to depend on factors external to us that originate in reliable processes, say. The philosophy of science, on the other hand, is (not surprisingly) more focused on the practices of science, some of which can be understood from an internalist perspective, while others seem more amenable to an externalist understanding. Although science is often held up as the ‘paradigm’ of knowledge-gathering activity, there has been surprisingly little engagement with these practices by epistemologists, and philosophers of science, in turn, have tended not to draw on general epistemological accounts as extensively as they might. The situation is undoubtedly unsatisfactory for both fields, and there is a clear need for a more positive and productive interaction.

So, when it comes to those three areas with which the philosophy of science might be expected to have the closest contact – the history of science, metaphysics and epistemology – significant rifts and divisions have appeared in the last 30 years. In each case, these have hampered work on both sides. However, recent developments have brought signs of hope and indications of new directions opening up.

Thus, the history of science is being engaged again, not only through award-winning books such as Chang’s account of the measurement of temperature (Chang 2004), or Ryckman’s acute analysis of disregarded views of Einstein’s general relativity (Ryckman 2007), but also by means of
interdisciplinary interactions between history and philosophy of science and science studies, and the international conferences of the ‘&HPS’ movement.2

Don Howard has played a leading role in these developments, and in his essay for this collection he maps out the history of the changing relations between the history of science and the philosophy of science. After providing a rich historical overview of the earlier close and productive relations between the two, Howard examines the role played by philosophers themselves in the divorce. Here we face a puzzle: it is usually the logical empiricists who are blamed for the split, but they themselves embraced serious history of science scholarship. Howard draws on the underlying political machinations to provide an answer. Although the ‘left wing’ of the Vienna Circle espoused both the value-ladenness of science and the importance of the history of science for the philosophy of science, it was the ‘right wing’ that gained the ascendancy. Through the promotion of the distinction between how theories are discovered and how they are justified – with philosophy of science to be concerned only with the latter – it pushed out values and denied the philosophical relevance of ‘real’, as opposed to reconstructed, history.

Howard also considers how we got to where we are now, with renewed interest in the history of the philosophy of science and, following the demise of logical empiricism, a new willingness to engage with the history of science. Friedman’s now classic work, *Kant and the Exact Sciences*, is taken as representative of this engagement, as it seeks to construct a framework that both respects the ‘historicity’ of science and also avoids the cheap kind of relativism that comes from insisting ‘different era, different science, different methodology’. This is not the only framework in town, of course, and the interest in appropriately capturing and advancing the ‘HS’ and ‘PS’ relationship is another indication of the health of the discipline.

The relationship between philosophy of science and metaphysics has long been fraught. Craig Callender has spent much of his career working on the metaphysical implications of modern science, particularly physics, and in his contribution he recalls the point made above that, as philosophy of science has moved closer to science, metaphysics seems to have drifted further away. His own view is that although metaphysics is deeply important to, and ‘infused within’, science, many current debates within this area of philosophy are sterile or even empty. In response, metaphysics needs to become more responsive to and connected with current science, just as its importance for science must be emphasized, particularly when it comes to explanation and understanding. As he notes, his provocative but thoughtful essay is a largely negative critique, but clarifying what is wrong with the current situation is necessary if we are to move forward. Thus, consider modality, for example—the study of possibility, necessity and the like. As Callender points out, metaphysicians often rely on intuitions in developing views of modality, yet these intuitions
are historically conditioned and may be unreliable, or even inconsistent. Instead, and more positively, he urges that our understanding of what is possible or necessary should be more closely tied to what science tells us, and that we should take such modalities only as seriously as the theories that generate them. This is a contentious line to take, but if metaphysics is not to become entirely intuition driven, or tied to scientific theories that were abandoned hundreds of years ago, it is a line that at least needs to be considered. Callender sums his position up in a slogan that appropriately reflects this new relationship: metaphysics is best when informed by good science, and science is best when informed by good metaphysics.

As we also noted above, philosophy of science, not unnaturally, tends to focus on the practices of science – a tendency that Alexander Bird calls ‘particularist’ – whereas epistemology offers a much more general account of knowledge and justification. Bird illustrates some of the difficulties in resolving these tensions by looking at the cases of Bayesian confirmation theory and inference to the best explanation (IBE). The former gives us a rule (Bayes’ Theorem) that tells us how we should update the probability of a hypothesis given the evidence, where such a probability is traditionally understood in terms of subjective degrees of belief. Because the formalism of Bayesian confirmation is that of probability theory, the rationality of updating one’s beliefs in this way is established via ‘internal’ factors, rather than, say, empirical investigation. IBE, on the other hand, is more restricted in scope, but licenses the truth of a hypothesis on the grounds of its providing a better explanation than its competitors. Its reliability depends on ‘external’ considerations, most particularly the existence of a correlation between truth and good explanations. Thus, Bayesianism and IBE relate to different tendencies within epistemology, and Bird outlines the problems with these, before emphasizing how a ‘naturalised’ approach allows a role for science to play in epistemology by setting our cognitive capacities in the appropriate evolutionary context. He concludes by pointing to recent work on ‘knowledge first’ epistemology, where there is significant potential for productive interaction with the philosophy of science.

Turning from these broader issues and concerns to the more narrowly focused topics found in Part II, ‘Current Research and Issues’, here, too, new developments and directions of research can be discerned. Indeed, by the time the claim that the philosophy of science was degenerate had been made, the realist consensus indicated by Suppe had been shattered by the publication in 1980 of van Fraassen’s classic work, The Scientific Image (van Fraassen 1980), which presented a plausible form of anti-realism based on empiricist principles. Taking the aim of science to be empirical adequacy, rather than truth, van Fraassen has subsequently developed an entire ‘empiricist stance’, that extends from the foundations of quantum physics to the nature of scientific representation.
Stathis Psillos is notable for presenting a robust realist response to this anti-realist tendency. In his chapter for this volume he both sets the debate between realists and anti-realists in the context of its rich historical legacy and examines the twists and turns of the debate itself. Thus, following the publication of van Fraassen’s book, there has been an increased focus on finding a principled way of distinguishing those parts or aspects of the world that we can know from those that we cannot. Psillos takes the drawing of such a distinction in a principled way to unite the constructive empiricist, such as van Fraassen, with various realists, such as the (epistemic) structural realist and the ‘semi-realist’, but he argues that there is, in fact, no good reason to draw such a distinction in a principled way, and hence no need to abandon standard scientific realism. In addition, there has been a further shift towards the incorporation of neo-Aristotelian metaphysics into these variants of realism, and here again we encounter some of the issues discussed by Callender. According to Psillos, such moves can create a tension with the desire to draw the above distinction, and in any case, these shifts to neo-Aristotelianism should be resisted in favour of a less extravagant metaphysical landscape that draws its inspiration from the empiricist philosopher David Hume.  

The notion of explanation features prominently in the realism debate. Realists think science explains things about the world – facts, phenomena, the data obtained through experiment – and that the explanatory power of theories is an indicator of their truth, while anti-realists regard such power as purely pragmatic. But whatever stance is taken, an understanding of scientific explanation is crucial. Henk de Regt presents an overview of different accounts of explanation, covering both those that emphasize causal relationships (and that appear in a number of the essays included here) and those that are relevant for the human sciences. Importantly, he explores the adoption of a pluralist stance with regard to explanation and notes that, on the one hand, from a pragmatic perspective, explanation can have different forms in different contexts, and, on the other, if understanding is taken as a universal aim of science, this can be achieved via different modes of explanation that vary according to context. Furthermore, de Regt insists, these two approaches are not necessarily incompatible. As he says, science itself is varied, and so we should respect the diversity of models of explanation. He concludes by tying his discussion in with many of the themes covered elsewhere in the volume and noting some directions for future research, particularly with regard to mechanistic models.

As de Regt notes, causal accounts of explanation are currently very much in vogue within the philosophy of science. But when attempts are made to spell out the notion of causality, we once more run up against the kinds of concerns indicated by Callender. It would be so easy if philosophers of science could take such a notion down off the shelves of metaphysicians, but the
latter’s debates seem driven by unscientific ‘intuition pumps’ and series of examples and counter-examples that appear to have little, if anything, to do with current science. Ned Hall acknowledges this in his paper on causation, but he takes a more conciliatory line, urging us to take the metaphysicians’ intuitions as clues that indicate where more fruitful concepts of causality might be found. Furthermore, he notes that although the metaphysicians’ obsession with singular causation may appear to be less interesting in a scientific context, these discussions may act as ‘seeds’ around which more interesting kinds of questions can crystallize. In particular, they may help us get a grip on the kinds of counterfactual claims that underpin the understanding of causal structure as a structure of dependencies. This is where an account of laws comes into the picture, and Hall delineates the two current rival accounts and their variants. It is in explicating these issues in the scientific context that we can see the potential for productive and fruitful interaction between the philosophy of science and metaphysics.

Moving from metaphysical to epistemological matters, one central question – dealt with by James Hawthorne in his essay on confirmation theory – is how can we come to know whether the claims of science are true? A theory of confirmation takes the relevant evidence and, on that basis, effectively tells us the extent to which a given hypothesis is confirmed. The most influential contemporary theory of confirmation is Bayesianism, which uses probability theory – in the form of Bayes’ Theorem – to express how the likelihood of the evidence given the theory (that is, how likely the evidence is) contributes to the extent to which that theory is confirmed by that evidence. Although this area can quickly become very technical, Hawthorne presents all the relevant details in a clear and succinct fashion. He also considers the issue of what these probabilities represent, noting that the traditional view (discussed by Bird), that they express subjective degrees of belief, runs into considerable difficulties. His own view is that we should stop trying to give an account of what they are and think about what they do. Here the answer is clear: they are ‘truth-indicating indices’, and conceiving of them this way suggests a pragmatic strategy of continually testing hypotheses and taking whichever has an index closest to one as our best current candidate for being true.

Scientific evidence, on the above account, enters only with regard to the expression for the likelihoods that feature in Bayes’ Theorem. Hence Malcolm Forster, in his chapter on evidence, refers to it as the ‘likelihood theory of evidence’. However, in his contribution he urges the adoption of a broader conception, drawing on work in epistemology, and presses the point that the goal of our accounts of evidence should encompass not just the truth of theories, but also their predictive accuracy. Even if this is only a transitional step towards truth, it is a hugely important one; in some cases, optimizing truth need not involve optimizing predictive accuracy, and vice versa. Attempts to
understand science in terms of a single goal may thus be doomed to failure, and Forster argues for a more nuanced understanding of the relationship between such different goals.

Furthermore, whether scientific theories should be thought of as being truth-apt to begin with is itself a problematic issue. One of the most significant trends over the past 40 years or so (flagged by Suppe in his 1977 volume) has been an emphasis on the nature and role of models in science. In his contribution on models and representation, Gabriele Contessa outlines these developments, focusing on the way in which such models represent. In this case, it is the philosophy of art that has been turned to for both the relevant frameworks and the supposedly pertinent examples, with work in this area spanning the realism–anti-realism divide (see, for example, van Fraassen 2008). Contessa notes how such representational models allow us to engage in ‘surrogative reasoning’ about the systems of interest and draws an important distinction between the issues of what makes something a scientific representation and what makes it a ‘faithful’ representation. This allows him to impose a certain order on recent debates between alternative accounts and, in particular, he argues that although representation comes quite cheaply, faithfulness is a more costly commodity. He also considers the important question, whether any difference can be found between scientific and non-scientific representations, and answers that there appears to be none, although the former will obviously place greater emphasis on the quality of being faithful.

Realism as a stance within the philosophy of science is also influenced by developments in the foundations of the sciences – as described by the philosophies of those particular sciences – and it sometimes feeds into these philosophies as well. Van Fraassen’s anti-realism, touched on above, was motivated in large part by concerns over the impossibility of maintaining appropriate forms of realism with regard to quantum physics and the foundations of spacetime. Philosophical reflection on these topics has flowered over the past 30 years or so, and the results that drove van Fraassen to anti-realism have now been reinterpreted in ways that are more amenable to at least some forms of realism. Presenting these results in an accessible way is a difficult task, but Nick Huggett has managed to do so, while also presenting an overview of the foundations of statistical mechanics, a topic that has recently become the focus of considerable interest. Thus, he covers recent discussions of the statistical grounding of the Second Law of Thermodynamics, which he takes to provide an arrow of time, but one that is minimal and non-fundamental. Huggett also examines the metaphysical implications of quantum statistics for the identity of indiscernibles and explores the differences between the geometrical and dynamical approaches to General Relativity, concluding with a note on its cosmological implications.

However, the philosophy of science has long been held to be too ‘physico-centric’, a feature that can be explained on historical grounds, but which has
Introduction

been seen as closing off the opportunity to garner both useful insights and alternative stances from other disciplines. The philosophy of biology has only emerged from this shadow in the past 40 years or so, but has rapidly established itself as very much a practice-oriented field, where there is fruitful interaction between biologists and philosophers. As Ingo Brigandt notes, the issues that have arisen have been both epistemological and metaphysical, and in addition to ongoing discussions pertaining to evolution, molecular and experimental biology have also recently moved into the spotlight. As in the case of the philosophy of chemistry, to be touched on below, issues of reduction and explanation arise, in association with biological case studies, that have significant relevance for the philosophy of science in general. Thus, whereas in the latter, different accounts of explanation may be seen as rivals, within the philosophy of biology they are typically seen as complementary, a view that chimes with de Regt's suggestion above. And here we find models of explanatory reduction that are better suited to the piecemeal nature of practice in biology itself.

Similar issues arise in the philosophy of chemistry, which is another area that has grown in interest and where considerable work has been undertaken in articulating what it is that is distinctive about this field. Here Robin Hendry focuses on conceptual issues having to do with substances, bonds, structure and, again, reduction. A common thread running through his discussion concerns the importance of bringing scientific practices to bear on philosophical issues and debates. Thus, it is notable that philosophical considerations of ‘water’ as a term denoting a natural kind, for example, typically proceed with little input from chemistry. There are also foundational issues particular to this field, such as those having to do with the appropriate characterization of chemical bonds, which have been comparatively little discussed. Reduction, on the other hand, has long been a central topic, both in the philosophy of chemistry and the philosophy of science in general. Here Hendry covers a number of important distinctions, most notably that between the reduction of one theory to another and ontological reduction, that is, from one set of entities to another. In the latter case, he notes, the possibility of emergent properties and sui generis chemical laws raises interesting issues and, of course, the potential for blocking moves to establish any kind of straightforward reduction between chemistry and physics.

The issue of how to characterize scientific reduction in general is covered in detail by Sven Walter and Markus Eronen. In their overview of attempts to develop an appropriate account, they draw on examples from the philosophy of mind and from neuroscience. They note, in particular, the shift from the supposed derivability of a given theory from a more fundamental one, to an appreciation of the practice of science where scientists mount explanations in terms of empirically discoverable mechanisms. These explanations are typically ‘multi-level’ and present both a ‘downward-looking’ and an ‘upward-looking’
aspect that has been taken to support both anti-reductionistic (and pluralistic) and reductionistic attitudes. Here, metaphysics can again find useful employment in providing an appropriate underpinning for the crucial notion of ‘level’, but in an area such as neuroscience and in the context of an emphasis on mechanism, that notion must be understood as local and case specific. Here, as Walter and Eronen indicate, there is considerable scope for future research.

Mechanistic explanations also feature prominently in Carl Craver and David Kaplan’s discussion of neuroscience, in which they defend the legitimacy of such explanations from top to bottom, as it were. Their distinction between ‘good’ and ‘bad’ explanations in terms of the revelation of causes meshes with neuroscience’s emphasis on control, rather than expectation or prediction of empirical results. This distinction then allows them to sort the various kinds of models one finds in this field, and it also underpins their argument that merely descriptive models have little, if any, explanatory value. In neuroscience, at least, explanation is a matter of situating neurobiological phenomena within the causal structure of the world, and hence we return to issues concerning causality covered by Hall.

Finally, although mathematics is clearly an area in which such moves would be inappropriate, nevertheless issues of explanation arise here, too. As Christopher Pincock records, for many years now there has been a distinct separation between the philosophy of mathematics and the philosophy of science. In some respects, this parallels the separation, decried by Callender, between metaphysics and the philosophy of science, with the philosophy of mathematics focused almost exclusively on matters of ontology. However, the recent attention paid to mathematical practice has engendered a form of rapprochement, as issues to do with epistemology and explanation, for example, come to the fore. Even more interestingly, perhaps, there has been close examination of the role that mathematics plays in science. This has generated another realism–anti-realism debate, now centred on the issue of whether the apparent indispensability of mathematics in science licenses a realist attitude towards the relevant mathematical entities. Pincock covers both sides of this debate, taking us through the latest developments, before concluding with a consideration of the role of mathematics in modelling and idealization. As he says, these issues place the role of mathematics firmly on the philosophy of science agenda.

We hope we have indicated here the various threads that run through the contributions to this volume, from particular concerns with causal explanations, reduction, representation and so on, to more general and, perhaps, more fundamental issues regarding the relationship between the philosophy of science and metaphysics, epistemology and also the history of science. Our aim in selecting both the topics of the essays and their authors was both to provide overviews of these ongoing discussions and also to indicate the new
Introduction

directions and research trajectories that have emerged in the past few years. We shall return to these in our essay ‘Travelling in New Directions’, but our further hope is that the reader – whether an undergraduate student taking her first course in the philosophy of science or the hardened veteran of these debates – will find something useful, interesting and thought-provoking in these essays.

Notes

1 The ‘Chronology’ charts some of the history here.
2 For a light-hearted account of the first conference of this new movement, see http://bit.ly/azkycm
3 Hume’s ideas also feature prominently in Hall’s essay on causation.

References

Part I

Philosophy of Science in Context
This page intentionally left blank
2 Philosophy of Science and Epistemology

Alexander Bird

1. Philosophy of Science between Generalism and Particularism

The relationship between philosophy of science and epistemology as practised at the heart of general philosophy has been variable. Philosophy of science is caught between potentially opposing forces. On the one hand, philosophy of science needs to be true to the (at least apparently) distinctive and even arcane practices of actual scientists. This I call the particularist tendency, because it tends to emphasize the particular, special nature of science (and maybe even of the individual sciences). On the other hand, philosophy of science needs to relate its account of scientific belief to the entirely general account of knowledge and justification provided by epistemology. This I call the generalist tendency, because it seeks to place the philosophy of science within a general epistemological framework.

Modern philosophy of science emerged in the middle of the nineteenth century. The most significant work of philosophy of science since Bacon’s Novum Organon, William Whewell’s The Philosophy of the Inductive Sciences, Founded Upon Their History (1840), exhibits the particularist tendency. As the title of his book indicates, Whewell’s philosophy of science is built on his earlier work, The History of the Inductive Sciences (1837). While he does construct his account within a general framework that bears a superficial resemblance to Kant’s, Whewell’s description of science and its processes of reasoning are clearly specific to science and are motivated in part by reflection on the details of particular episodes in the history of science. For example, he uses Kepler’s discovery of the elliptical orbits of the planets as an example of discoverer’s induction: the unfolding of fundamental ideas and conceptions permits the ‘colligation’ of the observed facts concerning the orbit of Mars to the conception of those facts as satisfying an elliptical orbit, which is then generalized for all planets. Whewell insisted that a philosophy of science must be inferable from its history. As a consequence, Whewell’s account is detailed and permits multiple routes to discovery and knowledge. Discoverer’s induction is itself a multistage process, and the colligation component may be achieved by
a number of different inferences. Likewise, the process of confirmation can involve three distinct components: prediction, consilience and coherence. Whewell addresses questions that are still much discussed in philosophy of science, such as whether novel predictions have greater value than the accommodation of old data – questions which have been scarcely addressed in general epistemology, either then or since.

While Whewell came to philosophy of science as a polymath scientist (he invented the very term ‘scientist’), John Stuart Mill came from the other direction, that of a general philosopher. Mill’s *A System of Logic* (1843) demonstrated the generalist tendency, seeking a unified account of all reasoning within an avowedly empiricist framework. For Mill, our inferential practices in science are essentially either enumerative or eliminative in nature (the latter being really deductive). Mill is generally sceptical regarding the methods Whewell describes, such as those that allow for inferences to the unobservable. All satisfactory science depends on enumerative induction, which is fallible, allied with elimination, which tends to reduce, if not remove, the fallibility of the induced hypotheses. For Mill, the inductions of science are extensions of the spontaneous inductions we naturally make. Science aims to improve on these spontaneous inductions, making inductions of a less fallible variety. But the pattern of inference is essentially the same. That being the case, Mill needed only a few, and simple, examples to illustrate his case. Whewell criticized Mill on precisely this point, that the paucity of his case studies suggests that Mill’s philosophy of science, unlike his own, could not be inferred from the study of its history.

Mill’s generalism was dominant in the first part of the twentieth century. The logical positivists’ conception of the relationship between scientific and everyday knowledge is best summarized in Einstein’s famous comment that ‘the whole of science is nothing more than a refinement of everyday thinking.’ To refine something is to remove impurities; the basic substance is unchanged. The logical positivists took the view that there is no fundamental difference between epistemology as applied to science and epistemology as applied to ordinary instances of knowledge and belief. In his *General Theory of Knowledge*, Moritz Schlick asserts, ‘knowing in science and knowing in ordinary life are essentially the same’ (1985, p. 9). Similar statements may be found in the work of other positivists, and beyond (from James Dewey and George Santayana, for example). For Schlick, science and everyday knowing differ only in their subject matter. It is natural that empiricists, such as Mill and the positivists, should take such a view. After all, empiricism emphasizes the foundational role of individual experience. Furthermore, it is sceptical (to some degree or other) of inferences that take us much beyond that experience. Consequently, they take the foundation and even the extent of all knowledge, scientific and everyday, to be fixed by something – individual sensory experience – that is
clearly a central concern of general epistemology, and not something abstruse and specific to science.

A corollary of the ‘science is refined everyday thinking’ dictum is that everyday thinking is scientific thinking with impurities. Because science and everyday thinking are essentially the same, except that science is purer, an understanding of their common epistemological foundation is best gained by looking at science. Thus, for the positivists, epistemology just is philosophy of science.

As philosophy entered a post-positivist phase in the second half of the twentieth century, the particularist tendency returned to the ascendancy, with an emphasis on the significance of the history of science that resembles Whewell’s work, rather than Mill’s. Thomas Kuhn gives an account of the development of science that is not deduced from some general epistemology, but is inferred from the historical facts. Although not an epistemology of science, Kuhn’s picture is epistemologically significant. First, he rejects a simplistic, empiricist conception of observation. Perceptual experience cannot be the foundation of science, because it itself is theory-laden. And observation, the process of generating scientific data, is not simply a matter of perception: ‘The operations and measurements that a scientist undertakes in the laboratory are not “the given” of experience but rather “the collected with difficulty”.’ (Kuhn 1970b, p. 126). Secondly, whereas Carnap (1950) took inductive logic to provide a generally applicable, rule-based approach to scientific reasoning, Kuhn argued that scientific cognition is driven by a scientist’s learned sense of similarity between the scientific problem she faces and some concrete exemplar – an earlier, exemplary piece of scientific problem solving. Thirdly, in the positivists’ picture, theory-neutral perceptual experience is supposed to provide a basis for the meaning of our theoretical terminology; but in the light of the first point, that basis is not theory-neutral at all. The languages of different scientific communities, working within different paradigms (i.e. employing different sets of exemplars), will not be immediately and simply inter-translatable – they will be incommensurable. Together, these points indicate an epistemological relativism. We do not have an unproblematic, agreed basis for settling scientific disputes. The language disputants use may be incommensurable, the data may be laden with theories that are themselves subjects of dispute, and our assessments of theories may be determined by differing exemplars. Such problems do not arise in an obvious way in the epistemology of everyday knowledge.

The debates between Popper (1959, 1970), Kuhn (1970a) and Lakatos (1970) illustrate the changing nature of the perceived relationship between the philosophy of science and the history of science. Although Sir Karl Popper was not a logical positivist, he shared many of the beliefs of logical empiricism. A key problem for Popper was a central issue in general epistemology – Hume’s problem of induction. It was also a central issue for others, such as
Carnap (1950) and Reichenbach (1971), but whereas they had some hope of providing a probabilistic solution to the problem, Popper took it to be insoluble. This fact, of being an inductive sceptic, combined with the view that science is rational, is the defining feature of Popper’s philosophy. The latter can be seen as the attempt to reconcile inductive scepticism with scientific rationalism. While Popper’s work is scientifically well-informed, and does refer to episodes in the history of science, it is clear that these have only a minimal evidential role. Popper’s falsificationism – his rationally acceptable alternative to inductivism – is primarily normative in character. But since Popper holds that science largely meets the norms of rationality, his combined view is answerable to history, for the majority of scientists should behave in accordance with Popper’s falsificationist method. Kuhn’s complaint is that they do not. According to Popper, scientists ought to reject a theory as soon as they discover a piece of evidence inconsistent with it. But Kuhn’s historically well-attested account of normal science shows that they do not do this: a theory can accumulate anomalies, including data flatly inconsistent with the theory, without being rejected. Kuhn emphasizes that a theory only begins to be questioned when anomalies of a particularly serious sort accumulate that resist the attempts of the best scientists to resolve them. One aim of Imre Lakatos’s ‘methodology of scientific research programmes’ is to reconcile a falsificationist approach with the historical record. Lakatos (incorrectly, in my view) held that Kuhn’s account of scientific development would, if correct, imply that science is irrational. Thus, while Popper’s view was founded in epistemology, with a relative disregard for historical detail, and Kuhn’s view was historically well informed but disengaged from general epistemology (there is just one, rather offhand, reference to the problem of induction in The Structure of Scientific Revolutions), Lakatos seeks to combine history with a normative approach derived from Popper’s general epistemology. Famously, Lakatos (1971, p. 91) adapted Kant’s words: ‘Philosophy of science without history of science is empty; history of science without philosophy of science is blind.’

I suggested above that a thoroughgoing empiricist philosophy of science is likely to be close to a general empiricist epistemology, because of (a) the central place given to perceptual experience, and (b) the disinclination of empiricists to go much beyond the evidence of our senses. Kuhn and others, such as N. R. Hanson (1958), emphasize the problematic nature of (a) in a manner that exacerbates the anti-realist tendency of logical empiricism. At the same time, a return to scientific realism in the philosophy of science (e.g. Grover Maxwell 1962) questioned (b). If science aims to get to know not just about perceptible items, but also about the unobservable, then epistemological questions may need to be asked about scientific cognition that do not apply to a central, everyday instance of cognition. Pursuing this line
of thought, it is open to the scientific realist to ask whether, since science has a subject matter that differs from the objects of everyday knowledge, the inference patterns of science are also distinctive.

2. Scientific Reasoning

Even among those who do think that science may employ forms of reasoning that are not found in everyday cognition, the opposing generalist and particularist tendencies are apparent. A long-standing disposition among many philosophers of science (and even more scientists) is to hold that the successful development of science has the scientific method to thank. The scientific method, which, one might hold, emerged during the period of the scientific revolution, is a single approach to scientific investigation and inference that underlies all good science in all domains. Belief in the existence of a single scientific method exemplifies the generalist tendency. Opposed to such a view, a particularist might deny that there is a single scientific method; instead there are many methods, reflecting the diversity of scientific theories and subjects.

If there is a scientific method, it must be largely tacit, because there is little agreement on what, exactly, the scientific method is. Since it is tacit, it is the job of the philosopher of science to reconstruct it, showing, on the one hand, how it is in fact used in successful science and, on the other hand, that its success is explained by some general epistemological justification. Lakatos’s methodology of scientific research programmes provides one account of the scientific method in action, but not the only one; Carnap’s inductive logic may be considered another.

2.1 Bayesian Epistemology

Of contemporary accounts of a single scientific method, the clear front runner is Bayesianism. Bayes’s theorem tells us how the conditional probability, \( P(h \mid e) \) of a hypothesis \( h \), given the evidence \( e \), is related to the probability of the hypothesis independently of the evidence, \( P(h) \), the probability of the evidence, \( P(e) \), and the probability of the evidence given the hypothesis, \( P(e \mid h) \):

\[
P(h \mid e) = \frac{P(e \mid h)P(h)}{P(e)}
\]  

(B)

Traditional Bayesianism interprets these probabilities as credences – subjective degrees of belief. On learning that \( e \) is in fact the case, the Bayesian says that one is rationally obliged to update one’s credence in \( h \) so that it is equal to \( P(h \mid e) \) as given in (B). This is known as conditionalization. Bayesians argue that
Bayesian conditionalization can resolve a whole host of problems and issues in the epistemology of science, such as the ravens paradox, the problem of induction, and the Duhem problem (Howson and Urbach 2006).

What is unusual about Bayesianism is that it is an epistemological tool or approach that was developed principally within the philosophy of science, and within science itself, and only thereafter transformed into a general epistemology, applicable not only to scientific but also to more general epistemological problems (Bovens and Hartmann 2003). There are a number of reasons for this. First, orthodox, subjectivist Bayesianism concerns fairly minimal rational constraints on credences, subjective degrees of belief. It does not tell us what those degrees of belief ought to be, only that they ought to be given previous degrees of belief. Consequently, Bayesianism does not provide any account of the relationship between belief and the world. But such relationships are at the centre of traditional epistemology; witness the centrality of concerns about scepticism, for example. Thus Bayesian epistemology has no place for the concept of knowledge, which expresses precisely such a relation. Arguably, it cannot account for the concept of justification. This point is more disputable, because unlike the concept of knowledge, the concept of justification provides no direct implication of a relationship between beliefs and the world. One might argue that a subject's degrees of belief are justified precisely when updated in accordance with Bayesian conditionalization. Also, one might think that the notion of justification cannot be entirely divorced from the notion of truth: justification ought normally to be truth-tropic (in the sense that better justified beliefs are more likely to be true), whereas Bayesian conditionalization is not, except in a very weak sense. Updating in accordance with Bayesian conditionalization puts a constraint on possible credences that is much like the requirement of logical consistency in the case of full-on belief. Such a constraint helps avoid some falsehoods, but does not go beyond that in pointing towards the truth. If one had chosen, before the process of conditionalization began, a sufficiently way-out set of credences, then one can obey conditionalization impeccably over a large data set, and still end up with credences that are far from reflecting the truth. Bayesians may point to the phenomenon of the ‘washing out of priors’. Subjects with differing priors who conditionalize on the same set of evidence will tend to converge in their credences. But the actual credences will only actually converge, that is, reach the same value (within some small tolerance), in the long run, which may be a very long run indeed, if the initial priors differ enough. In the short and medium run, the credences of the subjects may still differ significantly. And the question of justification concerns one’s current credences: is this subjective degree of belief justified?

Note that this case is not analogous to the possibility of justified false belief in traditional epistemology. A standard case is that of the subject who, looking
at a convincing barn façade, as one might have found in a Potemkin village, believes there is a real barn ahead. In this case, the subject’s failure to believe correctly is due to factors outside the subject’s control. On the face of it, the subject did exactly what she ought to do: form a belief on the basis of the way things appear to be, there being no reason to suppose that things are unusual. That seems to be a good, if informal, explanation of why she is justified in believing that there is a barn in front of her. Consider, on the other hand, the subject who starts with way-out initial priors and consequently now has credences that are far from the truth (i.e. he has high degrees of belief in falsehoods and low degrees of belief in truths). This subject is in that position not through any bad luck or fault with the evidence, but only because he started out with bizarre initial priors. It is debatable whether the well-behaved conditionalizer with way-out priors is justified in his credences that radically misalign with the truth. The extent to which one is willing to ascribe justification in such cases will depend on how strong a link between belief and world one thinks is signalled by the concept of justification. As we shall see, epistemological internalists will be more sympathetic to the Bayesian case than epistemological externalists: internalists insist on the epistemological primacy of internal relations among beliefs, whereas externalists prefer to focus on the belief–world relation.

2.2 Inference to the Best Explanation

In many respects inference to the best explanation (IBE) provides an important contrast to Bayesianism. Unlike Bayesianism, IBE is concerned with the objective truth of hypotheses rather than their subjective probability. Thus, IBE connects more directly to issues in traditional epistemology than does Bayesianism. On the other hand, supporters of IBE as a model for scientific inference typically do not regard every inference in science as conforming to the model. For example, it is not clear how one might construe many statistical inferences in science as IBE; likewise, inferences based on specific test procedures. One might attempt to force some of these into the IBE mould. For example, a litmus test may license the inference that a particular substance is acidic. This might be construed as an inference that the best explanation of the test result is that the substance is acidic. But many such Procrustean moves seem implausible, psychologically at least.

Instead, the IBE supporter should hold that only a particular subset of the inferences of science, albeit a particularly important one, exemplifies IBE. According to IBE, the truth of a hypothesis is licensed when that hypothesis provides a better potential explanation of the evidence than all its rivals (Harman 1965; Lipton 2004). That is, we select as being the actual explanation, an explanation from a pool of potential explanations that is (a) clearly better
than all its rivals at explaining the evidence, and (b) is in itself a good enough explanation. IBE faces a number of challenges and demands for clarification. What counts as a ‘good’ explanation? Isn’t goodness, whatever it is, too subjective to be a measure of truth? How can we be sure that we have considered the actual explanation among the pool of potential explanations?

The most obvious and important of the objections to IBE is that which Lipton calls ‘Voltaire’s objection’. Why should explanatory goodness be correlated with the truth? For IBE to be an inference procedure that leads to the truth, the actual world must be the best of all possible worlds, explanation-wise. Why should we think that is the case? On the face of it, it doesn’t seem to be a necessary truth that the actual world is explanatorily better (e.g. simpler, more unified, more elegant, ‘lovelier’) than all other possible worlds. Consequently, if IBE is a reasonably reliable form of inference, that fact is a contingent fact. Furthermore, it does not look to be knowable a priori.

In the light of this, one might ask, is it really the case that IBE can lead to knowledge, or even to justified belief? In using IBE, it appears that scientists are using a method that they appear to have no reason to think will lead to the truth. The case seems analogous to sceptical questions surrounding perception. On the (admittedly doubtful) assumption that we make inferences concerning our immediate physical surroundings from our sensory experiences, the sceptic points out that it is neither necessary nor a priori that there should be a match between the two. Hence, the inference is made upon a basis of a supposed match that the subject has no reason to believe is the case. How, then, can the inference lead to knowledge or even justification? Note again the contrast with Bayesianism. Bayes’s theorem itself is clearly a priori and may be inferred with ease from the intuitively compelling axioms of probability. Conditionalization is more contentious. Even so, Bayesians can quite plausibly claim that conditionalization is a priori, and they bring forward arguments of an a priori kind to defend conditionalization. For example, Bayesians appeal to diachronic Dutch book arguments (e.g. Teller 1976). Such arguments aim to show that a subject who updates credences in a way that does not obey Bayesian conditionalization is susceptible to a Dutch book: accepting a set of bets that can be shown to lead to a loss, whatever the outcome. This shows, it is suggested, that failure to obey conditionalization is a kind of violation of rationality.

3. Internalist and Externalist Epistemology

The contrast between Bayesian epistemology and IBE exemplifies a deep divide within epistemology, between epistemological internalists and epistemological externalists. There is more than one way of characterizing the
difference, and not all ways are equivalent, but most ways will have something in common with the following. The internalist requires that if a subject uses some method, such as an inference pattern, to come to have knowledge or a justified belief, then the subject must have some reason to believe that the method in question is likely to lead to the truth. The externalist will deny this. Another characterization is to say that the internalist holds that the state of epistemic justification supervenes on the subject’s intrinsic states – such as their sensations, perceptual experiences, beliefs and credences (ignoring issues of content externalism). Denying this, the externalist will hold that in some cases the difference between a justified believer and an unjustified believer may be some further feature that is extrinsic to the subjects, such as some relation between the subject and the world.

We can see why there is a connection between Bayesianism and internalism on the one hand, and IBE and externalism on the other. The internalist thinks that justification is a matter of one’s internal states, in particular an internal ability to discern the truth-conduciveness or rationality of one’s methods. The Bayesian says that the truth of Bayes’s theorem and the rationality of Bayesian conditionalization are a priori, that is, discernible by internal reflection rather than empirical investigation. The externalist will say that in some cases the justification provided by a belief-forming method will depend on factors of which the subject may not be aware. For example, according to the reliabilist version of externalism (Goldman 1979), a justified belief is one that is brought about by a method that is reliable. A reliable method is one that produces, or tends to produce, true beliefs. A method can be reliable without anyone being aware of its reliability. Consequently, a subject can have a justified belief without knowing that the belief was produced by a method that is reliable, so long as it is in fact reliable. For example, one might think that perception can justify one’s beliefs or give knowledge without anyone having considered the reliability of their perceptual capacities. The point is most clear with regard to small children or animals. It would seem obvious that they have perceptual knowledge, although they are not sophisticated enough to consider, let alone evaluate, the reliability of their senses. What is relevant in deciding whether they do gain perceptual knowledge is the actual reliability of their senses, not their knowledge of that reliability. Returning to IBE, perhaps it does not matter that one cannot know by any kind of reflection that explanatory goodness and truth are correlated. It may be sufficient for one to gain knowledge from IBE that there is, in fact, such a (non-accidental) correlation, as a result of which the central process of IBE – namely that of inferring from the fact that a hypothesis is the best explanation of the evidence to the truth of the hypothesis – is a reliable form of inference.

Externalist epistemology as applied to science faces the following problems. Reliabilism, a leading form of externalism, seems most plausible when
(a) regarded as a theory about knowledge, and (b) applied to relatively basic (e.g. innate) forms of belief production; reliabilism looks good as a theory about perceptual knowledge. Correspondingly, reliabilism may seem rather less plausible when considered as a theory of justification, as applied to the sophisticated methods developed by science. Could someone be justified in believing in neutrinos on the basis of Pauli's arguments from the details of beta decay and the conservation of mass, without understanding why those arguments justify the belief? It would seem that, in science, some degree of reflective awareness of the justificatory power of evidence and argument is required in order for the subject to be justified in holding a belief (when caused by that evidence and argument). Thus, internalism would appear to be on stronger ground when we think about scientific ways of producing beliefs than when the topic is simple perceptual belief-forming methods. This may be one reason why the internalism–externalism debate has been more intense among general epistemologists than philosophers of science.

This evidence in favour of internalism should not be exaggerated, however. First, note that reliabilism is only one externalist theory. Externalism itself says that it is possible in some cases to know or to have justified belief without reflective awareness of the belief-forming method employed; it need not say that such awareness is never required. Secondly, it should be recognized that the individually accessible justification that scientists have in their beliefs is very limited. For example, a chemist justifies her belief in a favoured hypothesis by referring to the fact that electrodes in a certain solution produced a current of 5.2 mA. But were she to be asked what justifies her faith in that fact, she may be able to do no better than to tell us that this is the reading that the ammeter showed, and that the ammeter was made by a reputable manufacturer. In another case, a physicist may be able to justify belief in Einstein's general theory of relativity by referring to the details of the precession of the perihelion of Mercury. But he may be able to tell us nothing about how the astronomers calculated that precession, nor about the observations they employed. In these cases, the justification of a belief seems to depend ultimately on the reliability of some instrument or the testimony of other scientists. (The social nature of scientific evidence and the importance of trust is emphasized in Hardwig 1991.) By the internalist's standards, this is insufficient to provide justification. One possible response would be to abandon internalism as a thesis about individual justification, but retain it as a thesis about group beliefs. The community as a whole is justified in the relevant beliefs, because as a whole it can provide an explanation of the functioning of the ammeter or the basis for the belief about the perihelion of Mercury.

Thirdly, internalism does not say that the subject should be aware of something relevant to the justification of their beliefs, but everything. The justification of a belief should depend solely on what is available to the subject. This is
importance because we may wish to make a distinction between the reasons
given that are specific to some particular case and the general form of the
justification. Let us imagine (as seems plausible) that Pauli’s reasoning
concerning beta decay is an instance of IBE. Certainly Pauli and anyone else
who uses this reasoning to gain a justified belief in the existence of neutrinos
ought not simply to be able to point to the relevant evidential facts (the
observed loss of mass-energy during beta decay, the laws of conservation of
mass). But they should also be able to articulate the idea that the existence of
the neutrino provides a simple but powerful explanation of the observed loss
of mass; and they ought also to be able to articulate, to some extent, why this
is a better explanation than some obvious alternatives, such as the proposal
that the law of conservation of mass is not strictly true. However, we would
not expect the scientist to be able to articulate, on a more general level, why
the fact that the neutrino hypothesis possesses the property of being the best
explanation we’ve come up with gives us reason to believe that hypothesis.
So science differs from perception insofar as it does require for justified belief
that the scientist be able to articulate reasons for her belief and to be able to
provide some explanation of how they support the belief. But that fact does
not fully vindicate internalism, for we do not expect the scientist to be able to
articulate why the general pattern exemplified by this case is a good one to
use; why this pattern is a pattern that makes it the case that the specific rea-
sons proffered by the scientist really are reasons for belief.

This line of thought then raises the following questions for the externalist
thinking about scientific reasoning. In order to have justified beliefs, the
subject is required to be able to justify herself – to give her reasons – to a fair
degree, but not beyond. Where does the cut-off point come? Why does it
come where it does? And why is sophisticated thinking different in this
respect from more direct, mundane or innate ways of forming beliefs which
do not seem to require self-justification? While these questions are signifi-
cant for the philosophy of science in particular, they are ones for general
epistemology to answer.

4. Naturalized Epistemology and the Philosophy of Science

Naturalized epistemology is one manifestation of epistemological externalism
(Quine 1969; Kornblith 2002). Cartesian epistemology takes it to be the task of
the epistemologist to show how we can come to know things from the vantage
point of the subject whose basic resources are limited to his or her reflective
capacities, plus the sensory impressions with which he or she is presented.
An optimistic Cartesian epistemologist, such as Descartes himself, might hope
to be able to deduce the veracity of his sensory impressions and thus get to
know about the external world. But such a project, whether in a foundationalist or coherentist guise, faces severe obstacles; it is susceptible to scepticism and threatens to collapse back to its solipsistic roots. The naturalistic epistemologist, on the other hand, does not start from the subjective experiences and reasoning capacities of the isolated subject, but sees the subject as an organism in an environment, an organism whose cognitive capacities have evolved in certain environmental conditions (which may include social conditions) and who uses those capacities to navigate and interact with that environment.

For the naturalized epistemologist, it is no surprise that our cognitive capacities have their limitations. They are evolved to cope with environments of a certain kind. Consequently, it cannot be regarded as a failure of those capacities if they would be unable to distinguish some common state of affairs from some hypothetical state of affairs that would obtain only in some reasonably distant possible world. Thus, for our perceptual systems to be operating properly, it is not required that they should be able to tell (in a normal circumstance) that the subject is not a brain in a vat or that she is not being deceived by an evil demon. If we further associate knowledge with what is produced by a properly functioning cognitive system in a propitious environment, then we have a rather more liberal account of when the subject knows than is provided by the Cartesian epistemologist.

Naturalized epistemology provides a role for science in epistemology. Under the Cartesian approach, it is rational (internal) reflection that assesses the ability of some method to produce knowledge – whether the reliability of the method is immune from rational doubt. The naturalized epistemologist, on the other hand, regards the knowledge-generating capacity of a belief-forming process to be an a posteriori matter whose assessment requires scientific investigation. Thus, naturalized epistemology can adopt a methodological pluralism that exemplifies the particularist tendency referred to above. There is no single scientific method – and certainly not an a priori one. Rather, as science develops, it refines and adds to its methods. For example, early clinical trials involved no form of blinding or placebo control. Techniques such as double blinding, randomization and placebo controls were developed to improve methods of assessing therapies as it became clear that outcomes could be biased by the both patients’ and doctors’ knowledge of who is being treated and who not.

Naturalized epistemology gives a context for and support to externalism. It also thereby provides some degree of support for scientific realism. Externalist epistemology is less susceptible to sceptical worries than internalism. The externalist will regard it as potentially sufficient for being knowledge that a subject’s belief stands in an appropriate relation to the world, for example, by being produced by a reliable belief-forming method, whereas the internalist will require, in addition, that this connection itself be known to the subject.
The empiricist displays this kind of internalism and is reticent about making inferences from sense experiences as to the nature of an independent physical world. Consequently, the empiricist inclines either towards epistemological anti-realism (our theories do not tell us about the way the world is, they aim only to provide accurate predictions of future sense experiences (Duhem 1914, c.f. van Fraassen 1980)), or they incline towards metaphysical anti-realism (our theories do tell us about the world, but only because the ‘world’ is constituted by our sense experiences (Carnap 1961)). Externalism, bolstered by naturalized epistemology, says that the reliability of a microscope, MRI scanner, radio telescope and so forth can suffice to give us knowledge of independent entities we cannot experience directly.

On the other hand, naturalized epistemology does present a prima facie puzzle. How is it that the cognitive capacities of Homo sapiens, which evolved to give us knowledge of our immediate physical and social environment, are also able to give us knowledge, as the realist claims, of facts of a very different kind, including mathematically sophisticated laws of nature, the age and structure of the universe, and the constitution of atoms? One (part of) a possible answer is the idea that our cognitive capacities continue to evolve: our innate capacities are supplemented by those that develop thanks to advances in science and in the social organization of science. Let us consider inference to the best explanation. As noted, IBE requires externalism – it is not a priori that the best explanations we devise are likely to be true. How exactly does externalism help IBE? Let us (the epistemologists) assume that the world is, indeed, one in which the best explanations are likely to be true. Then a subject who argues in accordance with IBE will be using a reliable method and will thereby acquire justified beliefs. While that justification may require the subject to have reasons for thinking that her inferred hypothesis is a better explanation than its competitors, it does not require that she has reasons for thinking that IBE is a reliable method or that the world is one in which the best explanations are likely to be true. Do we, as epistemologists, have such reasons? An externalist epistemologist might regard herself within her rights to decline to answer this question. This minimal response says that the importance of the externalist perspective is its response to scepticism: it can show that justification and knowledge are possible. Whether they are actual depends on the way the world actually is, and such contingent questions are not within the realm of philosophy.

The naturalized epistemologist, on the other hand, might feel emboldened to go further, regarding it as legitimate to use empirical results in answering questions originating in epistemology. Showing that IBE is reliable is a matter of showing that the world is one in which the best explanations are often true. And this looks like a tall order. After all, that would seem to make the world rather special, and why should that be the case? A possible strategy is to argue as
follows. There can be different standards of explanatory goodness. What is needed is that the world is such that explanations that are the best by our standards are likely to be true. Furthermore, it might be the case that standards are discipline-relative. So in order for IBE as used in, say, palaeontology to be reliable, it needs to be the case that the world is one such that the explanations that are the best by the standards used by palaeontologists are likely to be true. Now imagine that the standards in question are not fixed for all time, but evolve in response to the developments of science. In particular, it might be that the explanatory standards exemplified by a particularly successful hypothesis are more likely to be adopted than those exemplified by an unsuccessful theory. If so, it would be the case that the explanatory standards employed by a field of science are ones that are correlated with the truth and were adopted for that reason. Thus, we would have an explanation of why IBE as employed in advanced sciences is reliable. So far, what has been presented is a hypothesis that stands in need of evidence; evidence that can only be gathered empirically. Arguably, that evidence is exactly what is supplied by Kuhn’s historically informed account of normal science, according to which the evaluative standards, including explanatory standards, employed by scientists working in a particular field are embodied in an especially successful theory (or practice) that stands as a model or exemplar for subsequent science in that field (Kuhn 1970b). Kuhn’s history can be supplemented by the work of psychologists and cognitive scientists, who show that scientific thinking is heavily informed by thinking in terms of analogy with understood scientific problems, models, exemplars, cases and so forth (Dunbar 1996). (For this approach to work, it has to be argued that there is significant continuity of explanatory standards across scientific revolutions – as Kuhn allows.) This example illustrates how the naturalized epistemologist may be able to use the results of empirical investigations (in this case from history and psychology), to argue for an epistemological proposition (in this case concerning the reliability of IBE in advanced sciences).

5. Knowledge and Evidence

To conclude, I shall address a neglected topic in the epistemology of science, but one which, in my view (Bird 2010), offers considerable scope for philosophy of science to benefit from general epistemology. Despite the fact that we talk about ‘the growth of scientific knowledge’, very little epistemology in the philosophy of science is concerned with knowledge. There are a number of related reasons for this. Bayesianism simply eschews the notion of knowledge altogether – its sole epistemological notion is credence. (Bayesians do use a notion of evidence, but the latter is typically treated as matter of having credence equal to unity.) Even non-Bayesian accounts of confirmation tend to
focus on incremental confirmation – which is when a piece of evidence adds to the likely truth of a hypothesis. This sheds little light on the question of when one has enough evidence to know that the hypothesis is true. Although motivated by scepticism, Popper’s fallibilism has proved a popular view, especially among scientists. According to this view, our best theories are merely tentative, always open to the possibility of refutation. This is difficult to reconcile with the claim that we know some scientific theories to be true. A moderately sceptical attitude is common, even among realist philosophers of science. While they reject outright scepticism, they are sensitive to arguments such as the pessimistic meta-induction (Poincaré 1943). The latter argues that past theories that have been replaced by current theories were very often successful, even though ultimately refuted. This shows that success is no good indicator of truth, and that we ought not hold that our current theories are different in this respect from their predecessors. A common realist response is to accept that our theories are, indeed, always strictly false, while maintaining that they are nonetheless gaining in verisimilitude over time – later theories are getting closer to the truth, even if they do not reach it (Psillos 1999). But being strictly false, there is no possibility of knowing these theories to be the truth.

Science is sometimes held up as humankind’s most remarkable epistemic achievement. If so, it would be odd if it failed to reach the standard set by a very common epistemic standard, knowledge. Furthermore, according to one important epistemological approach, it would be impossible for science to reach any significant level of achievement without there being knowledge in science. Timothy Williamson’s ‘knowledge first’ epistemology emphasizes the centrality of knowledge (Williamson 2000). For example, Williamson argues that all and only what we know is evidence. So the denial of knowledge in science would be tantamount to the denial of evidence in science. In the absence of evidence, little or nothing in science would ever be justified. Williamson also argues that many of the problems facing Bayesian conditionalization are ameliorated if what is conditionalized upon is knowledge. The equation between evidence and knowledge also puts pressure on empiricism, which assumes that all our evidence is perceptual. This equation does not limit evidence to any particular kind of proposition. It thus supports scientists’ practice of citing heavily theoretically laden propositions as evidence (such as the anomalous precession of the perihelion of Mercury as evidence for General Relativity). The focus on knowledge also allows for a natural account of progress in science as the accumulation of scientific knowledge (Bird 2007).

6. Conclusion

Williamson’s ‘knowledge first’ approach was developed with general epistemology in mind. The fact that it has many important implications for the
philosophy of science shows that the relationship between general epistemology and the epistemological philosophy of science is still important, at least to the latter. Particularists will argue that such general approaches, while significant, can only provide a framework and partial answers to a question in scientific epistemology. There are still epistemological questions that are specific to science and which require specific answers: for example, how is it that the explanatory power of a theory is indicative of its truth, or can science be rational if it is guided by Kuhnian paradigms? Answers to such questions might be provided, in part, by empirical naturalistic investigation, drawing on history of science and cognitive psychology. The particularist–generalist tension is a significant one for philosophy of science. On the one hand, philosophy of science must be true to science itself; it cannot deal only with an idealization that bears only a superficial resemblance to anything with which real scientists are engaged. On the other hand, the parts of science do not exist in epistemic isolation, nor does science as a whole. Even if the particularists are right, and that the details of the particular patterns of inference found in science are not derivable from a single epistemological rule, it remains a significant task for epistemology in science to show how these details cohere with a general approach to epistemology. It is notable that few of the recent developments in general epistemology – virtue epistemology, anti-luck epistemology, contextualism and knowledge first, for example – have had much impact in philosophy of science. The preceding section gave a brief indication of the potential significance of the last of these. An important task for epistemologists and philosophers of science is to generalize this approach and to see how other developments in epistemology relate to science, on the one hand to test the fruitfulness of such approaches in the epistemically crucial case of science, while on the other hand seeing whether insights in general epistemology can help us advance our understanding of epistemological questions arising from the study of scientific discovery and inference.

References


1. Introduction

Philosophy of science has a complicated relationship with metaphysics. Studying topics such as the nature of causation, laws of nature and space-time, it clearly engages in activities that merit classification as metaphysics. Yet the academic discipline itself was born in opposition to metaphysics. The positivists were united in a shared distrust of metaphysics. Their suspicion ran so deep as to motivate a search for a demarcation between science and non-science, and science and speculative metaphysics in particular. Even today, philosophy of science appears caught in what Einstein (1933) called the ‘eternal antithesis between the two inseparable components of our knowledge – the empirical and the rational’ (p. 271). It wants to employ metaphysical speculation, but impressed with the methods of the subject it studies, it fears overreaching. Philosophy of science thus tries to walk a fine line between scientifically grounded metaphysics and its more speculative cousins.

Here I will try to draft some of the contour of this boundary, along the way introducing the reader to some of the relevant issues. Doing so is critical today, for we are in the midst of a major collision between two very large forces in philosophy that has a significant bearing on metaphysics. Whereas metaphysics and science were once one and the same field, natural philosophy, today there is a worrisome divide between the two.

This separation is no doubt due to developments within both science and metaphysics. Physics, for instance, in part due to its distribution of incentives since World War II, is far less ‘philosophical’ than it used to be (Holton 1986). Nineteenth-century physicists debated the reality of the electric field, but today few physicists debate the updated counterparts of this question for gauge fields. The same goes for the measurement problem in quantum mechanics. Sometimes dubbed the ‘reality problem’, the issue is really about the proper ontology suited to quantum theory, and it’s hard to imagine a question of comparable importance in previous times being shunted aside as it often is today. The same could be said for problems in many other fields of
science, too. As a result, metaphysical insight is especially needed now. Yet, instead of offering to fill the breach, many metaphysicians have adopted an approach to the field that makes it more or less autonomous from science. Not only is this a shame, given the current context within science, but it is also a bad idea, for it occasionally results in debates in metaphysics becoming sterile or even empty.

In what follows, I concentrate on the philosophical side of this increasing gulf between science and metaphysics. After tracing the origin of this gap, in part, to the resurgent idea that metaphysicians have a wider domain of study than scientists – and arguing against this – I suggest a rough and ready ‘systematization’ criterion that makes the above divide apparent. The criterion simultaneously allows that metaphysics is deeply infused throughout science, while also counselling that metaphysical investigations ignore science at their peril.

2. The Current Clash and Its Background

There is a long tradition of worrying about overreaching by metaphysics. Kant famously attacked metaphysics as an assortment of empty sophistical tricks, a kind of perversion of the understanding. Later, seeing themselves as Kant’s heirs, Carnap, Reichenbach and others took the measure of metaphysics and saw it as strikingly different from science:

Most of the controversies in traditional metaphysics appeared to me sterile and useless. When I compared this kind of argumentation with investigations and discussions in empirical science or [logic], I was often struck by the vagueness of the concepts used and by the inconclusive nature of the arguments. (Carnap 1963, pp. 44–5)

Metaphysics came under attack for having lost contact with the empirical and for its arguments being irredeemably unsettled.

Today, metaphysics is again the target of deep suspicion. Evidence of this comes from my bookshelf. Many recent books in philosophy of science possess entire chapters strongly condemning contemporary analytic metaphysics (see, for example, van Fraassen 2002; Maudlin 2007; Ladyman and Ross 2007; Maddy 2007). What’s especially remarkable about this is that the authors aren’t logical positivists. They don’t even embrace a common empiricist ideology – for which a distrust of metaphysics is to be expected. Rather, the authors run the full gamut of positions in philosophy of science, and each seeks to make room for some type of metaphysics. Evidence for this flare-up also exists in metaphysics. As I write, so-called ‘metametaphysics’ is all the rage in conferences, books and journals (see Chalmers, Manley and Wasserman...
2009; Braddon-Mitchell and Nolan 2009). Never a good sign for a field, here the literature is in part devoted to whether there are answers to certain types of metaphysical questions.

Clearly something is up. Just as earthquakes are evidence of tectonic plates colliding, so is this dust-up evidence of a collision between two large and slow-moving trends in philosophy. Let me briefly describe the positions that are at odds.

The logical positivists’ critique of metaphysics provides the backdrop. Recall that Carnap understood ontological questions as ultimately about which framework (theoretical structure) one should use. Crucially, he was a framework pluralist. Are there atoms? According to Carnap, one is always working with the entities presupposed by the framework. So if the framework presupposes atoms, the ‘Are there atoms?’ question doesn’t arise. The question to ask instead is why use the framework one is using, but Carnap thought this a purely practical decision. Ontology gets pushed into pragmatics.

The work of Quine and Kripke, however, pulled it out of pragmatics. In our cartoon-like history, we might say that Quine cleared the room for metaphysics, while Kripke furnished it.2

Quine’s part in this is primarily his famous assault on the analytic–synthetic distinction, the line between truths in virtue of fact and truths in virtue of meaning. If Quine (1951) is right, then there isn’t a sharp distinction between conceptual/linguistic truths and factual/contingent truths. This was bad news for the conceptual analysis that dominated philosophy at the time (since there would be no purely conceptual truths). However, it was good news for the possibility of metaphysics. The reason is that Carnap’s notion of a framework presupposed the analytic–synthetic distinction. No frameworks, no framework pluralism – and no place to banish metaphysics. For Quine, the concerns of metaphysicians are not any different than the concerns of scientists:

The question of what there is is a shared concern of philosophy and most other non-fiction genres . . . What distinguishes the ontological philosopher’s concern and [the zoologists’, physicists’ and mathematician’s concerns] is breadth of categories. (1960, p. 275)

Quine tells us that ontological ‘Are there X?’ questions make sense, but only once the statements involved are regimented in decent logical form. That done, one learns that a theory is committed to X’s just in case X’s are in the domain of the variables of the theory. Add to that the claim that the theory is true, and we are doing ontology. The positivist attack on metaphysics is repelled.

Having survived the attempt on its life, metaphysics was left dangling. Quine’s attack on positivism removed a reason not to do metaphysics, but it didn’t provide a particularly clear rationale for engaging in it. Nor was the
saviour much of a fan of metaphysics. In particular, Quine attacked one of metaphysics’ central subjects, (de re) modality. Modal claims about propositions such as ‘necessarily 2 + 2 = 4’ are bad enough, according to Quine, but modal claims about things themselves (that is, de re modality), such as ‘Jack is necessarily human’, are irredeemably confused, he thought.

However, in the 1960s and 1970s modality becomes respectable again. Modal logic was put on stronger foundations, and counterfactuals were given a rigorous semantics. Better than that (from the perspective of metaphysics), using various thought experiments, Kripke (1980) shows that we have robust intuitions about what is possible and that these intuitions carve out a realm of modality not obviously reducible to logical or scientific possibility, namely metaphysical modality. A kind of essentialism is resurrected. If water is actually H2O, we are told, then it couldn’t be anything else. The couldn’t represents metaphysical necessity, and Kripke is credited with discovering a posteriori necessities.

Emboldened by this success, metaphysicians found their subject matter, and one can now find claims such as:

metaphysics is most perspicuously characterized as the science of essence – a primarily a priori discipline concerned with revealing, through rational reflection and argument, the essences of entities, both actual and possible, with a view to articulating the fundamental structure of reality as a whole. (Lowe 2009)

Although not all metaphysicians would agree with Lowe, many would endorse a related division of labour, namely, that metaphysics differs from science in terms of its breadth. Whereas scientists excavate dusty field sites and mix potions in laboratories to tell us which states of affairs are actual, metaphysicians are concerned with what is actual and metaphysically possible. With philosophical intuition about metaphysical possibility unleashed, the journals gradually became filled with increasingly speculative metaphysics, much of it going well beyond Kripke’s a posteriori necessities. These philosophers, I hasten to add, do not take themselves to be exploring, Strawson-style, the architecture of their concepts, but instead feel deeply that their work is no less about mind-independent reality than science is.3

Meanwhile, a parallel set of trends grew – also emanating from Quine – that are, by their nature, suspicious of such metaphysics (see Maddy 2007, Papineau 2009 and Ritchie 2009.). I’m thinking here of the growth of naturalism, broadly conceived, in the forms of naturalized philosophy of science and Quine’s naturalized epistemology. One sees the attitude expressed nicely (and earlier) by Reichenbach:
Modern science . . . has refused to recognize the authority of the philosopher who claims to know the truth from intuition, from insight into a world of ideas or into the nature of reason or the principles of being, or from whatever super-empirical source. There is no separate entrance to truth for philosophers. The path of the philosopher is indicated by that of the scientist. (1949, p. 310)

There is, as Quine puts it, no ‘first philosophy’, no ‘supra-scientific tribunal’ justifying the results of science (Quine 1975, p. 72). Maddy (2007) calls on us to pursue ‘Second Philosophy’ instead. The Second Philosopher ‘simply begins from commonsense perception and proceeds from there to systematic observation, active experimentation, theory formation and testing, working all the while to assess, correct, and improve her methods as she goes’ (p. 2).

This perspective is reinforced by the attack on conceptual analysis by Stich (1991) that inspired later so-called ‘experimental philosophers’. Although contemporary analytic metaphysicians do not see themselves as engaging in conceptual analysis, still they lean heavily on certain modal intuitions. Experimental philosophers doubt the reliability and pervasiveness of many of these intuitions that guide much of contemporary analytic philosophy (Knobe and Nichols 2008).

Finally, another important strand is the increasing number of philosophers of science directly engaged with actual science. While this last group is a motley one, to be sure, many philosophers studying a particular scientific field feel themselves and their projects as closely allied, and even continuous, with the goals and methods of that field.

The collision between these two ‘plates’ was more or less inevitable. Knowledge of the modal structure of reality, when based largely on reflection and intuition, potentially offends against much of what those in the second group believe. Naturalists will want to know how creatures like us gain reliable modal knowledge, Second Philosophers will not see a separate pathway to ontology apart from science, experimental philosophers will challenge the pervasiveness of many of the modal intuitions needed for analytic metaphysics, and those engaged with actual science will see (I suspect) a radical difference between the explanatory and confirmatory aspects of science and of some metaphysics.

3. Metaphysics Walling Itself In

To evoke what he calls the ‘phenomenology of shallowness’ afflicting some of today’s metaphysics, Manley (2009) uses a metaphysical ‘problem’ that he
borrows from Eli Hirsch: when I bend my fingers into a fist, have I thereby brought a new object into the world, a fist? In contemporary metaphysics, a question such as this is viewed as deep, interesting and about the structure of mind-independent reality. Comparable questions in the literature are whether a piece of paper with writing on one side by one author and on the other side by a different author constitutes two letters or one (Fine 2000), whether roads that merge for a while are two roads or one, and whether rabbit-like distributions of fur and organs (etc.) at a time are rabbits or merely temporal parts of rabbits.

Outside metaphysics, many philosophers react with horror at the suggestion that these questions are deep and important. Instead, they find them shallow. The reason is that it’s hard to imagine what feature of reality determines whether a fist is a new object or not. How would the world be different if hands arranged fist-wise didn’t constitute new objects? And if there are debates, aren’t they easily solved? Call temporally extended distributions of fur and flesh in bunny shaped patterns ‘rabbits,’ and non-temporally extended such patterns ‘rabbits,’ Use ‘letter,’ for letters individuated by author and ‘letter,’ for those individuated by paper. And so on. Now, is there any residual disagreement about the non-semantic world? If fists really are new objects, then one imagines that philosophers of science bring two new objects into the world whenever they read this work.

It’s worth thinking through one example in some small detail. Consider the popular topic of simples. A simple is an object with no proper parts. One question that has attracted attention is whether simples with spatial extension are possible. Some philosophers argue that spatially extended simples are not metaphysically possible. Various arguments are marshalled for this conclusion. For instance, suppose the simple has heterogeneous properties, that at one region it is red and at another region it isn’t. Well, doesn’t it then have two parts, the red part and the non-red part, thereby contradicting the idea that it is a simple? Certainly that’s so if one invokes a principle to the effect that, necessarily, an object is red like that only if it has a proper part that is red simpliciter (Spencer 2008). Although we can easily find other examples in the literature, let’s use this no-extended-simples argument for an illustration.

The no-extended-simples argument makes claims about the actual world, namely, that anything actually extended with heterogeneous properties is not simple. Let’s now ask how this claim connects with science. On its face, it seems to contradict any science that posits non-point-like fundamental entities. For instance, on its most natural interpretation, superstring theory – one of the more promising attempts at a theory of quantum gravity – posits extended simples. I say ‘most natural’ because the theory was initially motivated by the fact that the topology of interacting continuous one-dimensional extended entities avoided the ultraviolet divergences that plagued graviton–graviton
scattering. The one-dimensionality of strings is a significant part of the original attraction of the theory. Despite criticism, string theory is a live possibility for describing the entities of our world; however, if you don’t like this example, feel free to switch to any other theory with extended simples.

If they exist, superstrings have some of their properties heterogeneously distributed, for example, nontrivial energy densities across a string. The no-extended-simples argument therefore applies to superstrings. Followed through to its conclusion, we know that superstrings are not the basic building blocks of the world, for they have parts. Reflection on the nature of parts and simples tells us that superstrings are composite. And to the degree that superstring theory leaves out the parts, it is incomplete and not fundamental. For, recall, this argument is not about the regimentation of our concepts; if the argument is right, then strings really are composite. No new colliders need be built to test this – witness all the tax dollars potentially saved!

How do philosophers view physicists’ claims that some simples are extended? Being charitable by nature, philosophers allow that physicists are confused: superstrings aren’t really extended simples. The theory must be reinterpreted in a manner compatible with the terms of art used by metaphysicians. Superstrings can thus be reinterpreted as composites of simple points. This theory of zero-dimensional entities is officially metaphysically possible, unlike superstring theory. But string theory is saved for practical purposes by being empirically equivalent to or best interpreted as a metaphysically possible theory, the metaphysician’s version of string theory. Yet note: the metaphysician’s version must posit strange new laws to ensure that the simples stay together in stringy configurations.

I’ll develop my complaint about these metaphysical parts and the like in subsequent sections. However, let me immediately highlight that I will not find anything directly objectionable about the metaphysicians use of intuition, nor their suggestion of a new interpretation of the physics. Researchers may have good reason to reinterpret, challenge and add to the physics, all in the name of achieving a greater balance of theoretical virtues. My own objection will instead focus on the gruesomeness of the resulting theory described above.

To begin to see the problem, compare the parts we have just found with the ‘partons’ Feynmann famously suggested in 1968. Partons are the point-like elementary constituents of hadrons that eventually became interpreted as quarks. Like parts, partons are supposed to be genuine elements of certain real wholes, discovered theoretically, and immune, in a certain sense, to direct observation (thanks to the later development of quark ‘confinement,’ free quarks never show themselves). But there the similarities end. The parton hypothesis is discriminating, applying to hadrons, and not everything with extension. Even though initially incomplete – how partons interacted via the strong force was missing – parton theory was very richly detailed, containing
both novel predictions and novel explanations, for example, especially explaining the ‘scaling phenomena’ found in inelastic scattering of electrons off protons at high energies. Very generally put, its virtues depended sensitively upon what the rest of the physical world looked like. Parts, by contrast, do not. Unlike the crumbs in cookies, biting the wholes of which they are parts will not reveal them; nor will anything in the physical theory signal their presence. Nor do they offer a theoretical improvement, for the resulting theory is far less simple than one without such parts. Partons emerged ‘red in tooth and claw’ from the competitive jungles of science, possessing all the virtues one would expect, for example, novel prediction/explanation, unification of some of the particle zoo and more. The metaphysical principle about parts, by contrast, arises from peaceful reflection on ordinary objects and language. Metaphysical parts increase the complexity of our systemization of the world without any compensating gain in generality or other theoretical virtues. Any decent theory of scientific confirmation threatens to weed them away.5

How did metaphysics come to this? While deeper diagnoses are certainly possible, I find the source in a subtle shift in what the subject matter of metaphysics is. It is the idea beautifully isolated (but not necessarily endorsed) by Conee and Sider (2005):

Metaphysics is about the most explanatorily basic necessities and possibilities. Metaphysics is about what could be and what must be. Except incidentally, metaphysics is not about explanatorily ultimate aspects of reality that are actual. (p. 203)

In metaphysical modality, metaphysics has found the subject matter over which it has ‘exclusive claim’ (ibid., p. 203). Notice the subtle change of emphasis from earlier metaphysics. Prior metaphysical investigations were primarily directed at providing reasons for believing that the actual world has particular entities or properties in it, for example, God, substantival space, creatures with free will, a moving now. Today, so limited a concern is passé. Metaphysics is after something bigger and more abstract, the structure of metaphysical modality. What it investigates can tell us about the actual world, but only – ‘incidentally’ – because the actual world is one possible world of many.

I submit that this shift in metaphysics’ direction is one major reason for the current clash between metaphysics and philosophy of science. This alternative style of metaphysical theorizing brings with it many unstated changes that offend those more connected to science. Being about what metaphysically must and could be, metaphysics on this conception is forced by the change of target into studying more general abstract principles, such as whether two objects can ever occupy the same place and same time. If the concern is whether this principle holds in the real world, science will be relevant to assessing its
truth. But why should science be relevant to assessing its truth in metaphysically possible worlds wherein science is very different? Plainly it’s not: science, after all, is mostly about the metaphysically contingent.

If Kant, Reichenbach and Carnap worried about metaphysics before, they would really agonize over its contemporary form. Shouldn’t intuitions of what is possible make some contact with science? (From the history of science don’t we learn that many ‘impossibilities’ end up possible, and vice versa?) Perhaps worse, as we’ve just witnessed, even if it pretends to have walled itself off, still this style of metaphysics does make threatening forays into the land of the actual. Independently of what science tells us about the actual world, it purports to tell us what must and must not actually be. When it does this, aren’t we entitled to inquire into the evidence base for such extra-scientific judgments of possibility? One needn’t be Kant or a logical positivist to worry about this development in metaphysics.

4. What’s Not Gone Wrong?

Kant, Carnap, Reichenbach and others criticized metaphysics for being superficial. Then they tried to do something about it, namely, forge a criterion that separates ‘good’ metaphysics from ‘bad’ metaphysics. However, none of these criteria, or any other attempts, have survived evaluation.

Recently, the field known as ‘metametaphysics’ has tried to diagnose what, if anything, goes wrong in these debates. Are two metaphysicians arguing over whether extended simples are metaphysically possible disagreeing about two genuinely different possible worlds? Or is the debate merely verbal? The metametaphysics community is currently divided on this question. Some think that debates like the above are genuine (Sider 2009), others that they are not (Chalmers 2009; Hirsch 2009), others that they are genuine but irresolvable (Bennett 2009), and still others believe that they’re genuine but only in the way debates about fiction are genuine (Yablo 2009).

Some ontological deflationists suggest a criterion to separate the verbal from non-verbal. A debate is verbal, Hirsch (2009) claims, just in case ‘each party ought to agree that the other party speaks the truth in its own language’ (p. 239). The idea is natural enough: those who deny extended simples can agree that people using ‘part’, ‘composite’, ‘simple’ in their opponent’s language speak truly when claiming that there are extended simples; but theists and atheists won’t agree that the other speaks truly. Interpretative charity will map part-talk into something true, but charity only goes so far: atheists won’t find a referent for God in their ontology.

While I admire much of this work, we shouldn’t expect to obtain practical guidance for detecting merely verbal debates from it. What is needed is, in
effect, a theory of *metaphysical equivalence*. When do two semantically distinguishable, but observationally undistinguishable, theories describe two truly distinct metaphysically possible worlds, and when are they notational variants? History with related equivalence criteria suggests that the problem is irredeemably tricky, that we won’t get anything like useful, necessary and sufficient conditions for equivalence. Philosophy of science has grappled with the related question about physically possible worlds for a long time. When do empirically underdetermined theories describe the same world? Positivists deflated the question: according to a verificationist theory of meaning, two theories that can’t be observationally distinguished ‘say’ the same thing. Absent such a criterion, however, we have a problem. We know many theories that are observationally equivalent don’t describe the same world. For instance, one could argue that Putnam’s theory that you are a brain-in-a-vat, stimulated to have the experiences you do, is observationally equivalent to the theory that there is an external world governed by the physics of the Standard Model, yet no one would take them to describe the same world. However, much harder cases lurk nearby. Do Einstein’s curved space-time and Weinberg’s flat space-time-plus-gravitons interpretations of general relativity describe the same world? Do Hamiltonian and Lagrangian versions of classical mechanics? These are open questions. The problem, in brief, is that there are too many moving parts. What is observable is partly theory-laden, what needs and gets explanation is partly theory-laden, and more. I expect all these problems will arise again at the metaphysical level. When the facts themselves are under dispute, interpretative charity for one group may be uncharitable for another.

Nor do I think we can claim that ‘bad’ metaphysics results from asking the wrong questions (which is what Kant thought) or from relying too heavily on speculative intuition (a common claim). It’s important to stress that these types of criteria might unnecessarily constrain science into taking too conservative a stance.

For example, Kepler’s model of the solar system, given the context, was perfectly good science or metaphysics, despite the fact that it was both wildly speculative and, from our perspective, asked the wrong questions. Kepler wanted to know why there are six planets (the number then known) and why they are spaced as they are. His answer, on which he struggled for years, was that planets are attached to concentrically placed spherical orbs, each one of which inscribed or circumscribed one of the five Platonic solids (three-dimensional polyhedral). (See Figure 1.) By ordering these spheres in a specific way, Kepler was able to devise a model that was within 5 per cent accuracy of the then-observed planetary orbits. The theory also made rich new explanations and predictions. For instance, with it he was able to explain features of the orbital period: proceeding from inner to outer planets, the difference in orbital period is twice the difference in orb radii.
Unfortunately for Kepler, there are more than six planets. Even worse, there is no grand symmetry principle dictating the number of relative distances between planets. The pattern of distances between planets is due to contingent initial conditions and isn't the result of any deeper principle. Intuitions about symmetry led Kepler to tackle the wrong questions and also to propose a truly wild metaphysics of the solar system. Yet intuitions about what patterns need explanation and what questions are fruitful are the lifeblood of science. In other cases, for instance, Gel-Man's 1962 symmetry argument for the omega-minus particle, intuitions of symmetry were successful: two new particles were successfully predicted. One attempt panned out, one didn't.

5. Levelling the Field

Instead of attacking our speculative abilities or pretending we know what questions are real ones, I submit that the basic problem with some metaphysics today is the idea that the philosopher and scientist doing ontology are performing fundamentally different and separate jobs. The metaphysician's picture that the scientist works in the lab, discovering the actual world's features, while the metaphysician discerns the wider universe of the metaphysically possible, isn't right. The error is thinking that the science of the actual world
doesn’t affect what one thinks is possible or impossible. The history of science and philosophy amply displays that what we think is possible or impossible hangs on science. Or going in the other direction, the error is thinking that modal intuitions are reliable if they are not connected to a systematic theory of a large domain, one possessing many theoretical and empirical virtues.

Analytic metaphysicians, of course, will grant that the science of the day affects what we think is physically or scientifically possible, but remind us that their claim is about metaphysical possibility and assert that their intuitions are about this wider domain. We have modal intuitions about parts and composites, and these intuitions reveal what is metaphysically, not conceptually or physically, possible.

Against this, I want to claim that there is no interesting species of metaphysical modality that is largely immune to science. Our modal intuitions are historically conditioned and possibly unreliable and inconsistent. The only way to weed out the good from the bad is to see what results from a comprehensive theory that seriously attempts to model some or all of the actual world. If the intuitions are merely ‘stray’ ones, then they are not ones to heed in ontology. In metaphysics we should take possibilities and necessities only as seriously as the theories that generate them.

Is metaphysical modality independent of the usual negotiation of virtues that occurs in the various sciences? I’m afraid that I cannot, in this short space, argue convincingly that it is not. However, in addition to the positive picture sketched below, I can make two relevant points.

First, metaphysical modality is murky. Currently, it is at the juncture of many disputes in philosophy of language, mind and logic. So-called modal rationalists debate modal empiricists (with many internecine disputes), and they in turn debate conventionalists and others. On many of these views, metaphysical modality won’t turn out to be separate from scientific modality and still be substantive.

Second, although Kripke gives us reason to believe in a category we might call metaphysical modality, there is nothing to be found in Kripke’s examples that would warrant thinking of metaphysical possibility as something immune to actual science. Kripke himself remarks that it may be possible to understand the intuitions he is trying to capture using only physical possibility. True, on a narrow reading of physical possibility, whereby chemistry and macrolanguage aren’t included, it can’t handle the claim that ‘water is H₂O’ – for arguably physics doesn’t have ‘water’ in its vocabulary. Yet this doesn’t provide any ammunition for one thinking of metaphysical possibility as immune from science. The interesting feature of Kripke’s necessities, after all, is that they are a posteriori. And the claim that water is H₂O comes from some science, if not solely physics. We might, following Edgington (2004), posit a realm of necessity that includes claims about the constitution of water, necessities from the
non-physical sciences, and more, and refer to it with the more inclusive moniker ‘natural necessity’. This natural modality will be sensitive to science.

In the absence of reasons for thinking that metaphysical modality is independent of science, I submit that we regard this species of modality as we do nomological modality. But isn’t this species of modality itself mysterious? What fact of the world makes it true that light can’t go more than 299,792,458 m/s? Whatever the right story, the answer doesn’t rely on our concept of light. We had that concept well before we knew how fast light could travel. We instead think it’s a feature of light, and even better, space-time structure, that makes this limitation on possibilities true. Let’s begin, then, with the most natural answer: the laws of relativity make this restriction true. What are laws? That, of course, is controversial. Yet note: no matter how they are understood, laws represent the central core of theories, and these are theories that try to systematize/explain the world. We only treat events as possible if they are parts of good systematizations of the world. We think it’s impossible that photons go faster than relativity claims. Why? Because our most powerful theories, the theories upon which we base our explanations and predictions — upon which we even stake our lives — say so. The possibilities for photons don’t arise from stray intuitions or attempts to systematize only semantic intuitions.

What is the source of the possibilities? Some, like modern day Humeans, will think the possibilities arise from the systematization itself. For Humeans, laws are the central core principles of the best systematizations of nature. The modality flows from the systematization (see Cohen and Callender 2009). We can conceive this as a specific version of Putnam’s 1962 claim that possibilities and necessities are always relative to a background theory. Never are claims possible or necessary simpliciter. Others, like non-Humeans, will proceed in the opposite direction: the systematization flows from the modality, not vice versa. Ontologically, the modality is basic and independent of a systematization. Space-time just doesn’t allow light to travel faster than 299,792,458 m/s. Nonetheless, non-Humeans think that explanations and theories appealing to the genuine modalities explain better than those that do not. In fact, that a law explains something well is taken as a symptom that it is representing a genuine modality.

Whatever the story is here regarding the deep question of the source of modality, all hands agree that the reason we have to think photons have certain properties arises from their role in a powerful, explanatorily and predictively accurate theory. Being connected to a good systematization of the world is either constitutive or symptomatic of serious possibilities.

We don’t have to be too strict about this. Scientists are free to devise models of the world wherein (say) the absolute speed of light is not constant. To be taken seriously, however, the comment is not an idle one, but rather one embedded in an alternative systematization of a comparable range of phenomena.
In fact, it’s interesting that one way this possibility is challenged (e.g. Ellis and Uzan 2005) is by pointing out how much the rest of the system hangs on the speed of light being constant – it’s a way of showing that the scientist hasn’t yet discharged her obligation to fit the new possibility into a large and equally good system. We may have all sorts of intuitions about the ‘essence’ of light, but my proposal is that we take such intuitions only as seriously as the theory of which they are a part.

6. The ‘Systems’ Demarcation, or: Are There Laws of Metaphysics?

I began this essay by describing the problematic attitude some philosophers have toward metaphysics. But in the ensuing discussion we have now seen a path that will help us steer between the Scylla of shallow metaphysics and the Charybdis of successful metaphysics. In principle, the division is quite simple; in practice it is difficult.

When trying to figure out what to believe about what there is, there are better and worse theories available to guide one. Not surprisingly, I urge that we rely on the best ones. How do we recognize these? A generation’s worth of philosophers sought and failed to find a clean demarcation between science and non-science. For our purposes, it’s better to describe this as the line between epistemically worthy and unworthy pursuits. No plausible necessary and sufficient conditions were ever found for being epistemically worthy. That doesn’t mean there isn’t a distinction, however. There is a large difference between the modern synthesis in biology and creationism, between chemistry and homeopathy, and so on. The failure to articulate a sharp division means only that what we count as epistemically worthy is quite diverse and assessed along so many dimensions that it’s hard to narrow the criteria down to something simply state-able. That’s why the demarcation project failed, why the instances of pseudo-sciences are easily picked out, but the criteria hard to state. The marks of success are clear: empirical adequacy, simplicity, novel predictions, novel explanations, unification, consilience and more. The metric by which we tolerate trade-offs among these virtues is less clear.

Let’s agree to call the theory or theories that strike the best balance among the above virtues the Best Theory. The Best Theory can be carved up according to various coarse classifications. For instance, if thinking of the standard model in physics, we might divide it up into its theoretical and experimental sides. Such a division is crude, of course, for the experimental aspects contain much theory, and the theoretical aspects must mesh well with experiment. As a result, the partition between the two is not sharp. Another division that one can make is between ‘metaphysics’ and ‘science’. I regard this as merely
a more extreme version of the theoretical versus empirical distinction. To a rough approximation, we can treat metaphysical claims as parts of the Best Theory that are more abstract and distantly related to experiment than the bulk of the theory, that is, science. Through experiment, confirmation and disconfirmation seeps upward through theory, but some bits – such as spatiotemporal continuity – are fairly well insulated. Bear in mind that there is, of course, a lot of theory and meta-theory even in empirical science, but at some point we start classifying the theory and meta-theory ‘metaphysics’.

I like this way of characterizing metaphysics because I am convinced that it is a mistake to think one can just see in isolation that an entity or claim is metaphysical. Modern science is filled with all kinds of odd entities, for example, quarks, but some of these play crucial roles in extraordinarily powerful theories. One can only see that an entity is metaphysical, and further, good or bad metaphysics, by looking at its role in the overall theory. Is a soul metaphysics? Good or bad metaphysics? What about a top quark? Stare at either in isolation and you can’t tell.

With these two divisions – that between epistemically worthy and unworthy pursuits and that between metaphysics and ‘science’ – I can make two claims. First, the metaphysics we ought to strive for should fall on the epistemically worthy side of the first divide. Or, using older terminology, it ought to count as ‘science’ rather than pseudo- or non-science. Here I hasten to add that this means only that it passes muster with our standards for good theories.

Second, I then claim that the metaphysics on the right side of this criterion almost inevitably will be responsive to and deeply connected with the ‘science’ also falling on the right side of this line. This result is almost inescapable, because in our theories we prize unification, cohesion and so on, but also empirical virtues. For a theory to be a good one, it had better meet with some empirical success; but since we value unification, cohesion and so on, the ‘metaphysical’ aspects of the theory will be sensitive to the aspects responsible for empirical success. Our demand for theories on the right side of the demarcation line means that our best theories will possess certain theoretical virtues. These virtues then provide a kind of glue between the more and less theoretical and empirical aspects of our best theories.

I say that this result is ‘almost’ inevitable because, of course, it’s logically possible to detach aspects of the best theory from the theory itself. Experimentalists, statisticians and theorists can also detach themselves from the big picture of the standard model of particle physics being tested at CERN. Similarly, mathematicians, scientists and philosophers can detach the Lagrangian framework or the propensity interpretation of probability from any particular theory and study it alone. This is simply the normal division of cognitive labour. Work on both of these examples is, to a large extent, independent of particular scientific theories. But if we’re actually going to believe in the Lagrangian
framework or propensities and their corresponding modalities, then they still need to earn their way into the best theory like everything else.

My picture is thus entirely symmetric between ‘metaphysics’ and ‘science’.10 Science ought to be on the right side of the demarcation line between epistemically worthy and unworthy pursuits. When it is, it, too, will inevitably be responsive to and deeply connected with metaphysics. Indeed, I think that what we conventionally call science in ordinary affairs is inextricably infused with metaphysics from top (theory) to bottom (experiment). Metaphysics is deeply important to science. Laying bare the metaphysical assumptions of our best theories of the world is a crucial and important part of understanding the world. And metaphysical speculation, when anchored in systematic theorizing connected to epistemically worthy pursuits, can aid our search for new and better theories of the world, and hence, better science.

One might reply that science proceeds perfectly well, while leaving many metaphysical questions unresolved. In a sense that may be correct, especially if one regards ‘perfectly well’ as only making good predictions. However, if we count explanation and understanding as crucial parts of a good theory, as we should, then I don’t agree. Bohr’s quantum mechanics is an excellent predictive theory, but it’s leaving so many metaphysical questions open or confused comes at great cost to explanation and understanding.

In slogan form, my claim is that metaphysics is best when informed by good science, and science is best when informed by good metaphysics.

Once the strict autonomy of metaphysics from science is abandoned, then it may be thought that claims about parts – used in my illustration of ‘shallow’ metaphysics – and such might be vindicated by the same methods that science uses. Perhaps there are laws of metaphysics comparable to the laws of particular sciences? I am here thinking of the metaphysician who claims to be using the same methods as the scientist, namely, a form of inference to the best explanation (Sider 2009).

The answer, of course, is that yes, indeed, in principle parts could play a role in laws of metaphysics. Posit gods, discrete spaces, universals, tropes, quiddities and more. So long as they pay their way, they are fine. There is nothing intrinsically wrong with any of them as posits about the world.

The question, then, is simply whether the putative laws of metaphysics truly survive the ‘red in tooth and claw’ selection scientific norms impose. Here there is nothing general to say. We must simply look at examples and see how they play out. Lacking a theory of ‘metaphysical equivalence’, we can expect cases wherein reasonable people sharing roughly the same epistemic values will disagree. Even in science, this happens regularly. Superstring theory, for instance, is currently under attack for being too distant from various theoretical, and especially empirical, virtues. So is neo-classical economics
under attack from behavioural economics. These debates are a normal part of discovering a systematization of a domain.

The ‘teeth’ of the criterion lay only in the fact – and I think it is one – that in many cases in contemporary metaphysics, the question of whether the possibilities envisioned survive the norms of scientific theory appraisal is as clear as can be. For roughly the same reasons that I don’t subscribe to the possibilities and necessities dictated by various pseudo-sciences – the theories lack too many virtues – I don’t treat as genuine the possibilities and necessities posited by some metaphysics.

What is known as Locke’s Thesis is taken by many to be effectively a law of metaphysics. Locke’s Thesis says that no two things of the same sort can be in the same place at the same time. Is this a core principle of a powerful theory? Give the generalization its due: its simply state-able and certainly seems true of most commonly acknowledged macro-objects. One needs to look hard for counter-examples. As a rule of thumb, certainly one could do worse than employ this generalization. Perhaps it plays a role in finding one’s keys in the morning. Maybe it is even a ‘law’ playing a role in the systematization of one’s life. So if one is interested in the metaphysics of the social world or macro-world, then perhaps a principle such as this may play a role in systematizing.

But the same can be said for the generalization that space is Euclidean. Indeed, the case of mereology in metaphysics is usefully compared with the case of Euclidean geometry. So ingrained in our thinking is this geometry that it took two millennia to see that space could be non-Euclidean. And still today, for local and macroscopic navigation, the possibilities and necessities in Euclid’s geometry hold pretty well. But if we’re interested in the fundamental modal features of space, and most metaphysicians are concerned with the world’s fundamental level, Euclid just isn’t right. The parallel postulate doesn’t have to hold, no matter how intuitive. How do we learn this? We discover that the world does not conform to our Euclidean intuitions by devising a comprehensive theory. Meeting the standards imposed by good theorizing can overturn even the most deeply felt and prima facie modal intuitions.

Assume metaphysicians are after the fundamental structure of reality. In that context, Locke’s Thesis plays no role. Not, at least, since the Pleistocene era has the concept ‘thing’ played a role in any putatively fundamental theory. ‘Things’ are way too vague and general to be useful kinds. Substitute ‘quantum field’ for ‘thing’ and then we can ask what QED says about the principle. The principle’s truth or falsity then follows from a broadly systematized area, not isolated intuitions about whether it’s true. Alternatively, one can choose to define ‘thing’ such that things are, when of the same sort, never in the same place at the same time. That kind of regimentation is fine, so long as one notes that it is indeed regimentation.
7. Conclusion

This chapter has focused largely on the negative. I haven’t had space to properly motivate a ‘scientific metaphysics’. Let me end by briefly defending scientific metaphysics from a common complaint and hinting at how much productive and exciting work there is to be done.

First, the complaint. Does a scientific metaphysics have room for philosophy, for metaphysics, or does metaphysics become the ‘handmaiden’ of science on my picture? My reply is that there is definitely room for philosophy, indeed, a demand for philosophy and metaphysics. As described, good science is informed by good metaphysics. Often, critics of ‘naturalistic’ philosophy paint a picture of scientific metaphysics as being reducible to science, lacking prescriptive force, or merely dotting the i’s in science. This picture has too narrow a view of science, and ironically, too modest a view of philosophy. It is too modest because sometimes just the reverse direction of influence has been the case: science has followed where metaphysics led. Metaphysical assumptions underlie science, and as Friedman (2001) argues, thinking about these (e.g. absolute simultaneity, infinitesimals) often drives revolutionary science. The view has too narrow a view of science because adopting (in ontology) the same general norms that operate in science leaves us an awful lot to do. Remember, these norms are very wide-ranging – they’re just ordinary reasoning ratcheted up in a systematic way. They permit wildly speculative theoretical science, such as inflationary cosmology, alongside experimental science. As for prescriptive force, look at science. Its norms call for unrelenting criticism of rival views, among other things. The journals are filled with critical reviews, analyses, meta-analyses and more. To be in favour of scientifically-informed metaphysics is not to endorse a merely descriptive – a glorified journalistic – take on science. Instead, people knowledgeable of science but trained in philosophy, with its emphasis on logic, clarity, norms of following an argument wherever it leads and so on, can offer distinctive and valuable perspectives on all these questions. The methods of any particular science at any particular time don’t exhaust the ways of properly studying the world.

Second, the advertisement. Science doesn’t cover everything metaphysical that it could or should. As I mentioned at the outset, science often leaves theories only partially interpreted or with significant questions unanswered. Serious gaps in our understanding of gauge fields, quantum theory, evolution and more require our attention. Let me also stress that metaphysics can be prospective as well as retrospective. It needn’t only follow where science leads. It’s very optimistic to think that a new quantum theory of gravity, for instance, won’t be, in part, sensitive to the ontology of quantum mechanics or electromagnetism. And by exploring different conceptions of time, philosophers open up new possibilities to consider in devising a theory of quantum gravity.
Finally, metaphysics can range generally over several scientific fields, asking distinctive questions about how they relate and what they have in common. These aren’t questions usually tackled in a science itself, for obvious sociological reasons, but they are no less important for it.

There are plenty of significant areas of metaphysics in which to work, philosophers are needed for this work, and one hopes that they can sometimes make a distinctive positive contribution.  

Acknowledgements

Thanks to Matt Brown, Nancy Cartwright, Jonathan Cohen, Carl Hoefer, Penelope Maddy, Tim Maudlin, Eric Schliesser, Chris Suhler, Christian Wüthrich, UCSD’s Philosophy of Science Reading Group, the Bay Area Philosophy of Science group, and especially Paul Teller.

Notes

1 The standard philosophy curriculum badly reflects the fact – which I believe is no accident – that the ‘great’ metaphysicians were each conversant with and participants in contemporary science. Although the cognoscenti won’t learn anything new, perhaps the following game will stimulate some readers to investigate this side of philosophy further. Connect the following scientific works with the metaphysician: (a) price theory in economics, (b) a steam engine and calculator (but also calculus, advances in physics, geology, embryology and hydrodynamics), (c) the (alleged) medical benefits of pine tar, (d) advances in thermodynamics and the vacuum, (e) optics and analytic geometry (but also almost everything else), (f) experimental properties of potassium nitrate, (g) the physics of elliptical nebulae and galactic clustering, and physical geography (but also much in the foundations of physics). Answers are in the second to last footnote.

2 For a more thoughtful account, see Price (2009) and the many fine papers on this time period by authors connected with HOPOS: http://www.hopos.org/.

3 The so-called ‘Canberra Plan’ (Braddon-Mitchell and Nola 2009) applied to metaphysics is a bit of a halfway house between traditional and Strawsonian metaphysics. Conceptual analysis determines the Ramsey sentence that best describes the role we want some X to play, for example, causation, but then science tells us what the world is like and whether there is anything that actually realizes that role. The enterprise of metaphysics is then very modest, for Canberra Plan metaphysics assumes that we know what the world is like. But that was what metaphysics originally was supposed to tell us!

4 Hudson (2005): ‘One can also find physicists who apparently endorse the actuality of extended simples, but I can’t help but think that this endorsement often arises from confusing the concept of an indivisible object with that of a mereological simple. Whereas having no parts may certainly be one explanation of the indivisibility of a material object – a law of nature prohibiting certain kinds of separation is another. . . . It may be the physicist’s job, for example, to tell us whether the fundamental entities that physics appeals to are physically indivisible one-dimensional strings, but it is the
job of the metaphysicians to tell us whether those uncuttable things are composite’ (p. 107).

5 As I believe Glymour’s 1980 bootstrapping theory would, for instance.

6 Kepler was hardly alone in thinking this. Even later astronomers were impressed by the distances between planets described by Bode’s Law. From the sun, the planets have distances in proportion to the numbers 4, 4+3, 4+2.3, 4+4.3, 4+16.3, and 4+32.3 (and later, with the discovery of Uranus, 4+64.3). Where is the planet corresponding to 4+8.3, the planet between Mars and Jupiter? Symmetry and intuition tell us it must be there; and to good measure, astronomers agreed. However, as pointed out by Hegel, who is unjustly accused of having decreed that there are necessarily seven planets (Craig and Hoskin 1992), another progression of numbers fits the data just as well, namely, the series 1, 2, 3, 4, 9, 16, 27 from the Timaeus. According to this series, there should not be a planet between Mars and Jupiter.

7 Hudson (2005), incidentally, argues that ‘objects’ can go faster than light, after all. Fortunately for relativity, none of these objects have well-defined masses, energies and so forth.

8 If you’re a philosopher of science who doesn’t believe in laws, you have no debate with me here. You probably still believe in causal principles, mechanisms, invariances or other counterfactual-supporting generalizations, and any of these can be substituted for laws in the following argument.

9 Compare with Leeds (2007, p. 463): ‘What gives the physical modalities their specific content—what makes them the physical modalities—are their rules of use: the kind of reasoning that the physicalist takes to be relevant to a claim of necessity or possibility. Most importantly, our physicalist will take as supporting a claim of necessity the kind of reasoning we all use when we argue that a particular statement is a law of nature... What leads us to classify a statement as a law are, in addition to our conviction that it is true, considerations having to do with its generality, its systematic import, its simplicity and explanatory power.’

10 Note how sharply this view therefore contrasts with Ladyman and Ross (2007). Although similarly motivated, they would make metaphysics inherently parasitic upon science: ‘Any new metaphysical claim that is to be taken seriously should be motivated by, and only by, the service it would perform, if true, in showing how two or more specific scientific hypotheses jointly explain more than the sum of what is explained by the two hypotheses taken separately, where a “scientific hypothesis” is understood as an hypothesis that is taken seriously by institutionally bona fide current science’ (p. 30).

11 Answers to footnote 1: a, Hume; b, Leibniz; c, Berkeley; d, Hobbes; e, Descartes; f, Spinoza; g, Kant.

References


History of science without philosophy of science is blind. . . . Philosophy of science without history of science is empty. Norwood Russell Hanson (1962, p. 580)

1. Introduction

When Norwood Russell Hanson penned that now oft-quoted line in 1962, the year in which Thomas Kuhn’s *The Structure of Scientific Revolutions* (Kuhn 1962) also first appeared, he was most definitely not speaking for the majority of his contemporary philosophers of science. Hanson was deliberately asserting a position contrary to the views of the neo-positivist, philosophy of science mainstream, explicitly contesting the mainstream dismissal of history of science as, in principle, irrelevant to the proper work of philosophers of science. The properly philosophical task was then widely taken to be the formal explication of core methodological concepts, such as explanation and confirmation, and the formal analysis of the structure and interpretation of scientific theories. Logic, not history, was the philosopher’s native discipline. Such properly philosophical work was assumed to be situated in what Hans Reichenbach had dubbed 24 years earlier the ‘context of justification’. Along with the sociology and psychology of science, the history of science was consigned to the ‘context of discovery’, with the small but telling exception – the exception that proves the rule – of ‘rationally reconstructed’ historical episodes that would be used to demonstrate the applicability of the philosophers’ formal explications and analyses (Reichenbach 1938).

In 1962, Hanson and Kuhn were part of a then small minority of dissenters from neo-positivist orthodoxy who believed that a philosophical understanding of science was possible only via the deep historical embedding of the philosopher’s work. The fellow dissidents included several philosophers later to be prized along with Hanson and Kuhn as prescient advocates of a historicized philosophy of science, such as Stephen Toulmin, Gerd Buchdahl,
Ernan McMullin and Georges Canguilhem. Different thinkers saw philosophy embedded in history in different, albeit sometimes complementary, ways. Buchdahl and McMullin would emphasize that philosophy of science, itself, has a history the understanding of which is relevant to the appreciation and assessment of philosophical models of science, partly because philosophical reflection on science is, itself, but a chapter of the larger history of philosophy, possibly a driving force in that history (see, for example, Buchdahl 1963 and 1969 and McMullin 1965a and 1965b). Like Kuhn, Hanson and Canguilhem would emphasize that the science the philosopher studies has a history the details of which complicate the venturing of easy, timeless, formal generalizations about how good science is and should be done (see, for example, Hanson 1958 and 1963, and Canguilhem 1962 and 1968).

Do these briefly sketched ‘dissident’ ideas about the relationship between philosophy and history of science seem platitudinous? We should probably be grateful that the radical dissent of 50 years ago strikes many today as plain common sense. Perhaps there is some progress in history. But the fact remains that, during the 30 or so years following World War Two, philosophical orthodoxy asserted the deep irrelevance of history to philosophy and philosophy to history. It is important that this was also the period during which both of the modern disciplines of philosophy of science and history of science developed their mature intellectual forms, and the period during which they became firmly established institutionally, in academic departments and programmes, in professional associations, in journals and in funding agencies. If, then, we want to understand why, still today, the relationship between history and philosophy of science is a vexed one, why, in spite of their common engagement with the subject matter of science, historians and philosophers do not speak a common language and do not collaborate to the extent one might expect, one must investigate the mid-twentieth-century historical origins of this tension.

Of course, there are two sides to the story. Philosophers of science today routinely complain that a history of science, in which cultural studies and social constructivism have elbowed aside once dominant traditions of intellectual history, no longer has much to offer the philosopher and pays too little heed to philosophical understandings of science. The charge is unfair on both counts, and historians of science share fully the desire to reopen the conversation with philosophers of science, as witnessed by the recent symposium in *Isis* on ‘Changing directions in history and philosophy of science’ (Richardson 2008, Daston 2008, Galison 2008, Friedman 2008). But it is true that the centre of gravity of contemporary history of science scholarship is no longer represented by work like that of Alexandre Koyré, that deeply engaged the philosophical dimensions of science (see, for example, Koyré 1939, 1961, 1965), or that of C. G. Gillispie, whose historiography borrowed heavily from then common philosophical norms of science (see Gillispie 1960). What is forgotten by philosophers who complain about a
divorce of history from philosophy of science is that a large part of the blame lies
with the philosophers’ mid-century repudiation of history. Which brings us back,
again, to the circumstances under which history and philosophy of science parted
ways back then.

The present volume being intended mainly for philosophers of science, the
focus here will be on the philosophers’ role in first causing the divorce and
then seeking a reconciliation, as philosophers of science in the later twentieth
century learned once again how to think historically. This emphasis also
reflects the author’s own intellectual biography as one of the offspring of a
failed marriage. Trained in the formal philosophy of science that dominated in
my intellectual adolescence, much of my career has involved a slow and not
always easy recovery of the ways of the historian (see, for example, Howard
1990 and 1997b). One would like to think that the result is both better philo-
osophy of science and better history of science.

2. Whewell, Mach, Duhem, and the Meyerson Circle

Why, then, the divorce of history and philosophy of science? Let us begin with
what should be an obvious fact, namely, that it was not always this way. If one
goes back to the nineteenth-century origins of history and philosophy of
science as both intellectual disciplines and academic, institutional structures,
one is struck by how common the assumption was then that history and
philosophy of science were two parts of a common intellectual enterprise.
Examples are everywhere to be found, starting with William Whewell, whose
first important book in the philosophy of science, published in 1840, carried
the telling title, *The Philosophy of the Inductive Sciences, Founded upon their
History* (Whewell 1840), and was based upon a massive, three-volume *History
of the Inductive Sciences* (Whewell 1837).

Surely the most important example, if only because of the irony of later
logical empiricist repudiations of history’s philosophical relevance, is that of
Ernst Mach. The chair to which he was called in Vienna in 1895, the first aca-
demic chair anywhere in the world specifically created for history and phi-
losophy of science, and the chair later occupied by the philosopher of science,
Moritz Schlick, founder of the modern Vienna Circle, was originally styled a
professorship for ‘Philosophie, insbesondere Geschichte der induktiven Wissen-
schaften’ [‘Philosophy, especially History of the Inductive Sciences’]. That
description of Mach’s responsibilities corresponds closely to Mach’s own
understanding of history and philosophy of science as a thoroughly hybrid-
ized mode of reflection on science.

Recall the titles of two of Mach’s most influential books: *Die Mechanik in
ihrer Entwicklung historisch–kritisch dargestellt* [The Development of Mechanics
Presented in an Historical–Critical Manner] and Die Principien der Wärmelehre. Historisch–kritisch entwickelt [The Principles of the Theory of Heat: Developed in an Historical–Critical Manner]. Think about what Mach intended to convey with the expression ‘historisch–kritisch’ ['historical–critical']. It would have been well known to every one of Mach’s readers in the later nineteenth century as the then common designation for the method of ‘critical’ Biblical hermeneutics and, beyond that, for all ‘critical’ textual studies, the most famous such exercise in the nineteenth century being Strauss’s Das Leben Jesu [The Life of Jesus] (Strauss 1835–1836). In general, the idea is to take seriously the historicity of a text, its having been written by specific individuals in a specific historical setting, in part for the purpose of querying what is contingent and what, if anything, is timeless or enduring in the text’s content. In many of its deployments, the historical–critical method is regarded as mainly a debunking strategy: to the extent that a text is helpfully regarded as responding to contingent, historical circumstance, to that extent its claims to express decontextualized truth are subverted. Think of it as, in a way, a precursor to the kind of twentieth-century, radical social constructivism that regards cultural artefacts, including scientific theories, as shaped entirely by context, and thus problematizes a simple realist interpretation of theory (for discussion, see Norris 1990 and Hacking 1999). In Mach’s version of the method, the point is to emphasize the moment of historical contingency in the production of scientific theories and, thereby, to encourage a healthy scepticism about the pretence that, say, Newtonian mechanics and gravitation theory is the last, best word in the physics of matter in motion. Mach is, of course, no social constructivist, but the implications of his thus emphasizing the historical contingency of science are too often underappreciated in later secondary literatures and later appropriations of the mantle of his authority, as with the Vienna Circle’s posturing as the inheritors of Mach’s philosophical legacy. Mach’s historical–critical method complements his evolutionary naturalism, or what he termed the ‘biologico-economical’ point of view, according to which scientific theory is helpfully assayed with respect to the mental economies that it effects, the tendency to seek such being understood as an adaptive trait of the human species. Most important for our purposes, however, is that the historical–critical method is, for Mach, a form of philosophical critical analysis precisely because of and through its being a historical mode of critical analysis. Philosophy of science and history of science are here wedded more tightly together than in the work of perhaps any thinker before Kuhn.

Mach’s younger contemporary and fellow positivist, Pierre Duhem, affords us yet another example. Philosophers of science know Duhem as the author of La Théorie physique, son objet et sa structure [The Aim and Structure of Physical Theory] (Duhem 1906), the book that famously propounds the philosopher’s ‘Duhem thesis’, according to which theory choice – at least in instrumentally
dense sciences like physics – is underdetermined by empirical evidence. Alert readers appreciate the wealth of historical examples in the book, but too many philosophers of science forget, or never knew, that Duhem is famous in the history of science literature for a totally different ‘Duhem thesis’, namely, the denial of a revolutionary break with Scholastic natural philosophy at the beginning of the early modern period and the assertion of deep continuities in the development of sciences like mechanics and cosmology from the middle ages, through the Renaissance and into the seventeenth century. This argument grows out of highly original historical research presented in tomes (they were tomes) such as L’évolution de la mécanique [The Evolution of Mechanics] (Duhem 1903), Les origines de la statique [The Origins of Statics] (Duhem 1905–1906) and Le système du monde [The System of the World] (Duhem 1913–1959). Duhem may not have theorized the relationship between historical and philosophical method as explicitly and in the same manner as did Mach, but thickly described historical episodes are again and again central to his way of arguing in his more purely philosophical works, and in some of his most influential books, especially his ΣΩΣΕΙΝ ΤΑ ΦΑΙΝΟΜΕΝΑ, Essai sur la notion de théorie physique, de Platon à Galilée [To Save the Phenomena: An Essay on the Notion of Physical Theory from Plato to Galileo] (Duhem 1908), the historical and philosophical threads cannot be disentangled.

Early twentieth-century France produces still other examples, surely the most important being what some now term the ‘Meyerson Circle’. Émile Meyerson is well known as one of the most influential French philosophers of science of his era. He is well known, also, for the manner in which, in books like Identité et réalité [Identity and Reality] (Meyerson 1908) and De l’explication dans les sciences [On Explanation in the Sciences] (Meyerson 1921), he adduces historical evidence for theses about a priori structures in scientific cognition, foremost among them the conservation laws whereby apparent diversity in natural phenomena is said to be explained in terms of underlying identities. Less well-known is the fact that Meyerson gathered around himself in Paris in the 1920s a group of thinkers who shared his commitment to according history an essential role in critical, philosophical analyses of science, a group that included Koyré; the philosopher, Léon Brunschvicg (see, for example, Brunschvicg 1912, 1922); and the historian of chemistry, Hélène Metzger (see, for example, Metzger 1918, 1923, 1926, 1930 and 1987). Much better known should be the fact that it is core members of the Meyerson circle – specifically Meyerson, himself, Metzger and Koyré – whom Kuhn credits in the preface to The Structure of Scientific Revolutions as having been ‘particularly influential’ in his intellectual development (Kuhn 1962, vi).

It is noteworthy that even though a neo-Kantian like Meyerson is a self-styled opponent of the ‘positivisme nouveau’ ['new positivism'] of Duhem and other radical French conventionalists, like Édouard Le Roy, he and his
companions agree with Duheim that history of science and philosophy of science are intimately related, history being productive of properly philosophical insight into science as a way of knowing. Another noteworthy fact is that there is a direct genealogical link between the historicist philosophers of the Meyerson Circle and the mid-century, French historicist tradition in the philosophy of science represented by Gaston Bachelard (see, for example, Bachelard 1934, 1937) and Canguilhem.

Is one surprised by this century-long history of close and respectful relations between history and philosophy of science? Perhaps most surprising to many who do not follow the burgeoning literature on the history of logical empiricism and the Vienna Circle,⁴ is the extent to which serious history of science scholarship was also promoted from within the Vienna Circle and prized by some of the members of the Circle for its relevance to logical empiricist philosophy of science. The best-known historian of science associated with the Circle is surely Edgar Zilsel (see, for example, Zilsel 1926 and 2000), whose typically Austrian marriage of Marxism and Machist logical positivism shaped his approach to history, as with his famous thesis that experimental science arose in the early modern period when capitalism brought craftspeople into productive interaction with scholarly elites. It was precisely the absence of such interaction in China that the later Marxist historian of science, Joseph Needham, cited – with explicit reference to Zilsel – to explain why a tradition of experimental science did not develop and take root there (Needham et al. 1954–). Zilsel had a complicated relationship with the Vienna Circle, but was not an outsider in the way in which, say, the young Karl Popper was (see Hacohen 2000). No. Zilsel was respected enough within the Vienna Circle to be invited to participate in the Circle’s most ambitious publishing venture, The International Encyclopedia of Unified Science, co-authoring with Giorgio de Santillana the fascicle on The Development of Rationalism and Empiricism (de Santillana and Zilsel 1941). But how could it be that the logical empiricism that later invented the context of discovery–context of justification distinction to enforce a divorce of history and philosophy of science once embraced serious historical work of this kind? The puzzle grows only more acute when we recall that, ironically, Kuhn’s Structure of Scientific Revolutions was first published in 1962 as yet another fascicle in the very same volume of the International Encyclopedia of Unified Science in which the de Santilla–Zilsel booklet appeared. Again, how could this be? The strongly leftist political valence of Zilsel’s history of science provides a clue.

3. The Context of the ‘Context’ Distinction

In the new historiography of the Vienna Circle, it is now commonplace to distinguish a ‘right wing’, at the centre of which stood Schlick and to which
Reichenbach was allied, from a ‘left wing’, whose members included the mathematician Hans Hahn and the physicist Philipp Frank, and whose leader was the economist and sociologist Otto Neurath. Many members of the Vienna Circle were democratic socialists, including Rudolf Carnap and Hans Reichenbach, but even those socialists affiliating with the right wing followed the liberal democrat, Schlick, in insisting on a strict separation between science and its philosophy, on the one hand, and politics, on the other hand. Schlick, Reichenbach and the Carnap who famously declared all normative discourse ‘cognitively meaningless’ (Carnap 1932) were already rehearsing in the pre-emigration days of the early 1930s the standard, postwar, liberal trope of the value neutrality of science and scientific philosophy. The Vienna Circle’s left wing was populated with Austro-Marxists like Neurath. They repudiated Marxist–Leninist revolutionary communism, being also committed to democratic social change, but they were Marxists nonetheless, espousing an Austrian home brew version of Marxism in which it was held that by combining Marx with Mach, that is, by giving Marxism a more sophisticated philosophy of science than one found in Friedrich Engels (see Engels 1878), one overcame the philosophical (and political) naiveté of ‘scientific’ socialism’s dialectical materialism. The left wing of the Vienna Circle argued that there was an ineliminable role for politics and values more generally in both the doing of science and the philosophizing about it. Moreover, like all good Marxists, the members of the Vienna Circle’s left wing took history very seriously, including the history of science.

The division between the right and left wings of the Vienna Circle was philosophical as well as political. The left wing tended to sympathize with Duhem’s theory holism and the thesis of the empirical underdetermination of theory choice, this contrasting importantly with the right wing’s defence of a verificationism that implied the complete empirical determination of theory choice. The right wing argued that, if each fundamental observation term possesses a strict empirical meaning, then so, too, does every cognitively meaningful proposition and scientific theory possess a determinate empirical content, and theory choice is, therefore, in principle, univocally determined by the corresponding experience. If this right-wing Viennese epistemology were correct, then there would be, in principle, no place for any non-empirical factors in theory choice, and so there could be, in principle, no place for politics or values of any kind in science. Neurath and the left wing disagreed. Neurath regarded empirical underdetermination as a fact of scientific life. He excoriated as ‘pseudo-rationalism’ any assertion that there is a purely empirical algorithm for theory choice (Neurath 1913), any claim that there is an ‘induction machine’ (Neurath 1935). Neurath argued that scientific objectivity required honesty about the role of social and political agendas in science, especially in the economics and social science that was his home
scientific terrain. Objectivity, for Neurath, required our exposing these agendas to public view and critical, even scientific, scrutiny. For Neurath, it was precisely the empirical underdetermination of science that opened the epistemological space in which values played an essential role. Liberals like Schlick argued, in the traditional Enlightenment way, that science for its own sake promoted human emancipation, on the theory that the unvarnished truth shall set one free. Austro-Marxists like Neurath argued that emancipation required science as a form of social action, a self-aware science that recognized a place for political agendas. Other things being equal, one was deliberately to choose the theory – in economics, say – that was judged more likely to advantage economically disadvantaged workers.

In its more narrowly philosophical aspect, the philosophical debate between the right and left wings of the Vienna Circle found famous expression in the protocol–sentence debate of 1932–1934. This was ostensibly only an argument about whether a phenomenalist observation language (the right-wing choice) was to be preferred to a physicalist observation language (the left-wing choice). In important ways, however, the protocol–sentence debate was a proxy war. The real issue was whether politics, and values more generally, was to be allowed any role in the logical empiricist picture of science. For if the phenomenalists won the argument, then the kind of verificationism favoured by Schlick would have prevailed, with, again, the implication that theory choice is, in principle, univocally determined by experience, so closing the door to any role for values in theory choice. It is widely believed that Neurath won the protocol–sentence debate, trouncing the empiricist foundationalism of Schlick and winning over Carnap, at least on the core philosophical issues of physicalism versus phenomenalism and theory-holism versus anti-holism. Neurath’s argument was that phenomenalism fails because one cannot reconcile the veridicality that the foundationalist empiricist needs in an empirical basis with the propositional form in which phenomenalist observation sentences would have to be cast were they to be capable of standing in logical relationships of consistency and contradiction with the theories among which we are supposedly to choose on the basis of those observation sentences (Neurath 1932).

With the demise of strict verificationism and foundationalist empiricism of the Schlick variety, the right wing had to retreat to some more defensible ground if it was to continue to oppose Neurath’s politicization of science and philosophy. That new ground was given its classic exposition in 1938 when, in the opening chapter of *Experience and Prediction*, Reichenbach premiered the context of discovery–context of justification distinction. Kindred distinctions were to be found in the literature for some decades prior to 1938, including Karl Popper’s rehearsal of related themes in his 1934 *Logik der Forschung* (Popper 1934). What is distinctive in Reichenbach’s version is not just his
giving the distinction its now-classic name, but also, and more importantly, his making clear – if not explicit – that the target was the left-wing Vienna Circle assertion of a systematic role for values in science and the philosophy of science.

The crucial moment in Reichenbach’s presentation occurs when he is explaining that the context of justification, the site of philosophy of science proper, comprises three tasks: the descriptive, critical and advisory tasks. The descriptive task involves only the aforementioned rational reconstruction of selected historical episodes through which to exhibit the essential structural aspects of science that are, themselves, to be revealed by the logical analyses carried out as the main part of the critical task. The logical analyses central to the critical task constitute the essentials of formalist philosophy of science in the context of justification. But what about the little bit of history that makes up the descriptive task? Recall its being explained above that the inclusion of rational reconstruction in the context of justification is the exception that proves the rule, the rule being the consignment of history to the context of discovery. For the whole point of rational reconstruction is to rewrite history in a way that eliminates all of the real history, foregrounding only the allegedly essential formal aspects of the situation – in explanatory structures, evidential inferences and so forth – and doing this in accord with already assumed formal norms.

What is the advisory task? As it turns out, it is an empty category. Neurath would have argued that the scientist and the philosopher of science have an important role to play in advising the public and the political leadership about goals. Reichenbach disagrees. He argues that, properly understood, the advisory task collapses into the critical task, for on his view, the scientist and the philosopher advise only about means, not about the ends, progress towards which those means serve. Here is exactly how Reichenbach puts the point:

We may therefore reduce the advisory task of epistemology to its critical task by using the following systematic procedure: we renounce making a proposal but instead construe a list of all possible decisions, each one accompanied by its entailed decisions. So we leave the choice to our reader after showing him all factual connections to which he is bound. It is a kind of logical signpost which we erect; for each path we give its direction together with all connected directions and leave the decision as to his route to the wanderer in the forest of knowledge. And perhaps the wanderer will be more thankful for such a signpost than he would be for suggestive advice directing him into a certain path. (Reichenbach 1938, p. 14)

While Reichenbach does not footnote Neurath here, it would have been obvious to any informed reader in 1938 that Neurath was the target, for the
metaphor of the lost wanderers was, famously, the starting point of the 1913 paper in which Neurath first introduced (a) his Duhemian, underdeterminationist epistemology, (b) his assertion that there is an essential role for social and political values in theory choice within the domain of underdetermination, and (c) his favourite term of abuse, ‘pseudorationalist’, for those who would deny such a role for values in theory choice. That paper was titled: ‘Die Verirrten des Cartesius und das Auxiliarmotiv. Zur Psychologie des Entschlusses’ ['The lost wanderers of Descartes and the auxiliary motive: on the psychology of decision'] (Neurath 1913).

So, for the Reichenbach who after World War Two would serve as a paid consultant to the Rand Corporation, one of the most influential, cold war think tanks, the scientist and the philosopher offer only technical answers to technical questions. For Neurath, by contrast, there is a proper role for the philosopher of science to play in articulating and assessing the values, the ‘auxiliary motives’, that drive theory choice within the domain of underdetermination. And advising about which social and political agendas thus self-consciously to deploy in theory choice is likewise, for Neurath, a proper responsibility of the philosopher of science.

This, then, is the context of the context distinction. Reichenbach’s introduction of the distinction and, therewith, his principled denial of the philosophical relevance of real history is, really, all for the purpose of legitimating a depoliticized, socially disengaged view of science and scientific philosophy. That a refugee from Nazism writing from exile in Turkey in 1938 would strike this pose makes obvious psychological sense, for historicizing science and philosophy, highlighting the historical contingency of science and philosophy, exposes them to the risk of the kind of political critique of science all too familiar to scholars, like Reichenbach, who had personally contended against the Nazi critique of Einstein’s theory of relativity as ‘Jewish physics’.

If the story ended there, it would be but a sad footnote in Reichenbach’s biography. But the story did not end there, for Experience and Prediction found a new, large audience after World War Two, and it was then that the doctrine of the discovery–justification distinction started to do its most important work. In the era of the ‘Red scare’ and McCarthyite political persecution, a new academic discipline like the philosophy of science was helped crucially in its ultimately successful struggle for institutional prominence by being able to present itself as inherently apolitical. The United States in the 1950s was not a welcoming home for the socially engaged philosophy of science of an Austro-Marxist like Otto Neurath, just as it was no longer a welcoming home for the comparably socially engaged, pragmatist theory of science of a democratic socialist like John Dewey. That the Marxist historiography of Joseph Needham was one of the more prominent varieties of history of
science in the postwar period only made worse the prospects for a histori-
cized philosophy of science to thrive in the 1950s.

The crucial philosophical move that effected the divorce between history
and philosophy of science was, thus, itself in many ways a highly political act
in a politically charged institutional and intellectual atmosphere. The divorce
of history and philosophy of science in the middle of the twentieth century
was one chapter in the larger political history of the cold war.9

4. History and Philosophy of Science Reunited

But that was then, this is now. The cold war is, itself, ancient history. More
than 20 years have passed since the collapse of the Berlin wall in 1989, and by
that time historicist assaults on logical empiricist orthodoxy – led, of course,
by Kuhn – had significantly weakened the philosophical foundations upon
which dogma like the discovery–justification distinction rested. Not surpris-
ingly, a new openness to history had begun to manifest itself in the philoso-
phy of science arena by the late 1980s. There were various expressions of this
new trend. It was, for example, just one year after the fall of the wall, in 1990,
that HOPOS – the International Society for the History of Philosophy of
Science was founded (see Howard 1997a), partly in response to the then
rapidly expanding new literature answering to that description, some of it
new editions and translations of original source material, such as was
presented in the “Vienna Circle Collection’ of the Boston Studies in the Phi-
losophy of Science, a collection launched in 1973 with publication of the first
of two volumes of Otto Neurath’s papers in English translation (Neurath
1973). Important new secondary literature was also beginning to appear, a
significant example being Alberto Coffa’s The Semantic Tradition from Kant to
Carnap (Coffa 1991). It was in 1992 that a new journal, Perspectives on Science,
was launched with the express purpose of serving as a venue for philosophy,
history and sociology of science, as well as scholarship combining and bridg-
ing these specializations. The trend accelerated during the decade of the
1990s, with richly hybridized work such as Michael Friedman’s pioneering
and seminal study, Kant and the Exact Sciences (Friedman 1992).

What made possible the philosopher’s new embrace of history? It was
partly, of course, that a new generation of scholars was emerging who were
largely innocent of the prejudices and fears engendered and reinforced by the
cold war intellectual politics of the immediate postwar period. It was partly
that, as mentioned, three decades of criticism of logical empiricist orthodoxy
made impossible the merely ritual invocation of shibboleths like the discov-
ey–justification distinction. It was partly that, in the space opened up by the
retreat of logical empiricist orthodoxy, innocent antiquarian interest in the
philosophical history of science and the history of philosophy in science – the kind of interest once expressed in the work of thinkers like Koyré and Metzger – could again thrive. In that same space, one could now again safely assert that even systematic philosophical problems have histories of such a kind that one does not really know what the problem is until one has understood how it came to be a problem in the first place. It was also partly that philosophers puzzled by how the philosophy of science could have thought itself into the cul-de-sac in which the field found itself by the end of the 1980s, with the near death of now old-fashioned general methodology and general philosophy of science, turned to the history of the philosophy of science for answers. Lastly, it was partly a matter of new philosophical projects discovering again that descriptive historical scholarship could function in an evidentiary manner with respect even to systematic philosophical claims with putatively normative import. This was the way in which Mach and Meyerson had once understood the bearing of history upon the philosophy of science.

Surely the most widely discussed such neo-Machian and neo-Meyersonian project of the last 20 years, a historically legitimated and yet systematic, philosophical project, is Friedman’s new defence of the contingent a priori as affording a philosophical meta-framework that respects the historicity of science while locating in science sufficient transient, but also paradigm-independent, structure so as to block the slide into relativism threatened by Kuhnian incommensurability claims (Friedman 2001). Friedman follows the very young Reichenbach (1920) in disentangling apodicticity, which is to say, necessity, from the constitutive or constructive aspect of Kant’s a priori, rejecting the former, and asserting that the latter is, nonetheless, still essential to understanding how scientific cognition works. Constitutively, a priori structures are still required, structures that are a priori in the sense of being necessary conditions for the possibility of our theories having empirical content, but what those structures are might change with progress in science. This is what happens when Kant finally assimilates the lesson of Einstein on relativity.

The present chapter is not the place in which to rehearse all of the details of Friedman’s project, nor to summarize the many critical responses it has evoked. What is important to emphasize here is the essential role of historical evidence in Friedman’s arguments for claims such as that the principle of equivalence plays a contingent a priori role in Einsteinian space-time theory and theory of gravitation, and that some such contingent a priori principles must be at play in all science, even if the specific principles change as one moves from, say, Newton to Einstein. Even more so than was the case with Meyerson’s historical argument for enduring a priori structures in the form of conservation laws (Meyerson 1908), Friedman’s claims about contingent a priori structures can be defended only on the basis of historical evidence.
For if the claim depends in detail upon the specific manner in which these contingent a priori structures change from one stage of history to the next, how else but through careful history can one carry on the argument?

Drill down deeper, and various more specific questions about Friedman’s historical method in the philosophy of science come into view. One of particular interest concerns the comparative evidentiary weight to be accorded to the historian’s reconstructions versus the historical actors’ self understandings.12 But, again, this is not the place to pursue even this important question in detail. The main point to emphasize is simply that the example of Friedman’s Dynamics of Reason project makes vivid, as would many others that one could now cite, the extent to which old assumptions about the alleged separation between history and philosophy of science are no longer valid. Common sense now holds that the marriage of history and philosophy of science is a stable and happy one.

Acknowledgement

A portion of the research upon which this paper is based was supported by NSF research grant no. SES-0724550.

Notes

1 It is worth remembering that Gillispie was the editor of the first edition of Dictionary of Scientific Biography (Gillispie 1970–1980).
3 For background, see Zimmermann (1967).
4 Stadler 1997 (English translation: 2001) is a good place to start.
5 Standard sources include Uebel (1992) and Cartwright et al. (1996).
6 For more on the political situation inside the Vienna Circle and its implications for logical empiricist philosophy of science, see Howard (2003).
7 Uebel (1992) and Cartwright et al. (1996) are good sources for understanding how politics sneaks into the protocol–sentence debate.
8 For a more extensive analysis of the way in which Reichenbach in 1938 is having an argument with Neurath’s more value-laden conception of science and philosophy, see Howard (2006).
9 Howard (2003) and Reisch (2005) provide extensive background on the political embedding of postwar, North American philosophy of science. McCumber (2001) tells a similar story about how specializations like meta-ethics displaced normative ethics owning to similar political pressures during the McCarthy period, and how the more general philosophical turn towards a socially disengaged analytic philosophical project benefitted the discipline’s institutional development in that political context.
10 Domski and Dickson (2010) is the place to go for more on both the Friedman project itself, and what the critics and fans are saying about it.
11 This specific argument is disputed in Howard (2010). As originally formulated by Einstein, the principle of equivalence asserts the physical equivalence of a uniform, homogenous gravitational field and a corresponding acceleration of a frame of reference. It is now customary to distinguish several different forms of the equivalence principle; see Ciufolini and Wheeler (1995).
12 On this point, with specific reference to Einstein and the role of the principle of equivalence, see again Howard (2010).

References

Reisch, G. A. (2005), How the Cold War Transformed Philosophy of Science: To the Icy Slopes of Logic. Cambridge: Cambridge University Press.


This page intentionally left blank
Part II

Current Research and Issues

A. General Issues in Philosophy of Science
Scientific Realism with a Humean Face*

Stathis Psillos

1. Introduction

This chapter offers an intellectual history of the scientific realism debate during the twentieth century. The telling of the tale will explain the philosophical significance and the prospects of the scientific realism debate, through the major turns it went through. The emphasis will be on the relations between empiricism and scientific realism and on the swing from metaphysics-hostile to metaphysics-friendly versions of realism.

2. From Verification to Confirmational Holism

The early stages of the debate were shaped by the verificationist criterion of meaning (VCM). This criterion brings together semantics (issues about meaning) with epistemology (issues about knowledge): meaningfulness is tied to verifiability and meaning to verification. The key idea is that the only non-analytic statements which are meaningful are those whose truth can be established empirically; concomitantly, the meaning of a synthetic statement is its empirical content, viz., whatever part of its content can be established empirically and no more. Verificationism was the principal way to capture the deep anti-metaphysical commitments of logical positivism. It was a way to show that there was no real excess content to metaphysical statements that were supposed go beyond what is verifiable in experience. A semantic criterion with a distinctively epistemological dressing was also supposed to be a criterion for demarcating science from metaphysics.

In Moritz Schlick’s (1932) hands, VCM was meant to deliver science from metaphysics, without, however, revising the rich conception of the world, as this is described by the sciences. Schlick was quite adamant that VCM, in its anti-metaphysics capacity, was leaving the world as described by science intact – a world populated by atoms and fields and whatever else our best science tells us there is – this is what he called ‘empirical realism’. VCM also dictated a certain
solution to the problem of empirically equivalent descriptions of the world. It may well be that the issue between realism and idealism is a pseudo-problem, since VCM licences no empirical difference between the two (see Carnap circa 1928), but – by the very same token – the issue between competing but empirically equivalent scientific theories of the world becomes a pseudo-problem, too. Take, for instance, the rivalry between the general theory of relativity, which has it that the structure of space is non-Euclidean, and a version of Newton’s theory, which keeps Euclidean geometry and posits universal forces, which acts indiscriminately on all moving bodies (e.g. moving rods) and makes them contract accordingly. These two theories, as Reichenbach showed, are empirically indistinguishable. Hence, according to VCM, the very choice between them would end up being a pseudo-problem. The natural reaction to this problem was to claim that, in line with VCM, empirically indistinguishable theories are cognitively equivalent – different but ultimately equivalent (inter- translatable) formulations of the same theory. If rival theories, properly understood, end up being the same theory, it is no longer the case that VCM leaves the world as described by science intact. (This, however, was not Reichenbach’s reaction, who adopted a probability theory of meaning precisely in order to break such deadlocks issued by VCM.)

In any case, verificationist semantics cannot really capture the fine structure of the relations of confirmation between evidence and theory. The challenge here is double. The first comes from the fact that the relations between evidence and theory are probabilistic and not deductive (as was pointed out by Hans Reichenbach 1938). The other comes from the fact that even if we were to stick to a strictly deductive account of confirmation, theoretical assertions always have excess content over their observational consequences (as was realized by Rudolf Carnap in his 1937a). The problem that Carnap uncovered in Testability and Meaning was precisely that insofar as verificationism is uncoupled from reductive versions of empiricism, semantics had to be liberalized in such a way that verifiability gives way to a weaker notion of confirmability.

The move from verification to confirmation was groundbreaking. The ensuing liberalization of empiricism didn’t challenge the subordination of metaphysics to semantics. But, in the longer run, it did pave the way for taking metaphysics seriously. Carnap favoured the ‘requirement of confirmability’: every synthetic statement must be confirmable (CT). This, of course, is an extremely liberal criterion which renders meaningful all kinds of theoretical assertions, as well as nomological statements. But Carnap did not think that the anti-metaphysical stance of empiricism (so far served by VCM) was thereby abandoned. The new liberalized criterion of meaningfulness (CT), Carnap (1937a, p. 35) says, ‘suffices to exclude all sentences of a non-empirical nature, e.g., those of transcendental metaphysics inasmuch as they are not confirmable, not even incompletely.’
This declaration was more like wishful thinking. The problem lied in a fact that Carnap himself was fully aware of, viz., that confirmation is holistic. With due acknowledgement to Duhem and Poincaré, he noted already in 1937:

It is, in general, impossible to test even a single hypothetical sentence. In the case of a single sentence of this kind, there are in general no L-consequences of the form of protocol sentences; hence for the deduction of sentences having the form of protocol-sentences the remaining hypotheses must also be used. Thus the test applies, at bottom, not to a single hypothesis but to the whole system of physics as a system of hypotheses (Duhem, Poincaré). (1937b, p. 318)

Hence, there is a tension in the offering. Confirmation-based semantics does leave the world as described by science intact, but it now seems that metaphysical assertions might not be sharply separable from scientific ones. Almost concurrently with Quine’s famous attack on analyticity and semantic atomism (the two dogmas of empiricism), Hempel (1951) showed in some detail that empiricists had failed to formulate general and precise criteria which can separate some (metaphysical) statements as isolated without rendering other meaningful statements isolated, too. Despite Carnap’s heroic efforts to formulate a criterion by means of which metaphysical statements would end up being isolated, while proper scientific ones, no matter how remote from experience, would not, he failed (see Psillos 2008, for the details).

By the middle of 1950s, there was full recognition among empiricists of the meaningfulness of theories and of the excess content that theoretical assertions have over their observational consequences. Yet, the very movement of thought that led to this conclusion had put in jeopardy the standard empiricist attempt to sharply separate science from metaphysics.

3. The Battle of Empiricism – Phase I

Here starts one of the most interesting battles within empiricism. A significant new addition was Herbert Feigl (1950), whose main point was that once it was accepted that theoretical terms have ‘excess content’, and once VCM is abandoned, it is but a short and harmless step to accept that theoretical terms have factual reference: they designate theoretical/unobservable entities. The so-called surplus meaning of theoretical terms – whatever is not characterized in terms of their observational consequences and their links with other theoretical terms – consists in their factual reference, where ‘in the language of empirical science all those terms (and only those terms) have factual reference which are linked to each other and to the evidential basis by nomological relationships’ (1950, p. 50).
This suggestion may not seem enough to guide ontological commitment. But this is not true. For one, it makes it clear that unobservable entities are no less real than observable entities, given that, as Feigl put it, ‘they are on a par within the nomological framework’ of modern science (cf. ibid.). For another, Feigl’s is an inclusive criterion of reality sharply different from verificationism’s. It states that to assert that something is real is to give it a place within the spatio-temporal-causal framework of science.

Feigl was sensitive to the idea that the adoption of scientific realism – and hence the concomitant criterion of reality – is, ultimately, a matter of convention: it is based on a decision to expand the conceptual framework through which we theorize about the world. This decision, he argued, required a Copernican turn. Whereas empiricism had traditionally started with the world of experience and had aimed to show how the object of science should be made to fit within the object of perception, realism should take the object of perception to fit within the object of science. Better put, perception is epistemically special, because it is through this that human beings get to know what the world is like, but the data of perception (as well as the perceivers) are part of the natural world, as this is described by science, and the question is how they fit into the thus described natural world. So, the move from empiricism to realism requires a change of perspective, which is not dictated by reason or evidence.

Are, then, electrons and the like real? The answer is clearly ‘yes’ if it is seen as being asked within the framework of realism. But if there were some further anxiety as to whether electrons and the like were really real, it would have to be quelled. Feigl shared the view with Carnap that if we take the empiricist critique of traditional metaphysics seriously, there is no framework-free standpoint from which what there is can be viewed. The question of what there is (better: the question of what one is committed to) can only be settled within a framework, and its answer has to do with what types of entity have to be assumed for the framework to play the role it is supposed to.

This kind of rapprochement requires the view that metaphysical (or ontological) issues can be clearly and forcefully distinguished from scientific ones. This was precisely Carnap’s view. Carnap (1950) argued that ontological questions could be asked in two distinct ways: as external questions and as internal ones. He went on to exclude external theoretical questions: questions about the reality of a general type (or category) of entity which are supposed to be settled by looking for (empirical) evidence for the reality of this type or by insight into the metaphysical structure of the world. Questions concerning the reality of a type of entity are legitimate and have content, but only if they are taken to be either external, practical questions concerning the benefits of adopting a certain framework which includes this type of entity in its basic ontic inventory, or as internal, theoretical questions concerning the evidence
there is for (or other reasons for accepting the reality of) certain tokens of this
type, but only after a framework has been adopted.

Although Quine (1951) was sharply critical of Carnap’s distinction, he did
agree with Carnap (and Feigl) on a fundamental point, viz., that there is no
theory-free standpoint from which what there is can be viewed. For him,
however, there is no sharp line between theoretical issues (or questions) and
practical ones. Ontological questions (questions about what there is) are
theoretical questions as well as practical ones: they are answered by our best
theory and there is no extra-theoretical court of appeal. The best theory
(if indeed there is a unique best theory) just is the theory that works suffi-
ciently well – in particular the theory that tallies with the evidence and satis-
fies a number of virtues, most notably simplicity. For Quine, the utility of a
posit and its reality go hand in hand. There is then, no difference between
a framework and the theories within it. The framework itself is a theory
(perhaps a general one) and is judged using the same evidential standards
and pragmatic considerations as in the case of ordinary theories. It follows
that the entities we are committed to are those that are required for the truth
of our overall best theory of the world.

Already in Two Dogmas of Empiricism, Quine had argued for the ‘blurring
of the supposed boundary between speculative metaphysics and natural
science’. But this is emphatically not the metaphysics of the traditional
metaphysician. Quine takes it that ontological questions are on a par with
questions of natural science, but adds that scientific questions are not purely
factual, either, in that they, too, concern the choice of a convenient ‘scheme or
framework for science’. This leads to a blurring of the distinction between the
factual and the conventional, and, in turn, it paves the way for a full-blown
commitment to the reality of the theoretical entities posited by science. Quine’s
master argument for the reality of molecules and the like was that they are
explanatorily on a par with the most ordinary physical objects. Hence, the
denier of theoretical entities is faced with a tu quoque: if you doubt the reality
of molecules, you should doubt the reality of the bodies of common sense.

The philosophical issue then becomes the following: given that theories
should be taken at face value (which is the gist of what Feigl called ‘semantic
realism’), can theoretical entities be dispensed with? And besides, given
semantic realism, can theories be taken to be true? If the answers are ‘no’ and
‘yes’, respectively, (as they were for Feigl, Quine, Sellars and others), the issue
of scientific realism is settled. To be sure, the second question cannot be settled
in an absolute way. The truth of scientific theories cannot be proved. But the
thought was – and it was a great thought – that what matters is the confirma-
tion of scientific theories; given semantic realism, if scientific theories are
well-confirmed, there are reasons to believe in the reality of the theoretical
entities they posit. As Sellars (1963, p. 97) summed this point up, to have a
good reason for holding a theory is ipso facto to have good reasons for holding that the entities postulated by the theory exist.

4. Ramsey-Sentences: A Failed Truce

A prima facie heavy blow to the ineliminability of theoretical terms came from some unexpected quarters and, in particular, from the application to philosophy of science of what came to be known as Craig’s theorem: for any scientific theory $T$, $T$ is replaceable by another (axiomatizable) theory $\text{Craig}(T)$, consisting of all and only the theorems of $T$ which are formulated in the observational vocabulary.

This theorem was readily seized upon by instrumentalists of all sorts. Broadly understood, instrumentalism claims that theories should be seen as (useful) instruments for the organization, classification and prediction of observable phenomena. A clear version of this view can be found in Philipp Frank (1932), whose instrumentalism, in modern terminology, is a form of non-cognitivism: theories are symbolic tools that do not (aim to) represent anything which is not antecedently given in experience. One important argument against non-cognitivism is that it is a reconstruction of science that turns a perfectly meaningful practice – where there is communication and understanding – into a meaningless manipulation of symbols underlied by problematic and context-dependent rules that connect some of the symbols with experience (and hence give them some partial meaning).

Perhaps for reasons such as this, instrumentalism was taken to require an eliminative dimension, associated with Ernst Mach. The idea is that theoretical discourse is, ultimately, eliminable: whatever in experience can be captured with it, it can be captured without it. This eliminative dimension – which had met only with failures – was given a new breath of life by Craig’s theorem. It was argued that theoretical commitments in science were dispensable; theoretical terms could be eliminated en bloc, without loss in the deductive connections between the observable consequences of the theory. If so, the question of whether they refer to unobservable entities becomes moot.

This predicament led Hempel (1958) to formulate what he called ‘the theoretician’s dilemma’. If the theoretical terms and principles of a theory do not serve their purpose of a deductive systematization of the empirical consequences of the theory, they are dispensable. But, given Craig’s theorem, even if they do serve their purpose, they can be dispensed with. Hence, the theoretical terms and principles of any theory are dispensable. But is this dilemma compelling? Hempel himself stressed that it is implausible to think of theories as solely establishing a deductive systematization of observable phenomena. He argued that theories also offer inductive systematizations:
they function as premises in \textit{inductive} arguments whose other premises concern observable phenomena and whose conclusion refers to observable phenomena. Accordingly, theories proper – as opposed to their Craig-transforms – become indispensable in establishing inductive connections between observations.

Two opposed reactions to Hempel’s dilemma were exemplified in the work of Sellars and Carnap. From the indispensability of theories, Sellars drew the obvious realist conclusion that theoretical entities are real. Challenging the empiricist idea that the major role theories have is to explain empirical generalizations (by deductive or inductive systematization), Sellars offered a direct route to commitment to theoretical entities: via their role in the explanation of singular observable events and, in particular, via the theory-based explanation of why some observable entities didn’t behave the way they should have, had their behaviour been governed by an empirically established observational generalization. In Sellars’s view, scientific explanation proceeds via the \textit{theoretical identifications} of observable entities with unobservables. Not only do the latter explain the behaviour of some observable entities; they really \textit{are} the constituents of observable entities. It’s not puzzling, then, that if we take our scientific image of the world seriously, we should be committed to unobservables. (For more on this, see Psillos 2004.)

Carnap, on the other hand, resisted till the end the demise of the distinction between science and metaphysics that the Sellarsian move had consolidated. As is now well documented (see Psillos 1999, chapter 3), he found in the Ramsey-sentence approach to theories a way to capture structuralism, and in particular, the thought that the proper content of a theory (beyond its empirical content) is fully captured by the logico-mathematical structure of the Ramsey-sentence of the theory and its existential implications. In this, let’s call it ‘Ramsey-sentence-structuralism’, Carnap thought he found a stone by which to kill two birds. He could defuse the debate between realism and instrumentalism as being merely about a choice of language, and he could secure the proper empirical content of scientific theories against the spectre of metaphysics. The key idea is that the Ramsey-sentence of a theory fares differently from the Craig-transform of the theory; hence it avoids the problems faced by the latter. The Ramsey-sentence \#T of a theory T has exactly the same observational content as T; it has exactly the same deductive structure; it can play the same role as T in reasoning by means of theory. Recall, however, that to get the Ramsey-sentence of a theory all predicates which are deemed theoretical are replaced with variables which are bound by an equal number of existential quantifiers. Hence, by its very construction, the Ramsey-sentence dispenses with theoretical predicates; it removes, at least prima facie, the issue of the reference of theoretical terms/predicates. Besides, as Carnap was first to note, the very theory T can be written down as a conjunction of two parts: the
Ramsey-sentence $^R \text{T}$ of $T$ and the conditional $^R \text{T} \rightarrow \text{T}$, which came to be known as Carnap-sentence.

The Ramsey-sentence $^R \text{T}$ says that there are classes of entities which are correlated with the observable events in the way the postulates of the theory describe, but it does not say what exactly those entities are; it does not pick out any such class in particular. It can be seen as capturing the structural-cum-empirical content of the theory: what the theory says of the world which can be assessed in terms of truth and falsity. Carnap took it to capture the synthetic component of a theory. The Carnap-sentence $^R \text{T} \rightarrow \text{T}$ should be read thus: if there are entities that satisfy the Ramsey-sentence, these entities are those that render the theory true. But though the Carnap-sentence appears to have genuine empirical content, it does not. Carnap took it to be a meaning postulate, hence, to capture the analytic component of the theory. To be more precise, the Carnap-sentence should be seen as a principle constitutive of the conceptual framework of a scientific theory; it defines (implicitly) its theoretical concepts and ipso facto the object of knowledge of the theory, viz., whatever satisfies its Ramsey-sentence. (For more on this, see Psillos and Christopoulou 2009).

If all had gone according to plan, Carnap would have achieved two things. First, he would have shown that the difference between Ramsey-sentence-structuralism and realism was only about the Carnap-sentence, which is without factual content anyway! Carnap went on to equate Ramsey-sentence-structuralism with instrumentalism and to declare that he had thereby shown that the difference between realism and instrumentalism is essentially linguistic. Second, he would have shown that there would still be room to adopt a form of realism (Ramsey-sentence structuralism) without giving up an essentially anti-metaphysics stance. For the further issue of the supposed excess content of the theory over its Ramsey-sentence (of the reality of theoretical entities, so to speak) turns out to be either an internal, and hence scientifically kosher, issue of what follows from the adoption of a set of implicit definitions offered by the Carnap-sentence, or else an external, and hence metaphysically impotent, issue of choice of a language.

Things didn't go according to plan, however. Ramsey-sentence-structuralism is not the proper way to capture instrumentalism, simply because it yields commitments to the entities (the satisfiers of the Ramsey-sentence) that are not allowed by standard versions of instrumentalism. What is the gain then? Carnap thought there can be a reading of the Ramsey-sentence such that its satisfiers – and hence the commitments that follow from it – are not standard unobservable entities of the sort favoured by realists. He very explicitly took it that, where the Ramsey-sentence says that there are non-empty classes of entities which are related to observable entities by the relations given in the original theory, we are at liberty to think of these classes as classes of mathematical entities (cf. 1963, p. 963). His radical view, then, was that Ramsey-sentence-structuralism

82
was compatible with instrumentalism precisely because it need not imply commitment to the reality of physical unobservable entities. Unless, however, commitment to mathematical entities can be taken to be metaphysically lightweight, the door to metaphysics is wide open again.

In any case, the neutral stance Carnap envisaged faced an unexpected difficulty which came from the fact that, without further assumptions, the truth of the theory collapses to the truth of its Ramsey-sentence. The Ramsey-sentence of a theory can certainly be false, since it might be empirically inadequate. But if it is empirically adequate, it cannot be false, provided that the universe of discourse has the right cardinality. Roughly put, if the world has enough objects, (at least as many as required for the truth of the Ramsey-sentence), the variables of the Ramsey-sentence can be assumed to take those objects as values – whatever they are. So, if the Ramsey-sentence is empirically adequate, the only way in which the world might fail to satisfy it is by not having enough entities to make the Ramsey-sentence true. The problem is particularly astute for Carnap's Ramsey-sentence neutralism, committed as it was to the satisfiers of the Ramsey-sentence being mathematical entities. It is a priori true that there are always enough of them to satisfy the (empirically adequate) Ramsey-sentence of any theory; hence, the very idea of an empirically adequate but false theory becomes an oxymoron – which certainly constitutes a fundamental revision of our conception of theories and of our give-and-take with the world. All this is a version of the much-discussed Newman problem that plagues most versions of structuralism. (For a recent useful discussion, see Ainsworth 2009). The relevant point here is that Carnap's attempt to reconcile empiricism with realism while avoiding metaphysics had to walk the tightrope of Ramsey-sentence structuralism, and its very survival depended very much on what Carnap was willing, in the end, to sacrifice: leaving the image of the world as described by science intact or giving way to metaphysics?

5. Explanation-based Metaphysics

In his review of Jack Smart's groundbreaking work (1963), Quine (1964) exclaimed: ‘With science dominating our lives and progressing ever faster on even more frontiers, it is strange that such a view [the realistic view of fundamental particles of physics] needs urging. Strange but true.’ By the 1960s, the tide had started to move the realists' way. The agonizing over semantic issues had led to a new consensus: realist (that is, face value) semantics. This makes plausible the claim that theories have ‘excess content’ over their observational consequences. In light of this, there is a straightforward answer to the following question: what is the world like, according to a given
scientific theory? (Or, equivalently, what is the world like, if a certain scientific theory is true?) The answer is clear and crisp: the world is the way the theory – literally understood – describes it to be.

The move towards explanation-based metaphysics is most clearly seen in Smart’s claim that the defence of scientific realism rests on an abductive argument. Smart argued, against instrumentalists, that they must believe in cosmic coincidence: a vast number of ontologically disconnected observable phenomena just happen to be, and just happen to be related to one another, in the way suggested by the theory. Scientific realism, on the other hand, leaves no space for such a coincidence; it is because the unobservable entities posited by theories exist that the phenomena are, and are related to one another, the way they are.

This kind of argument pattern bridges the gap between science and metaphysics from the moment it is generally accepted that it is, precisely, inference to the best explanation which is widely used by scientists when they come to accept scientific theories. A recognition such as this is by no means obvious. Pierre Duhem (1906), for instance, put forward an anti-explanationist form of instrumentalism which rested on a sharp distinction between science and metaphysics and claimed that explanation belongs to metaphysics and not to science. Driven by his opposition to atomism and his defence of phenomenological energetics, Duhem envisaged the ‘autonomy’ of physics, which was seen, by and large, as dependent on a strict conception of the scientific method, captured by the slogan, scientific method = experience + logic. And yet, Duhem went on to offer some of the most powerful arguments in favour of scientific realism, the most central being that the fact that some theories generate novel predictions could not be accounted for on a purely instrumentalist understanding of scientific theories. This is a precursor of Smart’s argument for realism and, yet, by insisting on the dichotomy between science and metaphysics and by equating metaphysics with the call for explanation, Duhem remained ambivalent as to the status of this argument for realism.

Once it is accepted that this dual stance is deeply problematic precisely because a metaphysics- (that is, explanation-) free science is a chimera, the door is open for the explanationist defence of realism. Hilary Putnam and Richard Boyd argued that, in light of the fact that inference to the best explanation is the very method scientists use to form and justify their beliefs in unobservable entities, scientific realism should be seen as an overarching empirical hypothesis which gets support from the fact that it offers the best explanation of the success of science. The Putnam-Boyd argument came to be known as ‘the no-miracles argument’ since, in Putnam’s (1975, p. 73) slogan,

The positive argument for realism is that it is the only philosophy that does not make the success of science a miracle.
All in all, the realist turn in the philosophy of science made metaphysics legitimate again. But what the empiricist critique of it left behind – despite its overall failure – can be captured by Feigl’s nice words: ‘if this be metaphysics, make the least of it!’ When it comes to the role of metaphysics in the scientific realism debate, it was confined mostly to a declaration of independence (the world that science aims to describe exists in a mind-independent way) and to a commitment to the independent reality of unobservable entities.

This was not an empty gesture, however. It cut a lot of ice against a species of non-sceptical scientific anti-realism which gained some currency in the 1960s and was motivated by the thought that, while the world as described by science should be left intact, this is not necessarily a mind-independent ready-made world. This thought can be traced to the work of the later Wittgenstein and has been advanced by Norwood Russell Hanson (1958). In this view, what there is and what one is committed to depends on the ‘logical grammar’ of the language one uses to speak of the world, where the ‘logical grammar’ was meant to capture the interconnections of the uses of key concepts that structure a certain language-game. Science is a ‘language-game’ which is characterized by its norms, rules, practices and concepts, though all these are internal to the game: they don’t give the language-users purchase on an independent world. One can then play the science language-game and adhere to its norms and practices. One can follow the scientific method (and in particular the abductive explanatory practices of scientists) and come to accept theories as true as well as believe in the existence of unobservable entities. One, that is, need not be a sceptic. But, on Hanson’s view, one need not (perhaps, should not) add to this non-sceptical approach any realist metaphysics. Nor should one build into the language-game a concept of truth that is evidence-transcendent.

The right realist answer to this challenge was to emphasize that the world comes already structured. That the world has a built-in natural structure is licensed as the best explanation of the friction there is between the world and our scientific theories or paradigms. The presence and persistence of anomalies in scientific theories is best explained by the fact that there is a mismatch between the actual natural structure of the world and the ways in which this structure is modelled by theories.

6. The Battle of Empiricism – Phase II

Non-sceptical anti-realism never became too popular among philosophers of science. The main rival of scientific realism in the last decades of the twentieth century was van Fraassen’s (1980) constructive empiricism. The debate took a distinctively epistemic turn – though this was occasionally disguised by the
fact that van Fraassen characterized both scientific realism and constructive empiricism in primarily axiological terms: realism takes it that science aims at truth, while constructive empiricism takes science as an activity that aims at empirical adequacy. In the background of this axiological characterization was a full endorsement of realist semantics. This endorsement, to be sure, did not dictate acceptance of scientific theories as true or truthlike; it is consistent for an empiricist to suspend belief in the truth of accepted scientific theories and take it that they can, at best, be assessed in terms of empirical adequacy. Why, however, would someone opt for this view unless one believed that truth was either unachievable or at least unnecessary for science?

There is a kind of oscillation between these two views. If it is claimed that truth is unachievable, some positive reason should be offered for this. In particular, the reason should be such that it challenges the ability of scientific method to produce a well-confirmed account of the unobservable structure of the world, and hence it undermines the rationality of belief in such an account. If, however, it is merely claimed that truth is unnecessary – in the sense that science can be made sense of without being taken to deliver truth (about the unobservable world) – belief in truth ends up neither irrational nor unwarranted. Accordingly, there are two ways to view the recent empiricist attempts to resist scientific realism.

The weak way (viz., the view that looking for truth is unnecessary) is meant to be ecumenical. Realism can coexist with constructive empiricism; neither of them is rationally compelling. But this weak way to resist realism faces two problems. The first is that constructive empiricism can be flanked from the left, as it were. For there are weaker positions available that render supererogatory even belief in the empirical adequacy of a theory. For instance, it could be argued that science aims at unrefuted theories and that acceptance of a theory involves only the belief that it is unrefuted. But an unrefuted theory is not necessarily an empirically adequate theory. The second problem is the one faced by Ramsey-sentence structuralism. As has been noted recently by Demopoulos (2003) and Ketland (2004), even with a model-theoretic understanding of empirical adequacy, like the one adopted by van Fraassen, unless further constraints are imposed on the models that render true an empirically adequate theory, empirical adequacy collapses to truth, in the sense that an empirically adequate theory cannot fail to be true. These extra constraints allow for the possibility that the world might not be among the models that satisfy an empirically adequate theory. If, however, such extra constraints are imposed – if, for instance, it is acknowledged that the world has a certain natural unobservable structure which might not be captured by an otherwise empirically adequate theory – constructive empiricism loses out on two counts: it puts its official agnosticism about the unobservable structure of the world in jeopardy and makes a rather significant metaphysical concession
to realism. A notable irony here is that, despite their eminent differences, the non-sceptical version of anti-realism, Carnap’s Ramsey-sentence-structuralism and constructive empiricism, all (and for different reasons) have to come to terms with the realist claim that the world has a certain natural structure.

What if constructive empiricism resists realism the hard way? What if it bases its resistance on the strong claim that the truth (about the unobservable) is unachievable or unavailable? This strong way to resist realism would be sectarian. Realism can no longer coexist with constructive empiricism: realism would be rationally bankrupt. But this strong way to resist realism faces two problems. The first is simply that constructive empiricism would also end up being rationally bankrupt. If the problem with realism was that theories cannot be proved to be true, the very same problem would hold for attempts to prove that theories are empirically adequate. Unless it is shown that assertions about unobservables are confirmed, or otherwise tested in ways that are essentially different from the ways in which assertions about observables are confirmed or otherwise tested, constructive empiricism would be no less precarious than realism. The second problem is that all attempts to show that there is a principled epistemic difference between observables and unobservables have been found wanting.

7. Resisting Epistemic Dichotomies

Van Fraassen’s critique of scientific realism was premised on making a natural distinction between observable and unobservable entities carry the weight of a sharp epistemic dichotomy between those aspects of nature that are knowable and those that are not. It is extremely interesting that many realists followed suit and developed positions that rested on epistemic dichotomies, which, however, were drawn within the realm of the unobservable. The key thought was that there is no problem with having epistemic access to the unobservable in general, but there is a problem with having such access to some aspects of it, or types of it, and so forth. A main reason for this selective scepticism comes from the so-called pessimistic induction. Before we make this link specific, let us have a summary of the most salient epistemically dichotomous positions that are meant to challenge scientific realism.

(Epistemic) Structural realism: The epistemic dichotomy is between knowing the structure of nature and knowing whatever is left to fill out the structure (the unobservable ‘fillers’ of the structure of the world). This is an epistemic distinction among bits of the unobservable world – its structure and its non-structure.
**Entity realism:** The epistemic dichotomy is between knowing entities (and perhaps some of their properties) and knowing the truth of (fundamental) theories.

**Semi-realism:** The epistemic dichotomy is between detection properties of particulars, that is properties of concrete causal structures, and auxiliary properties, that is properties attributed to particulars by theories, but for which there is no reason to believe in their reality, since they are not detected (though they might be detectable and become detected later on; cf. Chakravartty 2007).

**Neo-instrumentalism:** The epistemic dichotomy is between those entities to which there is an independent route of epistemic access (mediated by theories that cannot be subjected to serious doubt) and those entities to which all supposed epistemic access is mediated by high-level theories (cf. Stanford 2006).

As noted already, all these positions draw the epistemic dichotomy within the realm of the unobservable, therefore allowing that there is epistemic access at least to some unobservable parts of reality. That some knowledge of the unobservable is deemed possible is an epistemic victory for realism! But might it not be a Pyrrhic victory?

Not quite! The common denominator of all these epistemically dichotomous positions is precisely this: that there is a principled epistemic division between what can be known of nature and what cannot. So, there is a principled limit to the scientific knowledge of the world. The limit is different in the assorted positions, but it is always principled, definite and drawn by philosophical reflection and argument. What exactly, then, is the philosophical issue at stake? I take it to be this: given the ineliminability of theories from science; given the realist reading of the semantics of theories; given that theories, if true, imply commitment to unobservables; are there reasons to accept the existence of a strict, sharp, robust, and principled dichotomy between the epistemically accessible unobservables and the epistemically inaccessible ones? Only science can tell us what the world is like. Philosophy can only raise some principled challenges to the ability of science to tell us what the world is like. Science might, in the end, not succeed in revealing what the world is like. It might be able to disclose only part of the structure and furniture of the world. But this is as it should be. It would be a totally different matter if there were good reasons – mostly drawn by philosophical reflection on science, its methods and its limits – to believe that we, qua cognitive beings, or science qua an epistemic enterprise, are cognitively closed, in a principled manner, to some aspects of the unobservable world.

But there is no good reason (either a priori or a posteriori) to impose a principled epistemic division between what can be known of nature and
what cannot. There might be parts of nature that science might never be able to map out, but these do not fall nicely within a conceptual category which captures one side of a sharp epistemic dichotomy (the unknown X: the unobservable; the non-structure; the intrinsic properties; the auxiliary properties; whatever-there-is-only-theory-mediated-access-to, and the like). The argument for this realist reaction is book-length – dealing, as it has to, with all current attempts to resist realism. It is offered in Psillos (2009).

As noted already, an important motivation for resisting the epistemic optimism associated with scientific realism – the view that science does succeed in offering a truth-like account of the world – has come from the past failures of scientific theories. The key philosophical move that shaped much of the discussion in the last quarter of the twentieth century was based on the claim that there is no reason to think that current scientific theories enjoy any epistemic privilege over their abandoned predecessors. The so-called Pessimistic Induction over the history of science capitalized on the fact that, despite their empirical successes, many past theories were abandoned and replaced by others. Would it, then, not be natural to conclude that the current ones will face that same fate in due course?

Philip Kitcher (1993) and I (1999) have aimed to resist this ‘natural’ conclusion, by arguing that there are ways to distinguish between the ‘good’ and the ‘bad’ parts of past abandoned theories and by showing that the ‘good’ parts – those that enjoyed evidential support, were not idle components and the like – were retained in subsequent theories. This kind of response suggests that there has been enough theoretical continuity in theory-change to warrant the realist claim that science is ‘on the right track’. To be more precise, the realist strategy proceeds in two steps. The first is to make the claim of continuity (or convergence) plausible, viz., to show that there is continuity in theory-change, and that this is not merely empirical continuity: substantive theoretical claims that featured in past theories and played a key role in their successes (especially novel predictions) have been incorporated in subsequent theories and continue to play an important role in making them empirically successful. But this first step does not establish that the convergence is to the truth. For this claim to be made plausible, a second argument is needed, viz., that the emergence of this evolving-but-convergent network of theoretical assertions is best explained by the assumption that it is, by and large, approximately true.

Note that though this kind of realism is selective – in that it does not imply belief in everything a theory implies – its selectivity is not the outcome of a philosophically driven adherence to an epistemic dichotomy of the kind noted above. It is one thing to accept the (eminently plausible) view that not all parts of a scientific theory are confirmed by the evidence; it is quite another thing to impose a principled distinction between those parts that are confirmed and those that are not, which is supposed to mirror (or constitute) a sharp epistemic
dichotomy. Worrall’s (1989) structural realism – which initiated this selective realist epistemic attitude – should be blamed for a jump, from the sensible view that the actual historical development of theories should teach realists the lesson of being selective in what parts of theories they take seriously as warrantedly telling us what the world is like, to the controversial conclusion that only the mathematical structure of the theory can warrantedly tell us what (the structure of) the world is like.

The emergent consensus is that the unobservable is not, in principle, epistemically inaccessible. Actually, science has succeeded in telling us a lot about a lot of unobservables – a lot that we can warrantedly take to be part of a stable and broadly truthlike scientific image of the world. Why should we have expected anything more from the scientific realism debate?

8. From Neo-Humeanism to Neo-Aristotelianism

The story told so far has been premised on the assumption that it is not philosophy’s job to revise the description of the world, as this is offered by our best scientific theories, but rather to interpret these theories and tease out what the world is like according to them. The empiricists’ critique of metaphysics didn’t defy this assumption, though, as we have seen, it didn’t invariably succeed in steering a steady course between anti-revisionism and anti-metaphysics. But to some empiricists (notably Duhem and van Fraassen), the critique of metaphysics is tied to the critique of explanation by postulation – that is, explanation in terms of unobservable entities and mechanisms. This is supposed to be the pinnacle of inflationary metaphysics.

Why, one may wonder, is explanation-by-postulation inflationary? In a sense, it obviously is: it proceeds by positing further entities that are meant to explain the life-world and its (typically non-strict) laws. But in another sense, it is not. For, if you think of it, it proceeds by positing micro-constituents of macro-objects, whose main difference from them is that they are, typically, unobservable. That a putative entity is unobservable is, if anything, a relational property of this entity and has to do with the presence of observers with certain sensory modalities (of the kind we have), and not others. No interesting metaphysical conclusions follow from this fact, nor any seriously controversial ontological inflation.

If the issue is to avoid metaphysics, a lot depends on how much of it we should want to avoid and how. The empiricist tradition we examined in Sections 2 and 3 didn’t take an anti-metaphysical stance to imply (or to require) what Feigl once called the ‘phobia of the invisible and the intangible’. This tradition, which needs to be resurrected and further defended, avoids metaphysics not by taking a revisionist stance towards modern science (replete as
it is with talk about unobservables) but by noting that, at the end of the day, fundamental metaphysical questions are framework questions and are not dealt with in the same way in which questions about the reality of ordinary entities (be they stones or electrons) are dealt with – the relevant ontic framework must already be in place before questions about the reality of specific entities are raised. In light of this, an empiricist can accept the realist framework – in essence, a framework that posits entities as constituents of the commonsensical entities and relies on them and their properties for the explanation and prediction of the laws and the properties of commonsensical entities – without metaphysical anxieties. (For more on this, see Psillos forthcoming a). Here again, there is a nice compromise: along with empiricism’s anti-metaphysical stance, we accept that there is no framework-free standpoint from which what there is can be viewed; but along with realism, we accept that what there is is what is required for a coherent and unified causal-nomological scientific image of the world.

There is, to be sure, some residual metaphysical anxiety to be quelled. This is partly due to the following predicament: if explanation is the way to do metaphysics, and if metaphysical commitments, like ordinary scientific ones, are the product of the application of inference to the best explanation, where should we stop? Why not accept all the really metaphysically inflationary armoury of the neo-Aristotelians? Indeed, an increasing number of realists wed realism with a neo-Aristotelian view of the deep structure of reality, based on the claim that commitment to this rich metaphysics, which leaves behind the Humean barren landscapes, is licensed by the very method by means of which scientists form and justify their beliefs in the unobservable, viz., inference to the best explanation (cf. Sankey 2008). Where, exactly, one stops seems to be a matter of taste (and perhaps of some argument as to what exactly is the best explanation of whatever the explanandum is taken to be). But many (e.g. Brian Ellis and Alexander Bird) go all the way and adopt metaphysical accounts of causation and laws and dispositional essentialism about properties and natural kinds. As Sankey acknowledges, these are ‘optional doctrines’ of scientific realism. So why should they be bought? It’s not clear to me why. It is one thing to say, for instance, that the world has a pre-existing and determined natural structure – implying as it does that the world is not structurally amorphous. This might well be based on admitting objective similarities and differences in nature – even natural groupings of properties. It is quite another thing to hypostatize kinds or to be an essentialist about them. The same attitude could be had regarding laws of nature or universals or powers and potencies and all the other forbidden fruits of the neo-Aristotelian garden of Eden (for more on this, see Psillos 2005).

One possible problem with this heavyweight neo-Aristotelian conception of scientific realism stems from the fact that some of its advocates have
subscribed to an epistemically dichotomous position, noted above. This creates an important tension – which is very clearly seen in Chakravartty’s case. He, like many others, subscribes to the full panoply of neo-Aristotelianism. At the same time, he takes it that scientific realists should be committed only to the detection, as opposed to the auxiliary, properties of particulars. None of the extra stuff that Chakravartty finds in the world (de re necessities, ungrounded dispositions and the like) are detectable. They are taken to be part of the baggage of scientific realism because they play a certain explanatory role, notably, they distinguish causal laws from merely accidental regularities. So, we are invited to accept a certain set of double standards – one for scientific theories, and another for metaphysics. While in the case of scientific theories, epistemic optimism requires causal contact with the world, thus denying epistemic optimism merely on the basis of the explanatory virtues of theories, in the case of the metaphysical foundations of scientific realism, epistemic optimism is solely the function of explanatory virtues. To put the point somewhat provocatively, the metaphysics of scientific realism ends up being an auxiliary system whose detection properties are Humean regularities and other metaphysically less fatty stuff.

It might be concluded that if one wants to be a neo-Aristotelian scientific realist, one had better not rest one’s epistemic attitude towards theories on a too demanding criterion – and in particular one that cannot be honoured by metaphysical theories. Should, in any case, a realist adopt neo-Aristotelianism simply on the basis that it is the best explanation of, say, the neo-Humean account of the world? My own view on this matter is still developing, but in broad outline it comes to this. If we take IBE seriously, as we should, the answer to the above question should be positive. But, it can be contested that neo-Aristotelianism does indeed meet the best explanation test. One particularly acute problem is that all these denizens of the neo-Aristotelian world (powers, metaphysical necessities, dispositional essences and the like) are themselves unexplained explainers. Though everyone should accept some unexplained explainers, in this particular case, they are more poorly understood than the Humean facts that they are supposed to explain. Another problem is that it is not clear at all how all these heavy metaphysical commitments are related to current scientific theories. They are not born out of current theories. Actually, no particular science, let alone particular scientific theory, can yield interesting metaphysical conclusions, simply because each science has its own specific and particular subject matter, whereas the object of metaphysics (at least as understood by many neo-Aristotelians) is very general and domain-independent: it is the fundamental deep structure (or building blocks) of reality as a whole, abstracting away from its specific scientific descriptions.

Accordingly, neo-Aristotelian scientific realists face a dilemma. They have to proceed top-down, that is, to start from an a priori account of the possible
fundamental structure of reality and then try to mould the actual world as described by the sciences into it. The price here is that there is a danger of neglecting or overlooking important differences between sciences or scientific theories in the ways the world is described and in the commitments they imply. For instance, even though physical kinds might conform to the model of dispositional essentialism, biological kinds might not. Alternatively, they have to proceed bottom-up, that is to start with individual sciences or theories and try to form a unified account of the actual deep structure of reality by generalization or abstraction. The price here is that there is no guarantee that this account can be had.

Be that as it may, this is an important item on the agenda of the current debate over scientific realism. Far from being a dirty word, the m-word is hardly dispensed with. Again, however, at stake is how much of metaphysics we should buy into. In between the neo-Aristotelians and the neo-Humeans are the so-called ontic structuralists. A central motivation for ontic structuralism comes from problems in the interpretation of quantum mechanics and the metaphysics (and the physics) of individuality. But the antidote to issues of metaphysical underdetermination is metaphysical, too; to say (roughly) that objects consist entirely of relationships, or that objects do not exist independently of relational structures, hypostatizes relational structures and makes them the ultimate building blocks of reality. Even critics of metaphysics as a whole, like Newton da Costa and Steven French (2003) and Ladyman and Ross (2007), unashamedly take it that structure is all there is. Flirting with neo-Aristotelianism, (as in James Ladyman 2001), ontic structuralism might well create an explosive metaphysical brew, simply because it will have to amalgamate in the same position the thought that structures are abstract entities with the thought that they are nonetheless the locus of modality and causality.

9. Quo Vadis?

With more than a century of heated debate behind it, one would have expected that the scientific realism debate should have been exhausted. But it isn’t, despite some prematurely issued death certificates. The fact is that the fight over scientific realism is as fierce as ever, though the battleground has shifted to the more murky waters of metaphysics. It remains to be seen whether the resurgence of Aristotelian metaphysics will show that the Humean image of the world, mostly associated with the advent of the Scientific Revolution, will be seen as a small interruption in an essentially Aristotelian image of the world. I very much hope this won’t happen, but its very possibility shows that the scientific realism debate remains, as ever, in contact with the most fundamental philosophical issues.
Note

* Many thanks to my good friend Panagiotis Oulis for the Humean twist in the title.

References


Causation and the Sciences

Ned Hall

1. Introduction

The last several decades have seen a profusion of philosophical work on causation – most of it well outside the confines of philosophy of science. That’s odd. You would have thought that a central aim of any scientific discipline is to map some aspect of our world’s causal structure. So you would have thought that, concerned with illuminating the structure of scientific inquiry, philosophy of science would automatically take up the obvious epistemological and metaphysical questions about causation: How do we find out about causal structure? What is the nature of causal structure? And so, finally, you would have thought no argument is required for the view that philosophy of causation is simply a branch of philosophy of science – any more than one needs to argue, say, that philosophy of space and time is a branch of philosophy of science.

So much for what you would have thought. Since at least the early twentieth century, a strong undercurrent of suspicion has flowed through the philosophical world concerning the scientific legitimacy, or relevance, of any notion of causation. Sometimes this undercurrent bubbles to the surface, as in this famous quote from Russell (1953):

The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.3

Take such suspicions seriously, and you will naturally consider philosophical study of causation pointless – at least, for those who want to understand the structure and aims of scientific inquiry.4

Happily, a countercurrent has pushed in the other direction. Salmon’s important work on causation and explanation provides an obvious example (see, for example, his 1984). More recently, philosophers and scientists have taken an explicit interest in the epistemology of causation, making use of ‘structural equations’ to contribute important advances in this area (see, for example, Spirtes et al. 2001, Pearl 2009; more on structural equations in
A bit more tentatively, philosophers of science have delved into the metaphysics of causation, with results that have helped to clarify why, and in exactly what sense, the causal structure of the world ought to be a target of scientific inquiry (see, for example, Hitchcock 2003, Woodward 2005, Strevens 2009).

This chapter contributes to the anti-sceptical effort. But I differ from some of my allies in holding that an adequate account of causation needs guidance from a background metaphysical position about laws of nature. To do causation right, you need a Grand Metaphysical View. Better would be an awareness of the range of options for such a view, and of the ways in which choices among these options will affect the character of a theory of causation (see also Callender, this volume).

Here is the view I favour, quickly sketched:

Two distinct fundamental features characterize reality as a whole. First, there is a total history of complete physical states. Second, there are fundamental laws that dictate exactly how earlier states generate later states. It is the job of fundamental physics, and fundamental physics alone, to map this basic structure. All other natural facts are ultimately explicable in terms of it.3

Looking at the world just through the lens of fundamental physics, we won’t see the need for any interesting, richly structured concept of causation. True, we could say that each complete physical state ‘causes’ each later complete physical state – but why bother, once we have the fundamental laws in hand?

But the value of a concept of causation derives from details of our actual human predicament. First, we need control over our world. Second, we need to understand it. Third, while grasping a complete, correct physics would obviously facilitate understanding and control, we can only build up to such a grasp by way of piecemeal approximations.4 The scientifi c value of causal concepts is precisely that they facilitate control, understanding and piecemeal approximation. For the basic function of any causal concept is to record – with widely varying degrees of precision and informativeness – facts about how localized aspects of the world at one time nomologically depend on localized aspects at earlier times. Knowledge of such dependencies yields control; think of the case in which localized facts about the future depend on what you do right now. It yields understanding, for this requires, among other things, an organized knowledge of the localized patterns of dependency that the world exhibits (see especially Woodward 2005 and Strevens 2009). And it facilitates the project of constructing better and better approximations to a complete and correct physics, for to know a causal fact is to have a nugget of information.
about how the state of the world at an earlier time leads to the state of the world at a later time.

Finally, it is precisely one of the central aims of the special sciences – by which, we can now see, we should mean any science other than fundamental physics – to articulate the kinds of causal facts that facilitate control, understanding and piecemeal approximation.

That's the view, highly compressed. Now I’ll try to unpack it rather more systematically.

2. Remarks on Methodology

The contemporary causation literature has two off-putting features, especially for those whose interest in causation stems from an interest in science. In both cases, philosophers of science have legitimate cause for complaint, but shouldn't get carried away; the methodological corruption in this literature is not nearly so severe as to deprive it of real importance (again, see Callender this volume).

2.1 The Role of Intuitions

The first off-putting feature concerns the way in which the literature deploys intuitions, in the form both of judgments about specific cases and alleged ‘platitudes’ concerning the nature of the causal relation. A philosopher advances an analysis of the schema ‘event C is a cause of event E.’ Someone devises a fiendish counterexample: a case (typically hypothetical) about whose causal structure we all have a firm intuition not captured by the analysis. Then it’s back to the drawing board, either to add further bells and whistles (perhaps appealing to ‘platitudes’ for guidance), or to scrap the analysis, and try something new.

The following is a quick example. Start with the simple idea that for C to cause E is for it to be the case not only that C and E both occur, but that if C had not occurred, then E would not have occurred. But what if C not only brings about E, but also cuts off some backup process that would have brought about E, had C not occurred? Trouble, surely, since if C had not occurred, E would have occurred anyway, thanks to the backup. The canonical example makes use of the ‘neuron diagrams’ that populate much of the literature (thanks to the influence of Lewis; see in particular his 1986b).

Circles represent neurons; arrows represent stimulatory connections between neurons; lines with blobs on the end represent inhibitory connections. Shading a circle indicates that the neuron fires. The order of events is left to right. In Figure 1, A and C fire simultaneously; C sends a stimulatory signal to D,
causing it to fire, while A sends a stimulatory signal to B. But, since C also sends an inhibitory signal to B, B does not fire. Finally, D sends a stimulatory signal to E, causing it to fire. Figure 2 shows what would have happened if C had not fired.

So we tinker, declaring that what it is to for C (the firing of C) to be a cause of E is for there to be a chain of dependence linking C to E (essentially the analysis proposed in Lewis 1973a). In the diagram, E depends on D (for if D had not fired, E would not have fired), and D in turn depends on C. And we make the tinkering look less ad hoc by appealing to the supposedly a priori insight that causation is transitive.

But wait! Examples of late pre-emption show that this strategy won’t work. Here is a favourite, which for future reference I’ll call ‘Suzy First’: Suzy and Billy, two vandals, throw rocks at a window. Each has perfect aim. But Suzy throws her rock a bit quicker, hence it gets there first. Had she not thrown, the window would have broken all the same. Still, only her throw counts as a cause of the breaking. Yet no chain of dependence links her throw to the breaking.

You can no doubt see how things are going to go from here. Ever more clever analyses will confront ever more clever counterexamples. And you can easily come to wonder why we are bothering – what of value we would learn, even if we succeeded in constructing an analysis of causation that ran the gauntlet of every possible example. As the convolutions pile up, the hope of finding insights into causation that will ground insights into science is fading fast.
This is an overreaction. To be sure, we should repudiate the attitude that treats intuitions as non-negotiable data – an attitude on display in the following quote from Lewis:

When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble. (Lewis 1986b, p. 194)

It makes no sense to think that ‘common sense’ is such an infallible guide to the discovery of an account of causation that will illuminate the structure of scientific inquiry. All the same, common sense is a guide of sorts – namely, to where a potentially useful causal concept or concepts might be found. (Compare mathematics, where ordinary intuitions have oft served as guides to the construction of useful concepts.) The appropriate rule is really this one: If an analysis of causation does not deliver the common-sense answer, that is some evidence – defeasible, of course – that something of importance has been overlooked.

I think that hypothetical cases have important lessons to teach us philosophers of science – some of them, anyway, and in a fashion that may require some subtlety to discern. No reason to keep you in suspense: I think that cases with the structure of Suzy First highlight the importance of mediation by a causal mechanism, and that once we get the right account of what causal mechanisms are, it will naturally fall out that a central task of the special sciences is to discover them and accurately describe their structure. We’ll return to this topic in Section 5.2. For now, the crucial point is just this: We should reject the too uncritical Lewis stance towards intuitions, and adopt instead the moderate attitude that sees intuitions as clues to where potentially fruitful concepts of causation might be found.

2.2 The Proper Target of Analysis
The second off-putting feature of the literature is its almost exclusive focus on expressions of the form ‘event C is a cause of event E.’ Do scientists – at least, qua scientists – really pay much attention to statements of this form? Probably not. So an account of causation that limited itself to analysing sentences like this might seem to deprive itself of interest, at least to philosophy of science.

That’s hasty. Observe, first, that nearby claims clearly matter to science. Take a case where Suzy, by herself this time, throws a rock at a window, breaking it. We should all agree – because it’s true – that Suzy’s throw is a cause of the
window breaking. So what? But our interest might perk up if we are told, in
detail, what it was about her throw in virtue of which it was a cause of the
breaking. Or again, if we are told, in detail, what small variations in the manner
or circumstances of her throw would or would not have led to a breaking
(of exactly what kind). This sort of information can conceivably form the
ingredients for a causal generalization of the kind that the special sciences
traffic in. It is also the sort of information possession of which would give you
a properly fleshed out explanation for why the window broke – yielding an
understanding deeper than the shallow understanding you possess when you
know merely that the window broke because Suzy threw a rock at it. So clarity
about singular causal claims can set the stage for clarity about other kinds of
causal claims, ones that are the lifeblood of the special sciences.

2.3 Two Conceptions of Causation
I indicated above that an account of causation in the sciences should lean
on a grand metaphysical view for guidance. I won’t try to provide a survey
of such views. But one distinction really must be put on the table, between
two profoundly different ways of conceiving the subject matter of the theory
of causation. On one conception – which traces back to Aristotle’s notion
of ‘efficient causation’ – objects (or ‘substances’) in the world possess, in
virtue of their essential natures, sui generis causal powers or capacities, and
it is by reference to the operation of these powers or capacities that we
explain change. We can, if we wish, say that one event is a cause of some
later event, but this style of description misses the metaphysically important
features of the case; what is metaphysically fundamental is always the way in
which one or more objects in the world act on others in virtue of their respec-
tive causal powers or capacities. If you want to understand what causation
is, then you first need to understand what it is for an object to possess a
causal power or capacity.

The second, rival conception grants that we can speak of objects in the
world as possessing powers or capacities. But this is loose talk, explicable in
terms of something more metaphysically fundamental: patterns of depen-
dency in the way that the events that constitute our world’s history unfold.
So a philosophical theory of causation should focus on these dependency
relations.

Which conception one endorses will profoundly affect how one thinks
about scientific inquiry. Each deserves careful development. But for the
sake of tractability, I’m going to plant my flag with the second conception
and assume, from here on, that causal structure is, at bottom, dependency
structure.
3. Laws and Counterfactual Structure

And this dependency structure, in turn, I will take to be a *counterfactual* dependency structure – a structure reflected in the pattern of truth values for counterfactuals relating the state of some localized part of the world at one or more places and times to the state of some localized part of the world at one or more *later* places and times. This proposal raises three obvious questions:

(a) Why understand dependency as *counterfactual* dependency?
(b) What is the right account of the counterfactuals in terms of which dependency is to be understood?
(c) Why the temporal asymmetry? Why not also focus on future-to-past counterfactual dependence?

And one more question, less obvious but quite important: in giving an account of causal structure in terms of counterfactual dependency structure, does it matter what conception one has of the fundamental laws themselves?

Let’s take these questions in turn, starting with a very brief answer to the first: We should understand dependency as *counterfactual* dependency in part because, as this chapter tries to show, doing so is fruitful. And in part for the reason broached in the introduction: The fundamental laws of our world *endow* it with a counterfactual structure; it serves some of our most basic epistemic and practical needs to investigate that structure; causal concepts can be very naturally understood as the essential conceptual tools needed for such investigation. Now let’s turn to the remaining questions.

3.1 A Simple Account of (Simple) Counterfactuals

Take the canonical form for a counterfactual to be the following: ‘If it were the case that A, then it would be the case that B’, where A and B can be filled in by any declarative sentence. In typical cases, A will be false; hence the label ‘counterfactual’. Since the work of Lewis and Stalnaker in the 1960s and early 1970s (see, for example, Lewis 1973b, Stalnaker 1968), it has become standard to assume that truth conditions for counterfactuals are to be given by some kind of *similarity semantics*, for example, the following:

‘If it were the case that A, then it would be the case that B’ is true if among the possible worlds in which A is true, the one most similar to the actual world is one in which B is also true.

The search is then on for an analysis of the relevant notion of ‘similarity’.
To my mind, this has been a mistake, leading at times to pointlessly byzantine constructions (e.g. Lewis 1979). In fact, I think it is not so useful to look for a one-size-fits-all semantics, especially given the incredible variety counterfactuals can exhibit. Instead, I will simply give a plausible recipe (adapted from Maudlin 2007b; henceforth, “the Recipe”) for evaluating a limited, but important, range of counterfactuals.

Require that the antecedent A concerns events at (or around) a specific time t, and the consequent B matters after t. Consider an example. Suzy is standing near a window, not throwing anything. If Suzy had thrown a rock at the window, then the window would have broken. In outline, the recipe for evaluating this counterfactual proceeds in two stages. In the first stage, we construct an alternative state – a nomologically possible state – of the world at the given time t in which, instead of just standing around, Suzy is throwing a rock at the window. In the second stage we extrapolate this state forward, and check to see if the window breaks. If so, the counterfactual is true. If not, false.

How do we construct this alternative state? By making localized changes to the actual state of the world, sufficient to make it the case that Suzy throws, but involving no extra, gratuitous alterations. In part, the prohibition on gratuitous alterations gets taken care of by insisting that the changes be localized to the portion of the world described in the antecedent (i.e. Suzy and her immediate environment, at time t). Thus, we are not to imagine a possible world in which Suzy throws, but the window has been moved, or has had a brick wall erected in front of it, and so forth. But the prohibition demands more; for example, we are not to imagine Suzy throwing a rock with a tiny bomb embedded in it, which will explode well before the rock reaches the window. Don’t expect an exact account of what sort of alteration qualifies as ‘gratuitous’. It is fair to say, though, that the more sharply we can pin this issue down in any given case, the more rigorous use we can expect to get out of counterfactuals – and that will matter a great deal, in many scientific contexts.

Both indeterminacy and context-dependence arise naturally at this first stage. Many rocks lie scattered around Suzy, all ready to hand. So there is surely no fact of the matter about which one she would have thrown. Or suppose that she really hates breaking windows. In a context that makes this disposition salient, we can truly say that if she had thrown a rock at the window, she would have missed (intentionally); in a context that makes her skill salient, we can truly say the opposite. But it does not follow that either indeterminacy or context-dependence creeps in at the second stage, where we extrapolate from the counterfactual state of the world in order to see whether the consequent comes out true. And, indeed, there won’t be if we adopt – as I now propose we should – the obvious story to tell about how such extrapolation proceeds: the counterfactual state of
the world constructed at the first stage is to be evolved forward in accordance with the actual fundamental laws of nature.¹⁰

3.2 Temporal Asymmetry

The Recipe stipulates that the antecedent of a counterfactual concerns a time or times earlier than the time or times covered by the consequent. Why? Not because it’s a Revealed Metaphysical Truth that causation must be asymmetric (and we have an eye towards analysing causation in terms of counterfactuals). A more plausible and, frankly, much more interesting reason is the following: while we can certainly employ the Recipe in the future-to-past direction, the kind of dependence we thereby uncover is unfit for sustained scientific investigation.

Consider a simple example (here I draw on Elga 2000). Suzy throws a rock at a window, breaking it (by herself, say). After the breaking, shards of glass lay scattered on the ground, and the rock lies a bit further off. Imagine this process running in reverse – a physical possibility, given the time reversal symmetry our fundamental laws appear to display. Shards of glass and a rock are lying on the ground; then they fly up into the air in such a way that the shards seal themselves into an unbroken pane, which flexes in such a way as to give the rock (which has travelled in just the right way to be at the right location) a little extra push; it flies through the air; Suzy’s hand closes over it. At time 0, the rock and shards lie on the ground; at later time 1, the window contains an unbroken pane of glass. How does this time-1 fact counterfactually depend on the state of the world at time 0?

Extremely sensitively. Consider what it takes for the shards to fly up and form an unbroken pane. Vibrations in the earth converge beneath each individual shard so as to propel it upwards. Subtle air currents adjust the trajectory of each shard so that it reaches its appointed spot within the windowpane at just the right time. Processes internal to each shard prepare their surface molecules so that bonds will form as the shards come together. So if we examine the localized state of the world within any little patch of ground, or indeed any little volume of air in the surroundings of the window, we reach the following verdict: if the state of the world in that region had been even a little different at time 0, there would have been no unbroken pane of glass at time 1.¹¹

But that means that, while the state of things at time 1 counterfactually depends on the state of things at time 0, it does so in an utterly non-discriminating fashion. Compare the structure of the dependency relations when we run the example in the normal direction: the ultimate state of the window – broken – depends on facts about what Suzy is doing at an earlier time, but does so in a usefully insensitive fashion (i.e. if she had thrown just a tiny bit
differently, the window still would have broken). And the broken state of the window does not depend in any such systematic way on, for example, the state of the little patch of ground a few feet away from Suzy.12

In sum, when we ask, ‘How does the state of the window depend on localized states of the world at a slightly earlier time?’ we can give, in the normal case, a fairly interesting and richly detailed answer. But in the temporally bizarre world, we cannot; all we can say is that it is true of pretty much every localized region near the window that, had its state been different at time 0, there would have been no unbroken window at time 1. And what goes for breaking windows goes for any process that exhibits such thermodynamic irreversibility – which is to say, the vast majority of processes studied by the special sciences. So, while a world that was the time reverse of our own would, of course, have a quite interesting past-to-future fundamental physics, it would not have an interesting past-to-future localized dependence structure. And that is just to say that our own world does not have an interesting future-to-past localized dependence structure. So the reason for the stipulation in the Recipe is something more physically deep than metaphysically deep: given the overall thermodynamic structure of our world, there is no point to considering localized, future-to-past dependence.

The Recipe shows how a world’s fundamental laws can endow that world with a rich, past-to-future counterfactual structure. There is no reason to suppose that science cannot fruitfully investigate that structure. But I suspect that how one thinks such investigation ought to go will depend on the view one takes of the nature of fundamental laws.

3.3 Rival Accounts of Law
Philosophers who hold that there are fundamental laws of nature divide into profoundly opposed camps over their metaphysical nature (see also Psillos, this volume). The most basic split concerns the following questions: Are laws mere patterns in the phenomena (patterns that are in some way salient, to be sure – but still, nothing more than patterns)? Or are they something more, something that somehow governs or constrains those phenomena? I’ll call those who take the first view ‘Humeans’, their opponents ‘anti-Humeans’.

Take the following as illustration. Suppose that physicists had determined Newtonian particle mechanics to be the complete and correct physics for our world. Then we would have learned quite a lot about the world’s fundamental structure: say, that (a) there are (nothing but) finitely many point particles, (b) each possessing an unchanging value for mass and charge, and (c) each moving, at all times, in accordance with Newton’s second law of motion, together with the appropriate force laws. But questions about the metaphysical status of these very claims would remain.
For anti-Humeans, (a)–(c) capture basic facts about the metaphysical structure of nomological possibility and necessity. Now, views about the exact way in which the capturing gets done will vary: Some will think that magnitudes such as mass and charge are fundamentally modal, endowing particles that possess them with the power to affect and be affected by other particles in the way described by the Newtonian equations; others will think of these magnitudes as non-modal, but hold that the world contains some extra metaphysical ingredient that constrains or governs particle behaviour. Either way, an anti-Humean will view our Newtonian particle world as serving up a substantive answer to the question, ‘Why do all particles move, at all times, in conformity to the Newtonian equations?’

The Humean thinks that there is no such substantive answer, for she holds that the fundamental physical magnitudes are non-modal, and that what the laws are supervenes on de facto particle behaviour in a way that undercuts any talk of ‘governing’. The most that can be said is that the fact that the particles so move is a particularly useful or important fact to take note of. On the best version of Humeanism (for which see Loewer 1996, Beebee 2000 and Hall 2010), what makes this fact special is roughly that it captures, in an easy-to-state form, quite a lot of information about the world.

Anti-Humean views have a common feature that is very important to recognize. According to all of them, there are facts about what does, in fact, happen in our world – about what its total history of states is. But in addition, according to each of them the world has a determinate and objective modal structure. The views differ as to where exactly this structure comes from. But they all recognize it. And so they all can (and should) agree that the Recipe provides a valuable conceptual and linguistic tool for articulating this structure.

Matters are otherwise for the Humean. She can, of course, insist that what she calls ‘fundamental laws’ are to be used in giving truth conditions for counterfactuals in exactly the way described by the Recipe. Then she will mostly agree with her rivals about which counterfactuals are true or false. But by her lights, that looks like nothing more than an arbitrary semantic decision to use the ‘if it were the case that . . . then it would be the case that . . . ’ construction in a certain way.

Now, this point is tricky. After all, in some good sense, the way we use our linguistic constructions must involve some semantic decision. Still, if you have a view about the sorts of structures the world objectively contains, then that will motivate judgements as to whether certain ways of representing the world are more or less apt. In the case of our Newtonian particle world, anti-Humeans all think that this world has some sort of objective structure that goes beyond the facts about how particles happen to move. And so each of them will see the counterfactual construction as specially
suited to articulating this extra structure. But for the Humean, there is no such extra structure. There is just a decision that, when evaluating counterfactuals, we must always hold fixed those de facto claims about the world that she counts as ‘fundamental physical laws’.

One wonders at the basis for this decision. Why not allow much greater flexibility? You are wondering what would have happened, had you performed a certain experiment. Well, perhaps the answer will vary, depending on whether you hold fixed physical regularities, or rather chemical regularities, or (yet again) biological regularities . . . and so forth. An anti-Humean would insist that there just is some one thing that would have happened if you had performed the experiment (modulo imprecision in the antecedent of the counterfactual, and any indeterminism in the fundamental laws). That is just to say that the world has, here as everywhere else, a determinate counterfactual structure – in which case we are obliged to hold fixed, in our counterfactual reasoning, those aspects of the world that most directly reflect this structure. But our Humean can make no sense of such insistence. So, what mistake does she think we are making, if, in pluralist fashion, we hold that a question concerning what would have happened under certain counterfactual circumstances gets different (but equally correct) answers, depending on which patterns in the phenomena the context in which the question arises instructs us to hold fixed? Not a factual mistake, surely. What, then?

In what follows, I shall assume, with the anti-Humean, that our world has a determinate counterfactual structure. And it is exactly this structure that provides the materials for causal structure.

4. Causation as Counterfactual Covariation

Return to the guiding idea that was presented in the introduction: causal concepts matter to the sciences, for they provide tools by which those sciences can articulate – typically in an approximate and piecemeal manner – structures of dependency in the world. By means of such articulations we enhance our ability to control our world, and to increase our understanding of it.

4.1 The Central Idea

What account of causation should we favour, if we see causal structure as constituted by patterns of dependency? By this point, the question would seem to answer itself. We should favour a counterfactual account, with the counterfactuals understood by means of the Recipe. And the guiding idea behind the account ought, surely, to be the following: a causal connection
between goings-on in one localized region of space-time and goings-on in some later region of space-time is just constituted by a pattern of counterfactual covariation between these regions.

This is the account of causation Lewis presents in ‘Causation as influence’ (2004), intended to replace his earlier counterfactual account of causation. But there are important differences of emphasis. For Lewis, the name of the game was to give truth conditions for ‘C is a cause of E’, roughly as follows: both events occur, and a suitably wide range of alterations of C would have led to an appropriately corresponding range of alterations of E. Lewis was happy to leave as vague what ‘suitably wide’ and ‘appropriately corresponding’ should mean, since he thought that the ordinary concept of causation he was targeting for analysis was, itself, vague.

But from a scientific perspective, the interest in the account resides in the nature of this correspondence. Who – qua scientist, anyway – cares whether event C is a cause of event E? What we really want to know about is the detailed structure of the counterfactual covariation between the region in which C occurs, and the region in which E occurs – better still, if this detailed structure, or structures near enough like it, gets repeated in other regions of space and time. But this difference of emphasis should not lead us to lose sight of the fact that Lewis has drawn attention to a kind of objective structure the world possesses that clearly deserves to be called ‘causal structure’, and is of clear scientific interest.

4.2 Getting More Sophisticated

So we now have the following slogan: causal structure is localized dependency structure, of the sort articulated by the Recipe; investigating such structure is one of the main tasks of the sciences (the special ones, anyway). Getting more sophisticated requires looking a bit more closely at how we might go about representing localized dependency facts.

Suppose we have one region R – as Lewis might put it, the region in which C occurs – and some later region S (say, the region in which E occurs). How should we capture facts about how the state of things in S depends on the state of things in R? We will evidently need some taxonomy, in terms of which we can specify the actual states of R and S as being selected from a number of possibilities. I suggest that a complete theory of the structure of the special sciences must, in large part, be a theory of such taxonomies. Here I’ll rest content with some general remarks.

First, in very many cases, more than one taxonomy will be available. Something is going on in Suzy’s body. We could describe it in physiological terms, or biochemical terms, or chemical terms and so forth. For short, we should be pluralists about scientific taxonomies.
Second, while we should expect the possibilities articulated by any taxonomy to be exclusive – when some part of the world is in a given state, that is incompatible with it being in any other state – we should not expect them to be exhaustive. Something is going on inside this beaker. As it happens, it can be accurately described in chemical terms, as can many alternatives to it. But suppose we are wondering what would have happened if the beaker had contained a plasma. Then our chemical taxonomy gives out. That doesn’t mean we can’t describe this alternative – just that we need a different taxonomy, borrowed from a different branch of science. In short, typical scientific taxonomies will be limited in scope. Plausibly, the only place where we could find a taxonomy not limited in scope is in fundamental physics.

Third, there is at least a strong presumption in favour of using intrinsic taxonomies – ones that distinguish possible states of a region only on the basis of features that characterize how that region is in itself, independent of how any other region is. Within physics and chemistry, this requirement looks non-negotiable, though the softer sciences may relax it.14

Fourth, a mark of maturation in a scientific discipline is that its taxonomies, and its methods for deploying them in stating generalizations about localized dependence structure, work to substantially rein in the kind of context dependence and indeterminacy in counterfactuals that we touched upon in Section 3.1.

Finally, a sophisticated scientific taxonomy – and even many unsophisticated ones, including those that we more or less tacitly appeal to in ordinary life – will come equipped with standards for comparative similarity that allow one to locate a given possible state within a landscape of surrounding or ‘nearby’ alternatives. To see how ubiquitous this feature is – and to see, at the same time, its importance – consider our prosaic example of Suzy throwing a rock at a window. At one time and place, one thing happens: Suzy throws a rock. At a later time and place, something else happens: the window breaks. How do the goings-on in the second region depend on the goings-on in the first region? A crude way to map the patterns of dependency might deploy the following taxonomies. As for the window, it can either remain unchanged, or crack or shatter (not exhaustive, obviously!). As for Suzy and her throw, we can map the alternatives (some of them, anyway) by allowing the mass and colour of her rock to vary counterfactually. Then – even with these utterly simpleminded taxonomies in play – we will be able to say something explanatorily much more interesting than merely that the window broke because Suzy threw a rock at it. For we can begin to articulate the explanatorily relevant features of the throw, by noting, for example, that the mass of the rock mattered – in that had the mass been lower than a certain threshold, the window would merely have cracked, and had it been lower still, the window would have remained unaffected. And we can note that the colour of the rock
is irrelevant, in that, had the rock been any different colour, the window would have broken all the same.

In short, we highlight what is explanatorily important by contrasting what actually happens with alternatives that are similar and different in certain specified respects. So we can advance the following tentative generalization: A taxonomy contributes to the explanatory power of explanations proffered by means of it to the extent that it comes equipped with similarity metrics that enable such contrasts. Equipped with such a taxonomy, it may even be possible to write down an interesting equation relating the state of $R$ to the state of $S$. Notice that the sign that you are working with a taxonomy that is highly sophisticated in this respect is precisely that the space of alternatives it depicts has a natural mathematical representation.

### 4.3 Advantages

Several advantages accrue to thinking of scientific inquiry as organized around the discovery and description of causal structure, understood as localized dependency structure. Here I will sketch some of the more prominent ones.

#### 4.3.1 The Scientific Topic-Neutrality of Causation

The first point is one that John Campbell has emphasized (in conversation): on the conception of causal structure as localized dependency structure, causal structure turns out to be a highly generic feature of the world. Indeed, the variety of ways in which we can usefully represent such structure seems limited only by our creativity and imagination in coming up with clear and effective taxonomies. And it is evident that these taxonomies can overlap in their coverage: dependency patterns can be discerned in the same stretch of reality at a chemical level, a biological level, a psychological level and so forth. We can thus do justice to the attractive idea that the sciences are all in the business of investigating the causal structure of the world, while leaving ample room for significant differences in their methods of investigation.

#### 4.3.2 The Causal Character of Special Science ‘Laws’

Second, we’re now in a position to correct a tempting mistake concerning the content of generalizations in the special sciences. Let me explain, by means of a pair of contrasting examples.

Start with Newtonian particle mechanics. Presenting that theory, we make some de facto claim about the behaviour of all particles, at all times – say, that this behaviour conforms to such-and-such equations. But then (depending on our views on the metaphysics of fundamental laws), we go on to comment that this claim about particle behaviour has a special sort of status: it is nomologically necessary, or holds as a matter of law, for example. We therefore
respect a sharp distinction between the content of a law – a description of universal but purely de facto behaviour of stuff in the world – and the modal status of that law.

Now switch to, say, this example of a simple generalization from metallurgy: All metal bars that are heated expand in direct proportion to their change in temperature. Of course, this generalization needs to be qualified in order to be true: What if you heat a metal bar so much that it melts? What if you set off a bomb next to it as you heat it? And so on. Set these issues aside (for now – we will return to one of them in Section 5.2). Assuming we have an appropriately qualified version of the generalization, we might think that it likewise just says something about the de facto behaviour of stuff in the world.

Wrong, I think. Rather, what the metallurgical claim really aims to report is not merely that, in all of space and time, there are no examples of metal bars that are heated at one time, yet fail to expand by the appropriate amount shortly thereafter. I think it is intended to report, in addition, that such expansions counterfactually depend on the heatings.

To fail to notice that special science generalizations have this kind of content causes epistemological mischief (as Ward 2010 very nicely argues). Presented with a claim loosely stated in the form ‘all A’s are B’s’, you might wonder why it can’t be tested by looking for examples of non-B’s that are non-A’s: why can’t we test the metallurgical law by looking for examples of things that don’t expand and checking to see that they are not previously heated metal bars? Of course you can point out statistical asymmetries (it is antecedently enormously probable that a randomly chosen non-B will be a non-A, for example), but there is a better point to make: The idea that you could test a claim to the effect that the behaviour of a metal bar counterfactually depends in a systematic way on whether and how it is heated, without doing direct experimentation on such bars, is ridiculous on its face.

4.3.3 Foundations of Structural Equations
Among philosophers and scientists interested in causation, one idea has gained currency in recent years that would seem to offer an alternative to our approach. On this rival view, causal structure is to be analysed by means of causal models: roughly, systems of appropriately chosen variables, together with structural equations that relate them. Such models allegedly provide tools by which to analyse, in a controlled and rigorous fashion, certain specialized counterfactuals in terms of which causation is to be defined.

Alas, there are two painfully obvious foundational questions that, to my knowledge, the literature in this area has yet to adequately answer: What are variables? And what are the truth conditions for structural equations?16

Interestingly, our foregoing discussion suggests an obvious prescription for choosing variables and values, for an arbitrary system for which we might
wish to construct a causal model: First, find a way to ‘carve up’ the system into discrete, well-defined subsystems. Second, for each relevant subsystem, and each relevant time or time interval, introduce a variable to characterize the intrinsic physical state of that subsystem at that time, or during that time interval. That is to say: impose some scientific taxonomy.

Furthermore, there is a simple, attractive – I would almost like to say inevitable – story to tell about structural equations. We should take them to describe patterns of localized dependence of just the sort articulated by the Recipe – or rather, by the Recipe when it is extended in a natural way to accommodate counterfactuals whose antecedents specify alterations to the state of the world at one or more localized times and places. But this is easily done (see Hall 2007 for details). In short, the scandalous state of the literature on structural equations is easily corrected. But once the correction is in place, we can also see that that literature has overinflated the significance of structural equations and the causal models that make use of them. Far from providing a foundation for notions of counterfactual dependence and causation, these models turn out to be nothing more than useful tools for representing antecedently understood patterns of counterfactual dependence. Which is fine, given that their main value is for the epistemology of causation: after all, when it comes to testing some complicated causal hypothesis, half the battle consists in representing that hypothesis perspicuously enough that the way evidence bears on it can be sharply discerned. If anything, answering the foundational questions about causal models should remove any basis for what would otherwise be reasonable suspicion regarding their use in science.

4.3.4 Putting Metaphysics in Its Place

Adopt our conception of causation, and you are well positioned to guard against certain errors that have infected philosophical discussions in which theses about causation figure prominently. I will consider just one famous example, drawn from philosophy of mind.

Suppose you are a physicalist. You think that all mental phenomena are somehow grounded in physical phenomena. But not because you think that mental events are literally identical to physical events; rather, the latter ‘realize’ the former. And realization is not identity: when a given mental event M is realized by, say, some neurophysiological event P, it could have been realized by any of many quite different physical events.

Along comes an argument that your view has untenable consequences (see, for example, Kim 1993). You would surely like to say, for example, that a certain mental event M that an agent undergoes causes some later physical behaviour on her part. But then you are guilty of positing rampant causal overdetermination, for this physical behaviour is – given the ‘causal closure of
Causation and the Sciences

the physical’ – already caused by M’s realizer P. And that is unacceptable. It seems, then, that on pain of positing rampant overdetermination, physicalists about the mind face a dilemma: either hold that the mind is causally inefficacious, or hold that mental events, properties and so forth are strictly identical to their physical realizers.

The right response is to ask, ‘What’s so bad about rampant causal overdetermination?’ If you thought of causation as some kind of metaphysical juice, such that for anything to happen requires a certain amount of it, then you might have an answer: ‘Rampant overdetermination is implausible, because we should assume, in Razorish fashion, that the world contains no more juice than is necessary to get things to happen.’ But that answer relies on a confused metaphysical picture. Causation is, at bottom, localized counterfactual dependence of a certain sort. An agent’s body moves at a certain time. The way that it moves can depend in all sorts of subtle and interesting ways on aspects of her prior neurophysiological state, and likewise on aspects of her prior mental state, even if the latter is not strictly identical to the former. These patterns of dependence are underwritten in a straightforward way by the fundamental physical laws – and, crucially, underwritten in a way that carries no suggestion whatsoever that one pattern of dependence will somehow metaphysically ‘crowd out’ all others. There simply is no ‘exclusion problem’.

5. Problems and Open Questions

The account sketched here will only earn real plausibility once it is developed in detail. I want to close by mentioning two clear and serious problems that will need addressing, and by drawing attention to some philosophically interesting issues about the sciences that these problems bring into focus.

5.1 Causation in the Social Sciences

Consider counterfactuals used to describe features of social interactions. Billy and Suzy are lunching together. Billy is wearing a blue T-shirt. What would Suzy’s reaction be if he were wearing a green T-shirt? The sensible answer is ‘Nothing’; she doesn’t really pay much attention to how he dresses. The crazy answer is ‘Oh, my God! How did your T-shirt suddenly change colour?’ But notice that if we apply the Recipe to this example flat-footedly, we get exactly the crazy answer: make a localized change to the state of the world at the relevant time (i.e. switch the colour of Billy’s shirt from blue to green); leave the rest of the state of the world unchanged (which state includes Suzy’s memory of seeing him wearing a blue T-shirt just moments before). Update in accordance with the laws.
Obviously, we need to add to the Recipe that, somehow, appropriate adjustments are to be made to the surrounding state of the world in cases like this. It is a wide open question how to effect this modification of the Recipe in a clear and rigorous fashion. But it is also an urgent question, if we wish to understand counterfactual, and therefore causal, reasoning in any of the social sciences.

5.2 Late Pre-Emption and ‘No Interference’ Clauses
Next, there are two problems that superficially appear unrelated, but that have, I think, deep interconnections.

First, the idea that causal structure is localized dependence structure does nothing to capture the asymmetry that is so vividly on display in cases like Suzy First (in which, remember, Suzy and Billy both throw rocks at a window, but hers gets there first). One option is to try to discern this asymmetry in subtle features of the localized dependence structure. For example, we might observe that the exact time at which the window breaks is a bit more counterfactually sensitive to the timing of Suzy’s throw than it is to the timing of Billy’s throw. For various reasons, I think that is a dead end (see Hall and Paul 2003 for details). A second option is simply to ignore the intuitive difference, reminding any doubters that our aim is to find a concept of causal structure of use in illuminating the structure of scientific inquiry, not a causal concept that does justice to ordinary intuitions. I hold out hope for a third option, which is to construct a more refined account of causal structure that does justice, in cases like Suzy First, to the very natural thought that what distinguishes Suzy’s throw from Billy’s, causally speaking, is precisely that a process of the right type connects her throw to the breaking, whereas no such process connects his throw to the breaking. Here I will simply sketch what I think is the right approach (see Hall 2004 and 2005).

We should understand ‘processes’ simply to be sequences of events (not necessarily spatiotemporally contiguous). These can be compared with one another, and classified into types, on the basis of their intrinsic characteristics. Furthermore, some processes can be distinguished as ones that unfold from certain initial conditions, in highly sanitized circumstances in which nothing else is happening at the time that those initial conditions obtain.17

So, we can ask what a process looks like that begins with Suzy throwing a rock at the window, in circumstances in which Billy (and, for that matter, anything else in the environment whose presence might muck up the counterfactual relationships typically diagnostic of causation) is absent. Such a process will have certain distinctive features. Finally, it is because one can find, in Suzy First, a process sharing those features that connects Suzy’s throw to the breaking – but no process sharing those features that connects Billy’s throw to
the breaking – that Suzy’s throw counts as a cause of the breaking and Billy’s does not.

Suppose this approach works. Surprisingly, the tools it deploys can be brought to bear on the seemingly very different problem of making sense of the ‘no interference’ clauses that unavoidably adorn most generalizations in the special sciences. Metal bars, when heated, expand in proportion to their change in temperature. That is, provided that nothing interferes: If you set off a bomb next to a metal bar while you heat it, it will not expand in proportion to its change in temperature. Think about such examples for a few moments, and you will see that, outside of fundamental physics, it will be essential to tack on a ‘no interference’ rider to any generalization if it is to have a hope of being true and informative. And that has raised a philosophical problem: How do we make sense of this rider? (See, for example, Earman et al. 2002.) It’s no help simply to observe that special science generalizations are generalizations about localized dependence structure. Yes, but as such, they still need the rider to stand a chance of being true.

I think we make sense of it as follows. If we have a generalization of the schematic form ‘if, at time t, conditions of type C obtain, then conditions of type E will follow at a later time t∗’, we should understand this generalization as making a perfectly definite claim about what would happen, if C-conditions obtained at t, and nothing else were happening at t. For if nothing else were happening, then, a fortiori, nothing else would be happening that could possibly interfere. And so the generalization is true just in case, in all such stripped-down conditions, E-conditions obtain at the appropriate time t∗. Finally, I think we should add that in normal conditions – in which, of course, lots of other stuff is going on – we have a genuine instance of such a law not merely if the C-conditions obtain at some time, and the E-conditions obtain at the appropriate later time, but if there is a process connecting the C-conditions to the E-conditions that shares the relevant features displayed by the process that would have unfolded in stripped down conditions – where, roughly, these ‘relevant features’ are those in virtue of which this process has the counterfactual structure it does. Finally, call a type of process, individuated by such features, a mechanism. Then we can say not merely that the special sciences investigate localized dependence structure, but more specifically, that they aim to discover and articulate the structure of such mechanisms.

6. Conclusion

Let me close with a minor methodological moral. Philosophers who pursue an analysis of causation – especially one that stresses the importance of ordinary intuitions – face a challenge: Why should anyone (except perhaps those
interested in the semantics of the English word ‘cause’) care whether they succeed? The challenge can be met if the analysis yields a concept that has important and interesting connections to other things we care about. The connections between causation, counterfactuals, laws of nature and the taxonomic schemes and generalizations of the special sciences provide a case in point.

Philosophers of science should take very seriously the possibility that philosophical investigation into the nature of causation – of a sort that straightforwardly deserves the name ‘metaphysics’ – can provide them with valuable tools with which to probe and clarify the nature of scientific inquiry. Yes, such investigation should pay attention to what actually goes on in the sciences. But given that metaphysics, quite generally, aims to uncover basic facts about what reality contains, and what that stuff is like, how else should it proceed?

Notes

1 For a more recent expression of this sentiment, see Norton 2003, p. 2.
2 Maybe it has a point, if you want to sort out issues of legal or moral responsibility, and so forth.
3 The qualifier ‘natural’ merely signals a distinction between facts about the physical structure of the world, and facts – if such there be – about mathematics, or logic and so forth. In addition, this picture needs adjustment to accommodate relativistic space-times, for example, as follows: Pick an arbitrary point in space-time. Consider its past light cone. Then the laws dictate how the complete physical state of earlier slices of this past light cone generate complete physical states of successive slices. In what follows, I mostly won’t bother with this refinement.
4 And even with a complete, correct physics in hand, computational limitations would continue to make the use of approximations essential.
5 Note the now standard reading of the counterfactual as ‘non-backtracking’: thus, we do not reason that if \( D \) had not fired, that could only have been because \( C \) did not fire, whence \( E \) would have fired all the same. There is more in Section 4.1 on how to ground this reading in an account of the truth-conditions for counterfactuals.
6 For an example of a regrettably naïve appeal to transitivity along these lines, see Hall 2000.
7 With some obvious and important exceptions – for example, think of Feynman’s study of the Challenger disaster. More generally, it would be a foolish philosopher of science who insisted that investigating singular causal claims could never be of interest to science. It’s just that such investigation is not often the primary focus.
8 For example, the first conception lends itself to scepticism over whether our world is governed by simple physical laws at all; after all, why expect the operations of objects’ causal capacities to be so tidy? So it should come as no surprise that Cartwright, one of the most vigorous and creative contemporary advocates of the first conception, vehemently rejects a conception of physics as aiming to provide us with such laws. (See, especially, her 1999 and, for an earlier and forceful statement
of a closely allied view, Anscombe 1971.) The second conception of causation, by contrast, is the one that seems almost inevitable, if you adopt the sort of ‘fundamentalism’ about physics that Cartwright opposes: the view that fundamental physics should be in the business of looking for mathematically elegant laws, that describe with perfect precision and accuracy the evolution through time of the world’s complete physical state. (Which is not to say that these laws must be deterministic; precision and accuracy can take a probabilistic form, as well. For example, a ‘fundamentalist’ will think it perfectly appropriate to assume that decay phenomena, even if they are irreducibly stochastic, are characterized by perfectly numerically precise objective probabilities.)

9 There are several other ways in which context-dependence can creep in; it’s a useful exercise to try to come up with them.

10 If these fundamental laws are stochastic, then indeterminacy of a special sort will show up. Notice, though, how different it is from the ‘which rock?’ variety.

11 To put it a bit more carefully: for any such little region at time 0, the overwhelmingly vast majority of variations on its time-0 state are such that, if it had that variant, then there would have been no window at time 1.

12 The only obvious dependence is this: if the state of that patch had somehow been such as to interfere with Suzy, then the window would not have broken.

13 Mostly, but not completely. For example, an anti-Humean will be committed to counterfactuals like this one: if there had been just one particle, then it would have been the case that, had there been a second, the two would have interacted in precisely such-and-such a way. The Humean should, by contrast, probably say that if there had been just one particle, then the total history of the world would have been too simple to render it determinate what the laws were – and so there would have been no fact of the matter about what would have happened, had there been a second particle.

14 The obvious example (to philosophers, anyway) is a psychology that tries to discern patterns of dependence between an agent’s actions and the contents of her mental states – assuming, as has become commonplace, that such content is partly determined by how she is connected up to her environment.

15 It will always be possible to write down some equation: all you have to do is index the possible states of R by some numbers, and do the same for S; then you can gerrymander an equation to fit whatever the pattern of dependence between S and R happens to be. It goes without saying that the use of mathematical equations in such cases will hardly be illuminating.

16 Scan Pearl’s (2009) book-length treatment of causal modelling, for example, and you find nothing more substantive than scattered remarks such as the following: ‘The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behavior of a relatively small group of variables’ (p. 223). We are told nothing – anywhere in the book – that might yield an adequate understanding of such concepts as ‘autonomous’, ‘mechanisms’, or – most crucially – ‘constrains’.

17 Making sense of this notion of ‘nothing else happening’ will, I think, involve introducing a distinction between a default state of the world, and deviations from it. A situation in which the only things happening at a certain time consist in the obtaining of such-and-such conditions will just be a situation in which every part of the world save those involved in the instantiation of the conditions is in its default state. See Hitchcock 2007 as well as Hall 2007 for more discussion.
References


Norton, J. (2003), ‘Causation as folk science’, *Philosopher’s Imprint*, 3; http://hdl.handle.net/2027/spo.3521354.0003.004.


Scientific Models and Representation

Gabriele Contessa

1. Introduction

My two daughters would love to go tobogganing down the hill by themselves, but they are just toddlers, and I am an apprehensive parent, so, before letting them do so, I want to ensure that the toboggan won’t go too fast. But how fast will it go? One way to try to answer this question would be to tackle the problem head on. Since my daughters and their toboggan are initially at rest, according to classical mechanics, their final velocity will be determined by the forces they will be subjected to between the moment the toboggan will be released at the top of the hill and the moment it will reach its highest speed. The problem is that, throughout their downhill journey, my daughters and the toboggan will be subjected to an extraordinarily large number of forces – from the gravitational pull of any matter in the relevant past light cone, to the weight of the snowflake that is sitting on the tip of one of my youngest daughter’s hairs – so that any attempt to apply the theory directly to the real-world system in all its complexity seems to be doomed to failure.

A more sensible way to try to tackle the problem would be to use a simplified model of the situation. In this case, I may even be able to use a simple off-the-shelf model from classical mechanics, such as the inclined plane model. In the model, a box sits still at the top of an inclined, frictionless plane, where its potential energy, $U_f$, is equal to $mgh$ (where $m$ is the mass of the box, $g$ is its gravitational acceleration, and $h$ is the height of the plane) and its kinetic energy, $KE_f$, is zero. If we let go of the box, it will slide down the plane and, at the bottom of the slope, all of its initial potential energy will have turned into kinetic energy ($E_f = KE_f + U_f = \frac{1}{2}mv_f^2 + 0 = mgh + 0$) and so its final velocity, $v_f$, will be $(2gh)^{\frac{1}{2}}$. The final velocity of the box, therefore, depends only on its initial height and on the strength of the gravitational field it is in. But what does this tell me about how fast my daughters would go on their toboggan? And why should I believe what the model tells me about the real situation in the first place?
The practice of using scientific models to represent real-world systems for the purpose of predicting, explaining or understanding their behaviour is ubiquitous among natural and social scientists, engineers and policymakers, but until a few decades ago philosophers of science did not take scientific models very seriously. The received view (often also misleadingly labelled as the ‘syntactic view’) was that scientific theories were sets of sentences or propositions, which related to the world by being true or false of it, or at least by having true or false empirical consequences. In this picture, scientific models were taken to play, at most, an ancillary, heuristic role (see, for example, Duhem 1914, p. 117; Carnap 1939, p. 68; Hesse 1963).

As my initial example suggests, however, most real-world systems are way too messy and complicated for us to be able to apply our theories to them directly, and it is only by using models that we can apply the abstract concepts of our theories and the mathematical machinery that often comes with them to real-world systems.¹ In light of these and other considerations, most philosophers of science, today, have abandoned the received picture based on propositions and truth in favour of one of two views in which models play a much more central role. Those who adopt what we could call the model view (or, as it is often misleadingly called, the ‘semantic view’) deny that scientific theories are collections of propositions and prefer to think of them as collections of models.² Those who opt for what we could call the hybrid view, on the other hand, think that models are to a large extent autonomous from theories, but play a crucial mediating role between our theories and the world.³

Despite their differences, there are two crucial points on which the supporters of both views seem to agree. The first point is that scientific models play a central role in science. The second is that scientific models, unlike sentences or propositions and like tables, apples and chairs, are not truth-apt – that is, they are not capable of being true or false. So, whereas according to the received view, scientific theories can relate to the world by being true or false of it (or at least by having true or false consequences), they cannot do so on either the model view or the hybrid view, because, on either view, it is models (and not sentences or propositions) that relate (more or less directly) to the world.

But how do models relate to the world if not by being true or false of it? The most promising and popular answer to this question is that they do so like maps and pictures – by representing aspects or portions of it. As models gained the centre stage in the philosophy of science, a new picture of science emerged (or, perhaps, an old one re-emerged, as van Fraassen (2008, chapter 8) suggests), one according to which science provides us with (more or less faithful) representations of the world as opposed to (true or false) descriptions of it.⁴ In this chapter, I will discuss some of the philosophical questions raised by this representational model view of science.
2. Theoretical vs. Representational Models

There are at least two distinct senses in which scientists and philosophers of science talk of models. In the first sense, ‘model’ can be used to refer to what, more precisely, we could call a model of a theory or a theoretical model – that is, a system of which a certain theory is true. So, for example, the inclined plane model, which I used to represent my daughters on the toboggan, is a model of classical mechanics, in the sense that classical mechanics is true of the model.

In the second sense, ‘model’ can be used to refer to what we could call a model of a system or a representational model – that is, an object used to represent some system for the purpose of, for example, predicting or explaining certain aspects of the system’s behaviour. In my initial example, for instance, I used the inclined plane as a model of my daughters tobogganing down the hill.

These two notions of scientific model are easily conflated because, as the example illustrates, we often use theoretical models as representational models. However, whereas it would seem that any theoretical model can be used as a representational model, not all representational models need to be theoretical models. To represent my daughters tobogganing down the hill, for example, instead of the inclined plane model, I could have used an ordinary hockey puck sliding down an icy ramp. Alternatively, I could have gathered data about the final velocities of other toboggans going down the hill as well as other variables (such as the mass and cross-sectional areas of their passengers) and found an equation to fit the data and used that equation to predict the final velocity of the toboggan with my daughters in it.

As Hughes puts it, ‘perhaps the only characteristic that all [representational] models have in common is that they provide representations of parts of the world’ (Hughes 1997, p. S325). From an ontological point of view, for example, my three examples are a mixed bag. The puck is what we could call a material model, because unlike the inclined plane, it is an actual concrete physical system made up of actual concrete physical objects, just like the system it is meant to represent. The mathematical equation is what we could call a mathematical model: an abstract mathematical object that is used to represent (directly) a concrete system. Finally, the inclined plane model is what we could call a fictional model, because the objects in it, like fictional characters such as Sherlock Holmes, are not actual concrete objects, but are said to have qualitative properties (like having a mass or smoking a pipe).

Although there are a variety of devices that can be used as representational models, theoretical models still constitute the main stock from which representational models of real-world systems (or at least the building blocks for such models) are drawn. The following is a simple variation on a popular theme. Theoretical principles (the ‘laws’ of our theories) do not describe the world. They merely define certain classes of models – the theoretical models.
Theoretical models can be combined, specified or modified to be used as models of some real-world system. Such representational models can be used to represent either a specific real-world system directly (as in my initial example), or a type of real-world system (e.g. the hydrogen atom as opposed to some specific hydrogen atom), or a data model: a ‘smoothed out’ representation of data gathered from a certain (type of) system.

No matter how many layers one adds to this picture, the models at the bottom of this layer cake (be they data models or, as I will mostly assume here, representational models) would seem to have to represent directly some aspect or portion of the world if the gap between the theory and the world is to be bridged. It is to the question of how these ‘bottom’ models do so that I will now turn.

3. Disentangling ‘Representation’

You are visiting London for the first time and you need to reach Liverpool Street station. You enter the nearest tube station, pick up a map of the London Underground and, after looking at it, you quickly figure out that you have to take an eastbound Central Line train and get off after three stops. What you have just performed seemingly so effortlessly is what, following Swoyer (1991), I shall call a piece of surrogative reasoning. The London Underground map and the London Underground network are clearly two distinct objects. One is a piece of glossy paper on which coloured lines, small black circles and names are printed; the other is an intricate system of, among other things, trains, tunnels, rails and platforms. Yet you have just used one of them (the map) to infer something about the other (the network). More precisely, from ‘The circles labelled “Liverpool Street” and “Holborn” are connected by a red line’ (which expresses a proposition about the map) users infer ‘Central Line trains operate between Holborn and Liverpool Street stations’ (which expresses a proposition about the network).

The fact that a user performs a piece of surrogative reasoning from something (a vehicle) to something else (a target) is the main ‘symptom’ of the fact that the vehicle is being used as an epistemic representation of the target by that user. So, if you use the map you are holding in your hand to perform a piece of surrogative reasoning about the London Underground network, it is because, for you (as well as for the vast majority of users of the London Underground network), that map is an epistemic representation of the network. Analogously, if in my initial example I used the inclined plane model to infer how fast the toboggan would go, it was because I was using it as an epistemic representation of my daughters tobogganing down the hill.

In order for a vehicle to be an epistemic representation of a target, the conclusions of the surrogative inferences one draws from one to the other...
do not need to be true. For example, if you were to use an old 1930s map of the London Underground today, you would infer that Liverpool Street is the last stop on the Central Line, which is no longer the case. The difference between the old and the new map is not that one is an epistemic representation of today’s network while the other is not, but that one is a completely faithful epistemic representation of it (or at least so we can assume here) while the other is only a partially faithful one – some, but not all, surrogative inferences from the old map to the network are sound.9

Two things are worth noting. First, a vehicle can, at the same time, represent its target and misrepresent (some aspects of) it, for representation does not require faithfulness. This is particularly important for scientific models, which are rarely, if ever, completely faithful representations of their targets. Overall, the inclined plane model, for example, is not a very faithful representation of my daughters tobogganing down the hill. Nevertheless, it may be sufficiently faithful for my purposes. Second, unlike representation, faithfulness comes in degrees. A vehicle can be a more or less faithful representation of a certain target, but it is either a representation of a certain target (for its users) or it is not.

Once we distinguish between epistemic representation and faithful epistemic representation, it becomes clear that there are two questions a philosophical account of epistemic representation should answer. The first is: What makes a vehicle an epistemic representation of a certain target? The second is: What makes a vehicle a more or less faithful epistemic representation of a certain target? In the literature, these two questions have been often conflated under the heading ‘the problem of scientific representation’. This label, however, may in many ways be misleading. One way in which it can be misleading is that it suggests that there is a single problem that all contributors to the literature are trying to solve, while there are (at least) two.10 A better way to describe the situation is that some of the supposedly rival solutions to ‘the problem of scientific representation’ are in fact attempts to answer different questions. On this interpretation, one can find at least three rival accounts of epistemic representation, that is, three different answers to the question ‘What makes a vehicle an epistemic representation of a certain target?’ (the denotational account, the inferential account and the interpretational account) and two (somewhat related) accounts of faithful epistemic representation (the similarity account and the structural account). I will now sketch these accounts in turn, starting with accounts of epistemic representation.

3.1 Epistemic Representation

The denotational account suggests that all there is to epistemic representation is denotation.11 More precisely, according to the denotational account, a vehicle is an epistemic representation of a certain target for a certain user if and only
if the user stipulates that the vehicle denotes the target. The prototype of
denotation is the relation that holds between a name and its bearer. So, for
example, ‘Plato’ denotes Plato, but, had different stipulations been in place,
‘Plato’ could have denoted Socrates, and Plato could have been denoted by
‘Aristotle’. So, according to this view, if the London Underground map is an
epistemic representation of the London Underground network for you, or the
inclined plane model is an epistemic representation of my daughters tobog-
ganning down the hill for me, it is because, respectively, you and I have stipu-
lated that they are. You could equally well have chosen to use an elephant and
I a ripe tomato.

Whereas denotation seems to be a necessary condition for epistemic
representation, it does not, however, seem to be a sufficient condition. Nobody
doubts that you could have used an elephant to denote the London Under-
ground network, but it is not clear how you could have used an elephant to
perform surrogative inferences about the network. A user’s ability to perform
surrogative inferences from a vehicle to a target seems to be the main symp-
tom that she is using the vehicle as an epistemic representation of the target.
So, it would seem that other conditions need to be in place for a case of mere
denotation to turn into one of epistemic representation. But what are these
further conditions?

According to the inferential account of epistemic representation (see mainly
Suárez 2004), the answer is simply to explicitly add as a further necessary con-
dition for epistemic representation that the user be able to perform surrogative
inferences from the vehicle to the target. So, according to the inferential account,
a vehicle is an epistemic representation of a certain target for a certain user
(if and) only if (a) the user takes the vehicle to denote the target and (b) the
user is able to perform surrogative inferences from the vehicle to the target.

The inferential account thus avoids the problem that faced the denotational
account, but it does so in a somewhat ad hoc and ultimately unsatisfactory
manner. In particular, the inferential account seems to turn the relation
between epistemic representation and surrogative reasoning upside down. The
inferential account seems to suggest that the London Underground map
represents the network (for you) in virtue of the fact you can perform surroga-
tive inferences from it to the network. However, the reverse would seem to be
the case — you can perform surrogative inferences from the map to the net-
work in virtue of the fact that the map is an epistemic representation of the
network (for you). If you did not take this piece of glossy paper to be an
epistemic representation of the London Underground network in the first
place, you would never try to use it to perform surrogative inferences about
the network.

More seriously, the inferential account seems to suggest that the users’
ability to perform surrogative inferences from a vehicle to a target is somehow
basic and cannot be further explained in terms of the obtaining of deeper conditions, thus making surrogative reasoning and its relation to epistemic representation unnecessarily mysterious. Ideally, an account of epistemic representation should explain what makes a certain vehicle into an epistemic representation of a certain target for a certain user and how in doing so it enables the user to use the vehicle to perform surrogative inferences about the target. This is what the last account of epistemic representation I will consider here attempts to do.  

According to the interpretational account of epistemic representation, a vehicle is an epistemic representation of a certain target for a certain user if and only if (a) the user takes the vehicle to denote the target and (b*) the user adopts an interpretation of the vehicle in terms of the target (see Contessa 2007, Contessa forthcoming, and, possibly, Hughes 1997). So, the interpretational conception agrees with both the denotational and the inferential conception in taking denotation to be a necessary condition for epistemic representation, but it takes the adoption of an interpretation of the vehicle in terms of the target to be what turns a case of mere denotation into one of epistemic representation.

So, for example, a ripe tomato could be used as easily as the inclined plane model to denote the system formed by my daughters tobogganing down the hill, but it is not clear how I could use the ripe tomato to infer how fast the toboggan would go. In the case of the inclined plane model, on the other hand, there is a clear and standard way to interpret the model in terms of the system. In fact, such an interpretation is so obvious that it would seem to be almost superfluous to spell it out if it was not to illustrate what an interpretation of a vehicle in terms of a target is: the box in the model denotes the toboggan with my daughters in it; the mass, the velocity and the acceleration of the box denote the mass, the velocity and the acceleration of the toboggan; the inclined plane denotes the slope of the hill; and so on.

The main advantage of the interpretational conception is that it offers a clear account of the relation between epistemic representation and surrogative reasoning. The user adopts an interpretation of a vehicle in terms of a target (and takes the vehicle to denote the target) and this interpretation provides the user with a set of systematic rules to ‘translate’ facts about the vehicle into (putative) facts about the target. If the final velocity of the box is \( v_f \) and, if according to the interpretation of the model I adopt, the box denotes my daughters on the toboggan and the velocity of the box denotes the velocity of the toboggan, then, on the basis of that interpretation I can infer that the final velocity of the toboggan is \( v_f \). (Note, however, that this does not mean that I need to believe that the final velocity of the toboggan is going to be \( v_f \).)

It is tempting to think that, on the interpretational account, epistemic representation comes too cheaply. After all, nothing seems to prevent me from
adopting an interpretation of a ripe tomato in terms of the system formed by
my daughters tobogganing down the hill, one according to which, say, the
deeper shade of red the tomato is, the faster the toboggan will go. This may
well be true, but what exactly would be wrong with that? The objection might
be that from the tomato I would likely infer only false conclusions about the
system. This may well be the case, but the interpretational account is meant to
be an account of what makes a vehicle an epistemic representation of a certain
target, not an account of what makes it a faithful epistemic representation of the
target. Further conditions would need to be in place for the tomato to be a faithful
epistemic representation of the system, and it is plausible to assume that the
tomato (under this interpretation, at least) could not meet these conditions.

Maybe the objection is that using models for prediction is not all that different from using tarots, if all there is to epistemic representation is denotation and interpretation. Of course, there is an enormous difference between using models and using tarots to find out whether my daughters will be safe on the
toboggan, but the difference need not be that one, but not the other, is an epistemic
representation of the situation (after all, tarots are used to perform surrogative
inferences about other things); the difference could rather be that one provides a
much more faithful representation of the situation than the other.

If epistemic representation appears to come cheap, on the interpretational
account, it may be because epistemic representation is cheap. It doesn’t take
much for someone to be able to perform surrogative inferences from something to something else (but, at the same time, it does not take as little as the
denotational and inferential conceptions suggest). What does not come cheap are faithful epistemic representations, and even less cheap are epistemic
representations that we have good reasons to believe are sufficiently faithful
for our purposes. So it is to accounts of faithful epistemic representation that
I turn in the next section.

3.2 Faithful Epistemic Representation

Assume that the conditions obtain for a certain vehicle to be an epistemic
representation of a certain target for a certain user. What further conditions
need to be in place in order for the epistemic representation to be a faithful
one? According to the similarity account of faithful epistemic representation,
the further condition is that the vehicle is similar to the target in certain
respects and to certain degrees (where what counts as the relevant respects
and degrees of similarity largely depends on the specific purposes of the user;
see Giere 1985, Giere 1988, Teller 2001, Giere 2004). For example, in the case of
the toboggan going down the hill, what I am interested in is that the toboggan
will not go too fast. So, in order for the inclined plane model to be a (sufficiently) faithful representation of the system for my purposes, it must at least
be the case that the final velocity of the box in the model is sufficiently similar to the highest speed the toboggan will reach. But how similar is sufficiently similar? In this case, it would seem that the most important aspect of similarity concerns the highest speed of the box, on the one hand, and of the toboggan, on the other. The speed the toboggan will actually reach should not be (much) higher than the one reached by the box in the model, for if the speed of the toboggan were to be much higher than that of the box, I might inadvertently expose my daughters to an unnecessary risk.

This, however, still seems to be excessively permissive. After all, I might happen to employ a model that, on this particular occasion, happens to predict the highest speed of the toboggan accurately, but does so in an entirely fortuitous manner (say, a model based on some wacky theory according to which the speed of the toboggan depends on its colour). Would such an accidental similarity be sufficient to make the model into a faithful epistemic representation of the system for my purposes? This question, it would seem, should be answered negatively (for reasons analogous to the ones that make us deny that cases of epistemic luck constitute cases of knowledge). If accidental similarity was sufficient for faithfulness, then even tarots would sometimes be faithful epistemic representations of their targets. If faithfulness is a matter of similarity, then, to avoid accidental similarities, it would seem that the similarity between a vehicle and a target would need to be somewhat more systematic than the one between the fortuitously successful model and the system. For example, it seems natural to require that the model not only gives us an appropriate answer one-off, but that it does so reliably and under a range of counterfactual circumstances. However, it is not clear whether the unadorned similarity account has the resources to explain precisely what these more systematic similarities might be.

The structural account of faithful epistemic representation can be seen as trying to capture this more abstract and systematic sense in which a vehicle needs to be similar to a target in order for the former to be a somewhat faithful epistemic representation of the target. This account tries to avoid the problems that beset the similarity account, while retaining the basic insight that underlies it – that faithfulness is a matter of similarity. In fact, the structural conception could be considered a version of the similarity conception (for what would a structural relation be if not an abstract similarity (viz. a similarity of structure)?)?

According to the structural account of faithful epistemic representation, if a certain vehicle is an epistemic representation of a certain target for a certain user, then, if some specific morphism holds between the structure of the vehicle and the structure of the target, then the model is a faithful epistemic representation of that target. (A morphism is a function from the domain of one
structure to the domain of the other that preserves some of the properties, relations and functions defined over the respective domains).\textsuperscript{16}

The first problem a structural account of faithful epistemic representation encounters is that morphisms are relations between set-theoretic structures, and most vehicles and targets are not set-theoretic structures (see, for example, Frigg 2006). For example, neither my daughters tobogganning down the hill nor the London Underground network would seem to be set-theoretic structures. The most promising solution to this problem would seem to claim that, whereas most vehicles and targets are not structures, nevertheless they can instantiate structures.\textsuperscript{17}

Assuming a viable account of structure instantiation is available, the structural account roughly maintains that a vehicle is a faithful epistemic representation of the target only if some specific morphism holds between the structure instantiated by the vehicle and the one instantiated by the target. But which morphism? In the case of what I have called completely faithful representations, the morphism is arguably an isomorphism (or, more precisely, an ‘intended’ isomorphism). So, for example, the structure instantiated by the new London Underground map is isomorphic to the one instantiated by the London Underground network. (Among other things, this means that, for example, every circle on the map can be put into one-to-one correspondence with a station on the network, in such a way that circles that are connected by a line of the same colour correspond to stations that are connected by the same subway line).

The problem, however, is that most epistemic representations, and especially most scientific models, are not completely faithful representations of their targets. Supporters of the structural account generally agree that the solution to this problem is to opt for a morphism that is weaker than isomorphism, such as homomorphism (see, for example, Bartels 2006) or partial isomorphism (see, for example, French and Ladyman 1999). However, it is no easy feat to identify a type of morphism that is weak enough to capture all epistemic representations that are at least partially faithful (no matter how unfaithful) while leaving out all the ones that are completely unfaithful.

A more serious problem is that faithful epistemic representation is a gradable notion, but morphism is not. So, whatever type of morphism one opts for (no matter how weak), whether or not that morphism holds between the structure instantiated by the vehicle and that instantiated by the target is a yes-or-no question. How faithfully the vehicle represents the target, on the other hand, is a matter of degree.

These two problems may be solved in one fell swoop by denying that a structural account of faithful epistemic representation needs to identify a single morphism, one that is weak enough to allow for epistemic representations that are not completely faithful, while leaving out the ones that are completely
unfaithful. One could, instead, maintain that faithful epistemic representation is a matter of structural similarity, and that the more structurally similar the vehicle and the target are (i.e. the stronger the morphism between the structures instantiated by the vehicle and the target is), the more faithful an epistemic representation of the target the vehicle is (see Contessa forthcoming, chapter 4).

If, as I suggested, the structural account is nothing but a version of the similarity account, at present the similarity account of faithful epistemic representation would not seem to have any genuine rivals and the structural account of representation seems to be the most promising way for the advocates of the similarity account to avoid the charge of vacuity. More work, however, is needed to show that the structural account can be developed into a full-fledged account of faithful epistemic representation.

4. Models and Idealization

So far, I have mostly been concerned with the semantics of epistemic representation – what makes a vehicle an epistemic representation, or a faithful epistemic representation, of a certain target? In this section, I will focus on the pragmatics of epistemic representation, and in particular on what philosophers call ‘idealization’.

A recurring theme in the literature on scientific models is that successful representation often involves a great deal of misrepresentation. In the terminology used here, scientific models are rarely (if ever) completely faithful epistemic representations of their target systems. Idealization (which consists in the deliberate misrepresentation of certain aspects of a system) is often the key to successful application of the conceptual and mathematical resources of our theories.

The inclined plane model, for example, constitutes a highly idealized epistemic representation of my daughters going down the hill on their toboggan. Just consider one aspect in which the model is idealized. On their way down the hill, my daughters and their toboggan would be subject to an extraordinarily large number of forces (from air friction to the gravitational pull of the Sun). However, in the model, only two forces are acting on the box – a gravitational force (presumably exerted by a uniform gravitational field) and a normal force (exerted by the plane).

One may think that such a crude representation of a system is unsuitable for the purpose of successfully predicting (or explaining) the behaviour of that system. But according to what, following McMullin (1985), we could call ‘the Galilean account of idealization’, faithfulness and successfulness need not go hand in hand: a less faithful model of a certain system can be predictively or
Scientific Models and Representation

explanatorily as successful as (if not more predictively or explanatorily successful than) a more faithful one. How is this possible? First of all, it is important to note that the fact that overall the inclined plane is not a very faithful epistemic representation of my daughters on the toboggan does not mean that it is not sufficiently faithful for my purposes. After all, all I am interested in is that the toboggan would not reach an unsafe speed, and the model is doing its job insofar as it gives me good reasons to think that the toboggan won’t go too fast.

Once we draw this distinction, it becomes apparent that a model that is overall a more faithful representation of a system than another is not necessarily a more faithful representation for one’s purposes. A model that would take into account, say, the gravitational pull of the Sun on my daughters and their toboggan may well be a more faithful representation of the situation overall, but would not seem to be any better at predicting whether the toboggan will exceed a safe speed, as the contribution that that force makes to the speed of the toboggan is negligible compared to some other forces.

Some of the forces that have no counterpart in the model do have a non-negligible effect on the final speed, however. Frictional forces, for example, would seem to dramatically affect the speed the toboggan will reach. However, since these forces would only contribute to slowing down the toboggan, and since including them would result in a more complicated model, if all I am interested in is that the my daughters have a safe ride, I might as well use the simpler model.

If I don’t want to deprive my daughters of a potentially fun and safe ride, however, the inclined plane model may no longer be sufficiently faithful for my purposes, and I might have to abandon it in favour of one of the more complicated models that take into account frictional forces. If the slope is sufficiently smooth and icy, surface friction may be negligible compared to air friction. However, air friction on an object consists in the collision of billions of air particles with the object and is therefore a bewilderingly complicated phenomenon to model. The standard solution is to introduce a force that approximates the net effect of these collisions. For example, air friction on our box can be represented as a force of magnitude \(-1/2 \cdot C \cdot \rho \cdot A \cdot v^2\) (where \(C\) is a drag coefficient, \(\rho\) is the air density and \(A\) is the cross-sectional area of the box). In the new model, the box would reach a terminal velocity \((2mg \sin \theta)/(CpA)^{1/2}\).

As even this simple example illustrates, successful modelling depends on subtle interplay among different factors, including (a) the aspect of the behaviour of the system the modeller is interested in predicting or explaining, (b) the features of the system that most significantly contribute to the production of that behaviour, and (c) practical and principled limits to the representation of those factors. According to the Galilean account of idealization, the reason why a model need not be faithful in order to be (predictively
or explanatorily) successful is that usually not all aspects of a system are equally relevant to the production of the specific behaviour we are interested in explaining or predicting. Those aspects that do not significantly contribute to the relevant behaviour can thus be safely distorted or ignored by our models (at least insofar as introducing them in the model would not significantly alter that behaviour).

In some cases, only a handful of factors that can easily be modelled make a significant contribution to the behaviour we happen to be interested in. In many cases, however, Nature is less cooperative, and a certain amount of ingenuity and ad hoc manoeuvring is needed to get our models to work. In some cases, it even seems that idealizations are altogether essential to the success of models, in the sense that a de-idealized model (that correctly represents the previously idealized features of the system) would not be able to explain or predict the relevant behaviour. Understanding different kinds of idealizations, and their pragmatic virtues and (in some cases) indispensability, is an outstanding issue in philosophy of science.

5. Models and Realism

While the representational model view of the relation between theories and the world has widely supplanted the linguistic, descriptive one, it is still not completely clear what consequences, if any, this ‘representational turn’ has for some of the classic debates in philosophy of science. One of the most interesting examples is perhaps that regarding scientific realism, which has been traditionally framed in terms of theories and truth. There seem to be at least three views on how the representational turn has affected this debate. The first is that the representational turn favours scientific realists, by helping them to sidestep some of the notorious difficulties related to semantic notions such as reference and approximate truth (see, for example, Giere 1985 and Suppe 1989). The second view is that the representational turn leaves the terms of the scientific realism debate by and large unchanged (see, in particular, Chakravartty 2001 and Chakravartty 2007, chapter 7). The third is that the representational turn favours scientific anti-realism.

In the literature, one can find a host of arguments that suggest that, on a representational picture, the notion of empirical success of our theories is either too strong or too weak to sustain the scientific realist’s traditional arguments from success to truth. If one adopts a weak notion of empirical success, then, since models from incompatible theories are often successful at representing different aspects or parts of the same real-world system, it would seem that the inference from success to truth would leave us with many true, but mutually incompatible, theories (see, for example, Morrison 2000,
chapter 2, and Rueger 2005). If one adopts a strong notion of empirical success, on the other hand, since even the models of our best theories are successful only within relatively small pockets of the world, it would seem that we are only justified in believing that theories are true of those aspects or portions of the world they are successful at modelling (see Cartwright 1999, especially chapter 2). These kinds of considerations have motivated some to shift towards the anti-realist end of the spectrum. It is still a matter of debate whether the representational turn warrants such a shift.

Whatever the case may be, the consequences of the representational turn for the scientific realism debate as well as other traditional issues in philosophy of science are still mostly uncharted territory, and the exploration and mapping of such territory seems to be one of the most interesting and potentially fruitful projects for the near future. For the first time since the decline and fall of logical empiricism and the so-called ‘syntactic’ view of theories and the post-Kuhnian ‘Balkanization’ of philosophy of science, the representational turn may provide philosophers of science with a unified framework within which to work. While many important details still need to fall into place, the rough outline of the picture is already clear and seems to be far richer and more realistic than the one that preceded it. Much work still needs to be done to turn this rough outline into a detailed picture, but once most of the details have been worked out, this new framework will hopefully shed new light on old issues and reveal overlooked connections among them.

**Acknowledgements**

I would like to thank Steven French, Chris Pincock, and Juha Saatsi for their helpful comments on previous versions of this essay.

**Notes**

1 This point has been made most forcefully by Cartwright (see, in particular, Cartwright 1983, 1999).

2 The so-called semantic view originated with the work of Suppes in the 1960s (see, for example, Suppes 1960, but also Suppes 2002) and can be safely considered the new received view of theories counting some of the most prominent philosophers of science among its supporters (see, for example, van Fraassen 1980, Giere 1988, Suppe 1989, da Costa and French 1990). How exactly the so-called semantic view of theories relates to the view that theories are collections of models is an exegetical question that is beyond the scope of this essay.

3 This view, sometimes referred to as the models-as-mediators view, is developed and defended, for example, by many of the contributors to Morgan and Morrison (1999). The most developed defence of this view is perhaps to be found in Cartwright (1999).
4 How robust the distinction is between representing and describing obviously depends partly on one's views on language and truth. 

5 Cartwright (1983) was one of the first to suggest the analogy between models and fictions. See Godfrey-Smith (2009), Contessa (2010), Frigg (2010) and Suárez (2009) for different takes on this analogy. 

6 For a much more refined and detailed variation on the same theme, see, for example, Giere (1988, chapter 4). 

7 Note that to say that a representation is an epistemic representation is just to say that it is a representation that is used for epistemic purposes (i.e. a representation that is used to learn something about its target). So, for example, Pablo Picasso's *Portrait of Daniel-Henry Kahnweiler* represents the art dealer Daniel-Henry Kahnweiler. However, the main purpose of that portrait, presumably, was not that of being an epistemic representation of Kahnweiler (but rather what we could call an aesthetic representation of him). 

8 Here I take surrogative reasoning to be 'a symptom' of epistemic representation. However, there may well be 'asymptomatic' cases of epistemic representation – cases in which users perform no actual surrogative inference from the vehicle to the target, although they would be able to do so if they wanted to. 

9 The crucial distinction between what I call 'epistemic representation' and 'faithful epistemic representation' was first emphasized in this context by Suárez (2004). Another way in which the label 'scientific representation' could be misleading is that it seems to imply that something sets aside scientific representations from epistemic representations that are not scientific. See Section 2.5. 

10 If I understand them correctly, Callender and Cohen (2006) defend a version of what I call the denotational account. 

11 Suaréz seems to think that (a) and (b) are necessary but not sufficient conditions for what I call epistemic representation (see Suaréz (2004)). However, more recently, Suaréz and Solé (2006) seems to suggest that (a) and (b) may be jointly sufficient. 

12 Admittedly, the inferential account is meant to provide us with a deflationary or minimalist account of (epistemic) representation. However, it is not clear why one would opt for such a deflationary or minimalist account unless no more substantial account were available. 

13 It might be worth explaining why, according to the interpretational account, (a) is still needed even if (b*) obtains. Suppose that you find a map of a subway system that does not tell you which subway system (if any) it represents. Since most subway maps are designed on the basis of the same general interpretation, you would still be able to make a number of inferences about the subway network the map represents, even if you do not know what system it represents. According to the interpretational account, this is because, in this case, condition (b*) seems to hold, but condition (a) does not. 

14 It may also be important that the speed of the toboggan is not going to be much lower than that reached by the box (if the maximum speed of the toboggan were to be much lower than the one of the box, I might prevent my daughters from enjoying a fun and safe ride). 


16 For an account of structure instantiation, see Contessa (forthcoming, chapter 4), which draws on Frigg (2006) and Cartwright (1999).
See, for example, Cartwright (1983) and McMullin (1985). For a slightly different take on idealization, see Streves (2008, chapter 8), Jones (2005) and Weisberg (2007b).

For a few examples of ‘essential’ idealization, see Batterman (2001, 2010).

See, for example, Cartwright’s ‘patchwork’ realism (Cartwright 1999), Giere’s scientific perspectivism (Giere 2006) and van Fraassen’s empiricist structuralism (van Fraassen 2008), all of which are partly motivated by the representational turn.

References


French, S. (2003), ‘A model-theoretic account of representation (or I don’t know much about art . . . But I know it involves isomorphism)’, Philosophy of Science, 70, 1472–83.


—(2004), ‘How models are used to represent reality’, Philosophy of Science, 71, 742–52.

—(2006), Scientific Perspectivism. Chicago, IL University of Chicago Press.


—(2007), ‘Where are all the theories gone?’, Philosophy of Science, 74, 195–228.


—(2007b), ‘Who is a modeler?’, *British Journal for Philosophy of Science*, 58, 207–33.
1. Introduction

The idea of reduction has appeared in different forms throughout the history of science and philosophy. Thales took water to be the fundamental principle of all things; Leucippus and Democritus argued that everything is composed of small, indivisible atoms; Galileo and Newton tried to explain all motion with a few basic laws; seventeenth century mechanism conceived of everything in terms of the motions and collisions of particles of matter; British Empiricism held that all knowledge is, at root, experiential knowledge; current physicists are searching for a GUT, a ‘grand unified theory’, that will show that at very high energies the electromagnetic and the weak and strong nuclear forces are manifestations of a single unified field. Some of these projects are clearly ontological in nature (Leucippus and Democritus), others are more methodological (mechanism), and still others strive for theoretical simplification (the projects of Galileo and Newton or the search for a GUT). Nevertheless, as they all aim at revealing some kind of unity or simplicity behind the appearance of plurality or complexity, they may all be regarded as (attempted) reductions.

Section 2 surveys philosophical accounts of reduction, focusing mostly on theory reduction, but taking into account ontological aspects of reduction as well. Section 3 addresses the question whether (and if so, how) the special sciences are reducible to more fundamental sciences, in particular in the light of the fact that special science properties seem to be multiply realizable. Section 4 looks at some attempts to understand reductive endeavours in terms of mechanistic explanations. Section 5 explores interconnections between scientific reductions and the idea that our world is a layered one with distinguishable levels of organization. Finally, section 6 briefly highlights some worthwhile future research questions.

Space limitations prevent us from addressing some issues pertinent to the topic, like emergence (Bedau and Humphreys 2008; Stephan 2006), properties and powers (Molnar 2003), or laws (Carroll 1994; Mumford 2004). Although we
Reduction, Multiple Realizability and Levels of Reality

will mostly be concerned with discussions in the philosophy of mind having to do with the reduction of psychology to neuroscience, the issues we raise are rather general and arise in other disciplines as well, including physics (Batterman 2000, 2001), biology (Brigandt 2010; Brigandt and Love 2008; Schaffner 1993; Wimsatt 2007), chemistry (Hendry forthcoming) and the social sciences (Jackson & Pettit 1992).¹

2. Theory Reduction

2.1 Reduction as Translation

In the early twentieth century, logical positivists set out to understand the nature of science and the relations between the various scientific disciplines. One of their goals was to ‘unify science’ by finding a common language into which all meaningful scientific statements are translatable. Rudolf Carnap (1932a, 1932b) and Carl Gustav Hempel (1949) argued that the language of physics could serve as the universal language of science. Meaningful scientific concepts, statements and laws, they held, must be translatable into physical concepts, statements and laws. Psychology, for example, is ‘an integral part of physics’ in that ‘[a]ll psychological statements which are meaningful . . . are translatable into statements which do not involve psychological concepts, but only the concepts of physics’ (Hempel 1949, p. 18). For example, the psychological predicate ‘x is excited’, Carnap (1932b, pp. 170–1) argued, is translatable into a physical predicate like ‘x’s body . . . has a physical structure that is characterized by a high pulse and rate of breathing, by vehement and factually unsatisfactory answers to questions, by the occurrence of agitated movements on the application of certain stimuli etc.’ Carnap defended this claim by pointing out, first, that verificationism entails that predicates are synonymous if and only if they are applied on the basis of the same observations and, secondly, that the physical predicate simply enumerates the observations on the basis of which the psychological predicate is applied. Since Carnap took properties to be the intensions of predicates, he thought that properties were identical if and only if the corresponding predicates were synonymous and thus that the translatability claim also vindicated an ontological reduction of the property of being excited.

The hope that physics could serve as a lingua franca of science was soon dashed, however, because many prima facie meaningful statements of the special sciences, including those of psychology, were simply not translatable into physical language in a non-circular way: notoriously, someone who wants a beer will go to the fridge to get one only if she believes that there is beer in the fridge, does not attempt to stay sober and so forth. Synonymy thus seemed to be too strong a requirement on both the theoretical reduction of psychology to
physics and the ontological reduction of the mental to the physical. In the philosophy of science, translational reduction was therefore replaced by more sophisticated models of reduction (see Sections 2.2–2.4), while in the philosophy of mind, synonymy was abandoned as a prerequisite for property-identities, paving the way for the idea that psychophysical property-identities are what Kripke (1980) called ‘a posteriori necessities’ (see Section 3).

2.2 Oppenheim and Putnam: The Unity of Science
Although the dream of a wholesale translation of all scientific statements into the language of physics had to be given up, the ideal of a unified science in which special sciences like chemistry, biology, psychology and so forth are reducible to more fundamental theories was retained. Could something like a ‘unity of science’ not be attained even if higher-level predicates like ‘x is soluble’, ‘x is a Chinese wisteria’, or ‘x is excited’ are not translatable into purely physical terminology? Paul Oppenheim and Hilary Putnam (1958) suggested, as a working hypothesis, the view that all sciences are reducible to physics via a series of microreductions. Theory $T_2$ microreduces to theory $T_1$ if and only if (a) any observational data explainable by $T_2$ are explainable by $T_1$, (b) $T_1$ has more ‘systematic power’ than $T_2$, and (c) all the entities referred to in $T_2$ are fully decomposable into entities belonging to the universe of discourse of $T_1$. Oppenheim and Putnam’s approach faced severe difficulties. For example, it is unclear whether the observational and the non-observational can always be clearly distinguished; the notion of ‘systematic power’ is not clearly defined; and there are hardly any historical cases that satisfy the proposed conditions (Sklar 1967). Nevertheless, many of its key ideas are still visible in what became the standard model of intertheoretic reduction for decades to come: Ernest Nagel’s (1961) model of reductions as derivations via bridge-laws.

2.3 Nagel: Reductions as Derivations via Bridge-laws
Nagel (especially 1961, pp. 336–97) took the idea that reduction consists in the derivation of the reduced theory $T_2$ from a reducing theory $T_1$ seriously. Such derivations are possible, Nagel (1961, pp. 352–6) argued, if (a) the terms of $T_2$ are connectable with the terms of $T_1$ by means of suitable bridge-laws, that is, empirical hypotheses that express material rather than logical connections (the ‘condition of connectability’), and (b) given these connecting principles, all laws of $T_2$ can be derived from laws of $T_1$ (the ‘condition of derivability’). Reductions could thus be seen as deductive-nomological explanations, where $T_1$ explains $T_2$. Since in all interesting cases of reduction, $T_1$ and $T_2$ are going to be framed in partially disjoint vocabularies, the connectibility condition is essential: without connecting bridge-laws the required derivations would be impossible. The exact nature
of bridge-laws has been a matter of debate. Although Nagel allowed them to be material conditionals of the form $\forall x (F_{T_1}x \supset F_{T_2}x)$ (Richardson 1979), it was usually assumed that biconditionals of the form $\forall x (F_{T_1}x \equiv F_{T_2}x)$ are necessary for the ontological simplifications that were considered to be one of the main goals of reduction (see Section 3).

Unlike Oppenheim and Putnam’s approach, Nagel’s model of reduction was formally precise, but it also failed to fit standard cases of scientific reduction. Nagel himself acknowledged that even in his own example, the reduction of thermodynamics to statistical mechanics, the actual derivation would be immensely complicated and possible only under a set of idealizing assumptions (one has to assume, for example, that the gas is composed of a large number of perfectly elastic spherical molecules with equal masses and volumes that are in constant motion and subject only to forces of impact between themselves and the walls of the container). In fact, the derivation may not be possible at all (Richardson 2007; Sklar 1999), given that central thermodynamic concepts like ‘entropy’ are associated with a variety of distinct concepts in statistical mechanics which do not exactly correspond to thermodynamic entropy, neither separately nor taken together.

Another important problem for Nagel’s account was that the reducing theory often corrects the original theory, which entails that the original theory is false. For example, Newtonian physics showed that some principles of Galilean physics, like the assumption that uniformly accelerated gravitational free fall is the fundamental law of motion, were false. However, since logical deduction is truth-preserving, the new, reducing theory cannot both be true and logically entail a false theory. Problems like these led Paul Feyerabend (1962) to argue that no formal accounts of scientific reduction are possible or necessary. The majority of philosophers, however, responded by developing more sophisticated models (Causey 1977; Schaffner 1967), culminating in what became known as, using John Bickle’s (1998) term, ‘New Wave Reductionism’ (NWR).

2.4 New Wave Reductionism
Like its precursor, NWR is an allegedly universal model that takes reduction to be a relation involving logical derivations between theories (Bickle 1998, 2003; Hooker 1981; see also Churchland 1985; Churchland 1986; Schaffner 1993). However, what is derived from $T_1$ is not $T_2$ itself, but an ‘equipotent isomorphic image’ $T_{2a}$ of $T_2$, which renders the falsity of $T_2$ (see Section 2.3) unproblematic. The ultimate fate of $T_2$ and its ontological posits depends upon the exact relation between $T_2$ and $T_{2a}$. If the analogy between $T_2$ and $T_{2a}$ is strong, not much correction is needed. In that case, $T_2$ is reduced ‘smoothly’ to $T_1$, and $T_{2a}$ retains many of the entities posited by $T_2$. In contrast, if $T_2$ and $T_{2a}$ are only weakly analogous, the amount of correction needed is considerable.
In that case, the reduction is ‘bumpy’ and many or all of the entities posited by T₂ will be eliminated from the ontology of T₂a. It is not clear, however, how exactly to evaluate the strength of the analogy between T₂ and T₂a. Additionally, NWR inherits two problems that already plagued the earlier approaches.

First, NWR is still intended as a general model of scientific reduction. This renders it blind to certain fundamental differences between different kinds of reduction (McCauley 2007; Nickles 1973; Wimsatt 1976). Most importantly, it fails to account for the difference between intralevel (or successional) relations between competing theories within a particular science (e.g. Newtonian theory of gravity and general relativity theory) on the one hand, and interlevel relations between theories (e.g. cognitive psychology and cellular neuroscience) on the other hand. In particular, the examples of eliminative, or ‘bumpy’, reductions offered by NWR are all intralevel cases and thus provide no reason to expect eliminative reductions in interlevel contexts, for example between psychology and neuroscience.

Second, NWR retains the idea that the relata of reductions are formal or at least semiformal theories, phrased in first-order predicate logic or set-theoretic terms. Yet, some generally accepted cases of scientific reduction – the reduction of genetics to molecular biology, say – do not seem to involve such formal theories (Sarkar 1992). Quite generally, while formal theories that are suitable as starting points of logical derivations may be available in theoretical physics, most special sciences simply do not have any well-structured theories that could be handled formally (see, however, Schaffner (1993) for a defence of formal approaches to reduction in biology). Explanations and reductions in these disciplines can hardly be conceived as logical derivations. Instead, these disciplines typically look for descriptions of mechanisms that can serve as explanations for patterns, effects, capacities or phenomena (see Section 4).

3. Multiple Realizability and the Reduction of Special Sciences

3.1 Multiple Realizability and Kim’s Dilemma

In the philosophy of mind, the issue of reduction surfaces in the debate between reductionists and non-reductionists. While reductionists hold that the mental can be reduced to the physical – at least ontologically, if not conceptually – non-reductionists maintain that although such reductions fail, mental properties are nevertheless not non-physical in any ontologically threatening sense: the mental is irreducible, and thus ontologically and conceptually autonomous, but since it is realized by, dependent upon, or supervenient upon the physical, it is ‘naturalistically kosher’.

Once psychophysical predicate synonymies turned out be unattainable (see Section 2), early identity theorists famously argued for a posteriori identities.
The thesis that consciousness is a brain process, Place (1956, p. 45) held, is not a consequence of a successful conceptual reduction in which mentalistic statements are shown to follow a priori from statements couched in physical terms only, but a ‘reasonable scientific hypothesis’, on a par with other a posteriori theoretical identifications like ‘Water is H2O’. Putnam (1967) objected that the a posteriori identity of mental and physical properties is an ambitious and probably false hypothesis, because mental properties are multiply realizable by different physical properties in different species, conspecifics and even one individual at different times. Fodor (1974) provided further support for non-reductionism, arguing that Putnam’s considerations apply to all special science properties. According to what Fodor (1974, p. 97) called the ‘generality of physics’, all entities subsumed under special science laws must, at root, be physical entities. Yet, since a special science property M will typically be multiply realizable, statements like ‘(∀x) (Mx ≡ Px)’ linking M with a physical property P will usually be false, and hence fail to be laws. Statements like ‘(∀x) (Mx ≡ (P1x ∨ . . . ∨ Pnx))’ linking M with the complete disjunction of all of its physical realizers will be true, but cannot be laws either, because ‘(P1x ∨ . . . ∨ Pnx)’ fails to designate a scientific kind (see Section 3.2). Hence, there are no laws – and thus a fortiori no bridge-laws – connecting special science properties with physical kinds. This renders Nagelian reductions of special science properties impossible.

Due mostly to the arguments of Putnam and Fodor, non-reductionism achieved an almost hegemonic status during the 1970s and 1980s. Jaegwon Kim (1992), however, forcefully argued that far from making reductions impossible, multiple realizability actually engenders them. Non-reductionists, he maintained, face the following dilemma.

On the one hand, if Fodor is wrong, and a disjunctive predicate like ‘P1x ∨ . . . ∨ Pnx’ designates a kind, then nothing prevents us from reducing a multiply realizable special science property via a disjunctive bridge-law. Call this the Disjunctive Move. On the other hand, if Fodor is right, and ‘P1x ∨ . . . ∨ Pnx’ does not designate a scientific kind, then the predicate ‘M’ with which it is coextensive cannot designate a scientific kind either. If there are to be any special science laws at all, then they must thus be couched in terms of the only law-fit predicates left, viz., ‘P1’, ‘P2’, . . . ‘Pn’. This leads to so-called ‘local’, or ‘species-specific’, reductions via bridge-laws of the form ‘(∀x) (Sx ⊃ (Mx ≡ Px))’ saying that if x belongs to species S, then x has M if and only if x has P. Call this Local Reductionism (if M is multiply realizable below the level of species, ‘S’ refers to individuals, individuals at times, etc.). On either horn, Kim argued, reductionism carries the day.

Another important attack on non-reductionism has come from authors who argue that multiple realizability is, in fact, not at all common in the special sciences. William Bechtel and Jennifer Mundale (1999, pp. 176–7), for example, claimed that ‘a proper examination of neurobiological and cognitive
neuroscience practice will show that the claim that psychological states are in fact multiply realized is unjustified, and that what is usually taken to be evidence for it, is not.' In a similar vein, Bickle (2003, chapter 3, esp. pp. 131–58) argued that the cellular mechanisms underlying memory consolidation are the same in fruit flies, sea slugs and rabbits, and Batterman (2000) has made a similar point with regard to the alleged multiple realizability of macrophysical properties. (For a philosophical defence of the claim that the thesis of multiple realizability has been oversold, see Shapiro (2004).)

3.2 The Disjunctive Move
Although the inadequacy of bridge-law-based approaches to reduction was evident in philosophy of science and Nagel’s account had already been replaced by more sophisticated models, the debate between reductionism and non-reductionism in the philosophy of mind was concerned with the availability of bridge-laws until the late 1990s, before alternative models of reduction were finally explored (see Section 3.4).

Proponents of the Disjunctive Move, for example, assumed with Nagel that the existence of bridge-laws linking mental and physical predicates is sufficient for reductions, and then argued that multiple realizability is compatible with reductions because there will always be true biconditionals linking mental properties with the complete disjunction of their physical realizers. In response, opponents of the Disjunctive Move tried to show that such biconditionals cannot be bridge-laws.

According to a traditional (though not universally accepted) view of laws, they exhibit two features that have been said to cause trouble for the Disjunctive Move. (∀x) (Fx ≡ Gx) is a law only if (a) it is explanatory, and (b) ‘F’ and ‘G’ are projectible in the sense that observations of Fs which are G increase confidence that the next observed F will also be G. Opponents of the Disjunctive Move have argued that disjunctive ‘laws’ fail on both counts. (∀x) (Mx ≡ (P₁x ∨ . . . ∨ Pₙx)) is not explanatory (Pereboom and Kornblith 1991; but see Jaworski 2002), and the predicate ‘P₁x ∨ . . . ∨ Pₙx’ is unprojectible because it is causally heterogeneous: from a causal point of view, there is nothing in common to all and only the individuals satisfying it (Fodor 1974; Kim 1992; 1998, pp. 106–10; but see Walter 2006).

Given this, the prospects for the Disjunctive Move seem dim. In terms of Kim’s dilemma, however, this only leads to the second horn, viz., to Local Reductionism.

3.3 Local Reductionism
According to Kim, if the disjunction of M’s physical realizers is causally heterogeneous, unprojectible, and thus non-nomic, then M (say, the property
having pain) cannot be a nomic property either, given that these properties are instantiated by the same individuals in all nomologically possible worlds: ‘If pain is nomically equivalent to [a] property claimed to be wildly disjunctive and obviously non-nomic, why isn’t pain itself equally heterogeneous and nomic as a kind? . . . It is difficult to see how one could have it both ways – i.e., to castigate [the latter] as unacceptably disjunctive while insisting on the integrity of pain as a scientific kind’ (Kim 1992, pp. 323–4). This insight, Kim argued, leads to a positive account of reduction. Consider \( P_h, P_r \) and \( P_m \), the physical realizers of having pain in humans, reptiles and Martians. Suppose \( P_h, P_r \) and \( P_m \) considered individually are causally homogeneous and thus projectible, but so different that the disjunction \( P_h \lor P_r \lor P_m \) is causally heterogeneous, and thus unprojectible and non-nomic. Given Kim’s argument that having pain cannot be nomic if \( P_h \lor P_r \lor P_m \) is non-nomic, there can thus be no laws about pain as such. The only projectible pain properties left are \( P_h, P_r \) and \( P_m \), and so the only genuine laws about pain are laws about pain-in-humans, pain-in-reptiles and pain-in-Martians. Hence, ‘there will be no unified, integrated theory encompassing all pains in all pain-capable organisms, only a conjunction of pain theories for appropriately individuated biological species and physical structure types’ (Kim 1992, p. 325). The results are restricted bridge-laws \((\forall x) \) \((S_h x \supset (Mx = P_h x))\), \((\forall x) \) \((S_r x \supset (Mx = P_r x))\) and \((\forall x) \) \((S_m x \supset (Mx = P_m x))\), which sunder the psychological theory about pain into three different subfields, each of which is ‘locally reducible’ (Kim 1992, p. 328). The same holds mutatis mutandis for all other multiply realizable special science properties.

Kim eventually came to reject Local Reductionism, however. A successful reduction of \( x \) to \( y \), he held (1998, p. 96), should be explanatory by making intelligible how \( x \) can arise out of \( y \) and simplify ontology by getting rid of \( x \) as an entity in its own right. Bridge-laws, however, universal or restricted, fail on both counts. First, even if \((\forall x) \) \((x \text{ has pain} = x \text{ has c-fibre firing})\) were a law, this would not explain why having c-fibre firing feels painish rather than ticklish (Kim 1998, pp. 95–6). Second, bridge-laws do not simplify ontology. One reason is that even if \((\forall x) \) \((x \text{ has pain} = x \text{ has c-fibre firing})\) were a law, having pain and having c-fibre firing would still not be identical, because the law would be contingent, and its contingency could arguably not be blamed on a contingency involving an epistemic counterpart, as in all other cases of scientific identifications (Kripke 1980). Another reason is that even if restricted bridge-laws like \((\forall x) \) \((S_h x \supset (Mx = P_h x))\) are true, and ‘\( M \)’ is coextensive with ‘\( P_h \)’ relative to \( S_h \) and with ‘\( P_r \)’ relative to \( S_r \) and so forth, it seems that the property \( M \) cannot be identical with \( P_h \) relative to \( S_h \) and with \( P_r \) relative to \( S_r \); \( M \) is, in this context, typically construed as a functional property – it is the second-order property of having some first-order property (\( P_h, P_r \) etc.) that occupies a certain causal role. But then \( P_h, P_r \) and so forth and \( M \) cannot be identical, for first-order occupants of causal roles cannot be identical to the second-order properties whose causal
role they occupy. Therefore, ‘Nagel reduction gives us no ontological simplification, and fails to give meaning to the intuitive “nothing over and above” that we rightly associate with the idea of reduction’ (Kim 1998, p. 97).

Kim thus became convinced that only genuine property-identities can yield reductions. He therefore modified his Local Reductionism in a way that preserved the key idea that multiple realizability leads to species-specific reductions, while at the same time allowing for genuine species-relative property-identities. The result was his model of Functional Reduction.

3.4 Functional Reduction
Kim’s model of Functional Reduction is based on ideas from David Lewis (1980). Lewis argued that instead of looking for property-identities across all possible worlds, we should identify mental properties with physical properties relative to worlds, species or structures. The concept ‘pain’, he claimed, is a functional concept in the filler-functionalism sense, not in the more popular role-functionalism sense. ‘Pain’ is the concept of a property that occupies a causal role, not the concept of the property of having a property that occupies a causal role. In contrast to the usual role-functionalism reading, a filler-functionalism reading of ‘pain’ leads to property identities: ‘If the concept of pain is the concept of a state that occupies a certain causal role, then whatever state does occupy that role is pain’ (Lewis 1980, p. 218). According to Lewis, ‘pain’ is a non-rigid designator, defined relationally in terms of the causal role of pain, which picks out different physical properties relative to different species. The gerund ‘being in pain’, in contrast, is the role-functionalist predicate that picks out the same property, viz., the functional property of having a property that occupies the pain-role, in each world, species or structure (Lewis 1994, p. 420). Thus, according to Lewis’ (not at all uncontested) view, ‘being in pain’ rigidly designates the same functional property in all creatures, whereas ‘pain’ non-rigidly designates different physical fillers of the pain-role in different species.

If ‘M’ means ‘the occupant of the M-role’ and if there is variation in what occupies the M-role, Lewis argued, then not only the contingent laws relating ‘M’ to physical predicates but the property-identities themselves are restricted: ‘not plain M = P, but M-in-K = P, where K is a kind within which P occupies the M-role. Human pain might be one thing, Martian pain might be something else’ (Lewis 1994, p. 420). Since these are genuine property-identities, Lewis’ account yields the ontologically simplifying and explanatory reductions Kim was looking for. Since M-in-K is identical to P, there is no need to recognize M-in-K as a property in its own right, and if P is the property that plays the M-role, there is no question of explaining why M-in-K is correlated with P – having M-in-K just is having P.
Lewis-style reductions are essentially three-step procedures: A special science property \( M \) is first construed via conceptual analysis as the property characterized by a certain causal role; then the physical property \( P \) occupying that causal role in a world, species or structure \( S \) is identified by means of empirical investigation, and finally \( M \) and \( P \) are contingently identified, resulting in an identification of \( M \)-in-\( S \) with \( P \). This became the key idea behind Kim’s model of Functional Reduction:

For functional reduction we construe \( M \) as a second-order property defined by its causal role . . . So \( M \) is now the property of having a property with such-and-such causal potentials, and it turns out that property \( P \) is exactly the property that fits the causal specification. And this grounds the identification of \( M \) with \( P \). \( M \) is the property of having some property that meets specification \( H \), and \( P \) is the property that meets \( H \). So \( M \) is the property of having \( P \). But in general the property of having property \( Q = \mathrm{property} \ Q \). It follows then that \( M \) is \( P \). (Kim 1998, pp. 98–9)

Kim’s new model allegedly avoids the two problems that prevented bridge-laws from yielding genuine property-identities (see Section 3.3): (a) Kripke’s argument concerning the necessity of identities, and (b) the fact that first- and second-order properties cannot be identical. Kripke’s argument is ineffective because it works only for identity statements containing rigid designators, while ‘\( x \) has pain’ is supposedly non-rigid. Furthermore, instead of talking about second-order properties, it would be more appropriate to talk about second-order designators or predicates. Second-order designators express role-concepts that are filled by first-order physical properties, and they (non-rigidly) designate these first-order physical properties, rather than a second-order property common to all individuals that satisfy them. The predicate ‘\( x \) has pain’ thus expresses the concept ‘pain’, but it neither denotes the property having pain nor any other property common to all and only the individuals that have pain (here Kim disagrees with Lewis, who acknowledged such a property, viz., the role-functional property expressed by the gerund ‘being in pain’).

Kim’s model of Functional Reduction is a kind of eliminative reduction (see Section 2.4). Having pain is abandoned as a genuine property which can be exemplified by creatures of different species; there only remain the predicate ‘\( x \) has pain’ and the concept ‘pain’ which equivocally pick out distinct properties in different species. Although mental predicates and concepts group physical properties in ways essential for descriptive, explanatory and communicative purposes, we have to learn to live without universal mental properties like having pain (Kim 1998, p. 106). It is thus clear why Kim thought the multiple realizability argument for non-reductionism fails: the differences among the physical realizers of special science properties do not show that
these properties are multiply realizable; they show that the corresponding predicate non-rigidly picks out more than one property.

One important problem with the model of Functional Reduction is that mental properties might be multiply realizable not only in different species, but also in conspecifics or even single individuals, so that having pain would be one physical property in Paul and another in Peter, or one in Paul at \( t_1 \) and another in Paul at \( t_2 \). But further narrowing mental kinds into ever more restricted physical structures seems theoretically self-defeating, as with the increasing loss of generality the identifications will be theoretically uninteresting and purely ad hoc.

Another problem is that properties which cannot be construed relationally in terms of their causal role – in particular phenomenal properties like having a reddish visual experience, having a lemonish gustatory experience and so forth – will not be susceptible to functional reductions. The problem, however, is not only that such properties turn out to be irreducible, but also that Kim’s own Supervenience Argument (Kim 1998, 2005; see also Walter 2008) is designed to show that irreducible properties cannot be causally efficacious. Kim (2005, p. 173) has thus reluctantly admitted that phenomenal properties are causally otiose epiphenomena, so that the ‘fact that blue looks just this way to me, green looks that way, and so on, should make no difference to the primary cognitive function of my visual system.’

Finally, Kim presents his model as a realistic general account of reduction in science (Kim 1998, p. 99), but does not show that scientific cases of reduction actually conform to it (Eronen forthcoming). Rueger (2006), for example, argues that Kim’s model is inapplicable in physics. Kim’s own favourite example is a biological one – the reduction of the property of being a gene to strands of DNA. But even this example is presented only very schematically and in a way that does no justice to actual history and to the philosophy of biology (Hull 1972; Schaffner 1969; Wimsatt 2007, chapter 11).2 This illustrates once again that discussions about reduction in the philosophy of mind have been largely unconstrained by, and are effectively lagging behind, developments in the philosophy of science.

4. Mechanistic Explanation, Explanatory Pluralism and Ruthless Reductionism

Mostly due to the reasons outlined in Section 2, theory reduction is nowadays not considered to be the norm in the special sciences. What has become something like the new received view on the nature of interlevel and intertheoretic relations is rather what is known as ‘mechanistic explanation’ (Bechtel 2008; Bechtel and Richardson 1993; Craver 2007; Machamer et al. 2000). The basic
insight of this approach has already been noted at the end of Section 2: if one
takes into account actual scientific practice in neuroscience and many of the
life sciences, it turns out that instead of focusing on formalizable theories and
their derivability from more fundamental ones, practising scientists try to
formulate explanations in terms of empirically discoverable mechanisms.3
Broadly speaking, mechanisms are ‘entities and activities organized such that
they are productive of regular changes from start or set-up to finish or termi-
nation conditions’ (Machamer et al. 2000, p. 3). Or, as Bechtel (2008, p. 13) puts
it, a ‘mechanism is a structure performing a function in virtue of its compo-
nent parts, component operations, and their organization.’ A mechanistic
explanation then describes how the orchestrated functioning of the mecha-
nism is responsible for the phenomenon to be explained.

Consider the example of memory consolidation (Bickle 2003; Craver 2002,
2007). A mechanistic explanation of memory consolidation describes the
cellular and molecular mechanisms underlying it by showing how the relevant
parts of the memory system and their activities together result in the
transformation of short-term into long-term memories. Central to this expla-
nation is Long Term Potentiation (LTP), a well-studied cellular and molecular
phenomenon that exhibits features that make it very likely the central part of
the memory consolidation mechanism.

Typically, mechanistic explanations have to be multilevel, because focusing
on a single level does not allow for a full understanding of the explanandum.
In the case of memory consolidation, for instance, Craver (2002) identifies four
relevant levels which, crucially, are not to be understood as general levels
of organization, but simply as the levels of the mechanism in question (see
Section 5): (a) the behavioral-organismic level (involving various types of mem-
ory and learning, the conditions for memory consolidation and retrieval, etc.);
(b) the computational-hippocampal level (involving structural features of the
hippocampus, its connections to other brain regions, and the computational
processes it supposedly performs, etc.); (c) the electrical-synaptic level (involv-
ing neurons, synapses, dendritic spines, axons, action potentials, etc.); and
(d) the molecular-kinetic level (involving glutamate, NMDA and AMPA
receptors, Ca²⁺ ions, and Mg²⁺ ions, etc.).

Mechanistic explanations have both a ‘downward-looking’ and an ‘upward-
looking’ aspect. In the LTP case, one is looking upward when, in order to
understand the computational properties of the hippocampus, one is taking
into account its environment, or when, in order to understand the role of the
molecular processes of LTP, one is looking at the larger computational-
hippocampal framework. In contrast, one is looking downward when memory
consolidation is explained by appeal to the computational processes at the
hippocampal level, or when the synaptic LTP mechanism is explained by
appeal to activities at the molecular-kinetic level.
On the one hand, since mechanistic explanation does not necessarily grant primacy to lower levels, it can be seen as supporting a kind of anti-reductionist explanatory pluralism (Craver 2007; McCauley 2007; Richardson and Stephan 2007). This anti-reductionist conclusion receives further support from the interventionist account of causation (Woodward 2003, 2008), according to which higher-level entities can have causal and explanatory relevance even if lower-level explanations in terms of implementing mechanisms are complete.

On the other hand, the process of ‘looking downward’ and invoking parts of the mechanism to understand its behaviour as a whole is close enough to what scientists generally take to be a reductive explanation to warrant treating the downward-looking aspect of mechanistic explanation as a kind of reductive explanation (Bechtel 2008; Sarkar 1992; Wimsatt 1976). Carl Gillett (2007), for example, has argued that mechanistic explanations in fact imply ontological reductions. Also, John Bickle (2003, 2006) has taken the reductive aspect of mechanistic explanations seriously, arguing for what he calls a ‘ruthlessly reductive’ analysis of explanation in neuroscience. According to Bickle, when we look at experimental practices in molecular and cellular cognition, we find a two-step strategy: a researcher (a) causally intervenes into cellular or molecular pathways in order to (b) track statistically significant differences in behaviour resulting from these interventions. If successful, this strategy establishes a scientific reduction by forging a mind-to-molecules linkage. Importantly, once the lower-level explanations are completed, higher-level sciences are retained only for heuristic and pragmatic purposes: ‘psychological explanations lose their initial status as causally-mechanistically explanatory vis-à-vis an accomplished . . . cellular/molecular explanation’ (Bickle 2003, p. 110). One problem for Bickle’s account is that while advocates of explanatory pluralism can appeal to the interventionist account of causation, it is unclear which account of causation or causal explanation Bickle could appeal to. Mechanistic explanation pluralistically understood thus seems to have a stronger case, so that explanation in neuroscience seems, if at all, only ‘somewhat’ reductive – not ‘ruthlessly reductive’, and not eliminative.

5. Reduction and Levels of Reality

Talk of levels is ubiquitous. Philosophers talk about levels of nature, analysis, realization, being, organization, explanation or existence, to name just a few. In science, the list is even longer. In the neurosciences alone, at least the following uses of the term ‘level’ can be found: levels of abstraction, analysis, behaviour, complexity, description, explanation, function, generality, organization, science and theory (Craver 2007, pp. 163–4).
Talk of levels has, of course, also been important in debates about reduction. Early on (see Section 2), when the goal was to reduce all ‘higher-level’ sciences to ‘lower-level’ sciences, one important question was how to sort the various sciences into levels. Oppenheim and Putnam (1958) proposed a preliminary division into six hierarchical levels – social groups, (multicellular) living things, cells, molecules, atoms and elementary particles – which were supposedly related mereologically in the sense that the entities at any given level are composed of entities at the next lower level.

A similar appeal to mereology can nowadays be found in Kim’s work with regard to levels of properties. The level of a property, Kim (1998, p. 92) argued, depends upon what it is a property of: properties of objects with parts are higher-level with regard to the properties of their parts, and properties of objects with no parts are fundamental properties. In addition to that, every level of reality has different orders of properties, generated by the supervenience relation: second-order properties are generated by quantification over the first-order properties that form their supervenience base (Kim 1998, p. 20). Each level thus contains lower- and higher-order properties; higher-order properties are properties supervening upon lower-order properties of the same level, not upon lower-level properties. Supervenience thus generates an intralevel hierarchy of lower- and higher-order properties, while the interlevel micro/macro hierarchy between properties of wholes and properties of their parts is not generated by supervenience, but by mereology.

The mereological appeal to composition can be found in nearly all philosophical accounts of levels of organization. In addition, size or scale are often presented as criteria (Churchland and Sejnowski 1992), where organization by size is obviously related to compositional criteria, as parts are smaller or at least no bigger than wholes. However, these criteria lead to anomalies and unwanted conclusions. A pile of snow, for example, is composed of smaller piles, but this does not mean that the larger pile is at a higher level than the smaller ones. Regarding size, there are bacterium-sized black holes and raindrop-sized computers, but it does not seem natural to say that bacteria are at the same level as black holes, or that raindrops are at the same level as tiny computers.

Perhaps the most comprehensive account of levels of organization has been developed by William Wimsatt (1976, 2007). Wimsatt’s starting point is that levels of organization are compositional levels that are non-arbitrary features of the ontological architecture of the world. Wimsatt is not aiming at a strict definition of levels, but rather at establishing sort of a ‘prototype’ idea of levels, by describing several characteristics levels typically (but not necessarily) have. For example, levels of organization are constituted by families of entities usually of comparable size, and the things at a level mostly interact with other things at the same level, so that the regularities of the behaviour of a thing are
most economically expressed in terms of variables and properties appropriate for that level. As a kind of a preliminary definition, Wimsatt (2007, p. 209) suggests that ‘levels of organization can be thought of as local maxima of regularity and predictability in the phase space of alternative modes of organization of matter.’ Roughly speaking, this means that at the scale of atoms, for example, there are more regularities than at scales just slightly larger or smaller, so that at the scale of atoms there is a peak of regularity and predictability, and thus a level of organization.

However, Wimsatt acknowledges that instead of a neat hierarchy of the kind envisaged by Oppenheim and Putnam (1958), these criteria yield a complex and branching structure of levels. Furthermore, at higher levels, for example in psychology and neuroscience, neat compositional relations break down. According to Wimsatt (2007, pp. 227–37), levels become less useful here for characterizing the organization of systems, and it becomes more accurate to talk of ‘perspectives’. Perspectives are subjective, or at least quasi-subjective, views of systems and their structures that do not give a complete description of all aspects of the systems in question, and that do not map compositionally onto one another as levels of organization do. When even the boundaries of perspectives begin to break down, perspectives degenerate into so-called ‘causal thickets’ where things are so intertwined and multiply-connected that it is impossible to determine what is composed of what and which perspective a problem belongs to (Wimsatt 2007, pp. 237–40). According to Wimsatt, the neurophysiological, the psychological and the social realms are, for the most part, such causal thickets. Unfortunately, the notions of perspectives and causal thickets remain rather vague and unclear in Wimsatt's account.

Levels also play a central role in the context of mechanistic explanations (see Section 4). The levels of mechanistic explanations are a special variety of levels of composition whose relata are mechanisms at higher levels and the components of these mechanisms at lower levels (Craver 2007, chapter 5). This notion of level is, in one respect, fundamentally different from general levels of organization. Levels of mechanisms are not universal divisions in the structure of the world (à la Oppenheim and Putnam). Rather, different mechanisms have different hierarchies of levels. The levels in the spatial memory system, for example, are different from those in the circulatory system. According to the mechanist, these local and case-specific levels are sufficient for understanding reductive explanations and interlevel relations in many fields (Bechtel 2008; Craver 2007). One limitation of this is that global comparisons become impossible. We cannot say that cells are, in general, at a higher level than molecules. All we can say is that cells in a certain mechanism are at a higher level than the molecules that are part of the same mechanism. We cannot even say that a certain molecule in a certain brain is at a lower level than the hippocampus of that brain, unless the molecule is involved in the same mechanism as the
hippocampus. Even within a certain mechanism, it is not possible to say whether subcomponents of two different components are at the same level or not, since they do not stand in a part-whole relation to each other.

Wimsatt-style levels of organization and levels of mechanisms are not necessarily incompatible. As seen above, levels of organization are said to ‘break down’ in the neurophysiological and the psychological realms, and these are exactly the realms in which levels of mechanisms are typically applied. In this sense, the two accounts may simply complement each other.

### 6. Directions for Future Research

There is a huge gap between formal approaches to reduction (e.g. NWR, Functional Reduction) and non-formal approaches that are closer to scientific practice (e.g. mechanistic explanations). What is still unclear is whether non-formal approaches are able to replace formal analyses entirely or whether they just have to be seen as complementing them. Answering that question convincingly requires a clearer picture of the possible fields of application of non-formal and formal approaches than we have as of yet, and a better understanding of the limitations, theoretically and practically, of the non-formal approaches.

Another cluster of important open questions concerns the idea of levels, a topic that has received comparatively little philosophical attention. Are there universal levels of organization or just local, case-specific, levels? What are the criteria for assigning things to levels? In what sense, if any, are lower-level explanations and theories more fundamental than higher-level ones? These questions, in turn, are related to some well-known but still unresolved metaphysical debates in the philosophy of mind that could benefit enormously from finally taking into account actual scientific practice: Is there any multiple realizability, and if so, what is its import? What are the ontological implications of successful mechanistic explanations? Is explanatory pluralism and higher-level (interventionist) causation compatible with physicalism? In particular, what are its implications with regard to the causal closure of the micro-physical and the denial of overdetermination?

Some of these questions are purely philosophical, and others have clear empirical aspects; but they are all of prime importance to the philosophy of science and should be targets of further research.

### Acknowledgements

We are grateful to Vera Hoffmann-Kolss, Bob Richardson and Dan Brooks for helpful comments.
Notes

1 See Brigandt and Hendry (this volume) on reduction in biology and chemistry, respectively.
2 Again see Brigandt (this volume) for more detail on reduction in biology.
3 Mechanistic explanation is also covered in Chapters 9, ‘Explanation’ and 14, ‘Towards a Mechanistic Philosophy of Neuroscience’.

References


1. Introduction

Science is in the business of explanation and scientific explanations can take quite different forms. Philosophers of science confront the task of finding out what is special about scientific explanations. What, for example, is the difference between mere descriptive knowledge and explanation? What are the criteria for a good scientific explanation? Is a general model of explanation possible? This chapter reviews the debates about such questions and outlines the state-of-the-art in the philosophy of scientific explanation.

For quite some time, explanation was regarded as suspect by philosophers. Following Ernst Mach, logical positivists held that the aim of science is to describe and predict phenomena, not to explain them. With the decline of logical positivism after WW II, this attitude changed, and explanation became an acceptable notion. The first full-fledged model of scientific explanation was Carl Hempel's 'covering law model', first presented in 1948. Although now rejected by most philosophers, it has remained influential and is the starting point of most discussions of explanation. Since the 1950s, an enormous literature on scientific explanation has been produced, which has resulted in a bewildering plethora of models of explanation. Most have their roots in criticism of Hempel's model, and were presented as replacements for it. But while Hempel believed that all scientific explanations (should) conform to a single model, it now seems more plausible that some models of explanation need not exclude each other. Instead, they provide alternative ways to explain phenomena in different disciplines and domains, or even different ways to explain the same phenomena. Rather than being rival accounts of explanation, they may be regarded as alternatives that may peacefully coexist.

In Sections 2–5, four main types of explanation will be discussed. Section 2 presents Hempel's model and related models that centre on explanation via laws. Section 3 deals with causal explanation, an important alternative type of explanation that is often encountered in the natural sciences. Section 4 discusses functional explanation, which is typical of the life and social sciences. Section 5 briefly investigates whether types of explanation exist that are specific to the human sciences. One may ask, of course, whether these
quite different types of explanation have something in common: what is it that
makes all of them explanations? Is an overarching analysis of explanation pos-
sible and desirable? This question will be taken up in Section 6. Subsequently,
in Section 7, I will discuss how explanation is related to other topics in the
philosophy of science. The concluding section will list some directions for
future research.

2. Nomological Explanation

2.1 The Covering Law Model of Explanation

Hempel’s model of explanation presupposes that scientists explain phenom-
ena by showing that they are consequences of general laws. The criterion for
explanation is ‘nomic expectability’: by citing laws, we give grounds for
expecting the occurrence of the phenomenon. Hempel articulated this idea as
follows: an explanation can be formulated as a logical argument in which the
conclusion (a description of the phenomenon-to-be-explained, or
explanandum) is derived from a set of premises (the explanans). The explanans contains
all relevant initial and background conditions, and at least one general law. Schematically:

\[
\text{Explanans: } \quad C_1, C_2, \ldots C_k \quad \text{particular conditions} \\
L_1, L_2, \ldots L_r \quad \text{general laws} \\
\hline \\
\text{Explanandum: } \quad E \quad \text{phenomenon}
\]

This schema presents the deductive-nomological (DN) model of explanation, so
called because the explanandum is deduced from an explanans that contains
one or more laws (nomos = law, Gr.) that cover the type of phenomenon to be
explained. (For this reason the model is also known as the covering law model.)
Since the notion of law is central to Hempel’s model, he needs an adequate
characterization of laws, a notoriously difficult question. Hempel (1965,
pp. 264–70, 338–42) took laws to be true universal generalizations, but did not
pretend to solve the problem of how to distinguish between laws and accident-
generalizations. An implication of the DN scheme is that there is no essen-
tial difference between explaining and predicting a phenomenon. This is the
thesis of structural identity (or symmetry) between explanation and predic-
tion (Hempel 1965, p. 367).

In its standard form, the explanandum of a DN explanation is a particular
event. Real science, however, is often concerned with explanation of general
regularities or laws. Think, for example, of the explanation of Boyle’s law by kinetic theory. Hempel (1965, p. 247) suggested that DN explanation of laws is possible by subsuming them under more general laws. Thus, the kinetic explanation of Boyle’s law invokes the more general laws of Newtonian mechanics plus particular assumptions about the nature of gases. There is a problem with such explanations, however: it will always be possible to deduce a law from the conjunction of itself with another law, which prima facie qualifies as a more general law. But such a derivation surely does not constitute an acceptable explanation. Hempel never succeeded in solving this problem, which hinges on the problem of giving an adequate criterion for the generality of laws.

Sometimes, when we ask for an explanation of a particular event, we can give an argument that confers a high degree of probability on the event’s occurrence, but does not deductively entail it. For example, I can explain the fact that my son has contracted the measles from the fact that several kids in his class have the measles and the law that one has a high probability of contracting measles if one has been in touch with carriers of this disease. Such a derivation, which contains a statistical law among its premises, is called an inductive-statistical (IS) explanation. It is a species of covering law explanation that is not deductive-nomological: the explanans does not deductively entail the explanandum, but lends inductive support to it.

Although it may seem natural and innocuous to demand that IS explanations confer high probability upon the explanandum, a famous example refutes this idea. It is well known that a small percentage of people with syphilis contract paresis. If someone asks why Mr X contracted paresis, we may want to explain this by saying that he suffered from syphilis and that some syphilitic patients may contract paresis. However, according to Hempel’s IS model, this does not qualify as a sound explanation, because it does not meet the high-probability requirement. So it seems that this requirement fails, and that nomic expectability is not a necessary condition for explanation.

2.2 Unificationist Models

As noted above, Hempel suggested that laws and regularities should be explained by subsuming them under more general laws, but he failed to provide a satisfactory criterion for generality. Hempel’s suggestion was the basis of a new model of explanation: the unificationist model. Like the covering law model, the unificationist model asserts that explanations are arguments in which phenomena are subsumed under general laws. In addition, however, it requires that these arguments provide a unification of the phenomena. This basic idea squares with the fact that many successful explanatory theories unify phenomena, for example, Newtonian mechanics, Maxwell’s electrodynamics, chemical atomic theory and molecular genetics.
The first unificationist model was advanced by Friedman (1974, p. 15), who argued that ‘science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given.’ He tried to elaborate this intuition into a precise account of explanatory unification: explanations would have to be unifying in the sense that they reduce the number of independently accepted phenomena by subsuming them under a more general law. Friedman’s approach proved to be untenable, however; see Salmon (1998, p. 70) for a review of objections. In response, Kitcher presented an alternative account, in which unification is achieved by deriving descriptions of different phenomena using the same ‘argument patterns’. Consequently, the value of a candidate explanation cannot be assessed in isolation, but only by seeing how it ‘forms part of a systematic picture of the order of nature’ (Kitcher, 1989, p. 430). This is explicated as follows: a successful explanation must belong to the ‘explanatory store’ E(K), where K is the set of statements endorsed by the scientific community, and E(K) is the maximally unifying systematization of K, that is, the set of derivations that employs fewer argument patterns than any other systematization of K. The unificationist model allows for the explanation of laws in terms of more comprehensive laws, while it evades Hempel’s problem of finding a criterion for generality. Moreover, Kitcher (1989, p. 447) suggests that unificationism solves the notorious problem of characterizing laws: these can be simply defined as ‘the universal premises that occur in explanatory patterns’. In sum, the unificationist model of explanation is a worthy successor to Hempel’s model, preserving the basic features of deductive argument and subsumption under laws, but avoiding the problematic aspects of its predecessor.

In recent years, much further work has been done on unification. A notable alternative model is offered by Schurz (1999). Nevertheless, Kitcher’s work, particularly his comprehensive 1989 essay, remains the *locus classicus*, and most criticisms of the unification model are directed to his model. Woodward (2003, pp. 359–73), for example, argues against its implication that in any domain only the most unified theory is explanatory at all. Craver (2007, pp. 42–9) objects that the model cannot be used to assess individual explanations here and now. Both Woodward and Craver argue that not unification, but causality, is the essence of explanation.

3. Causal Explanation

While the scientific literature is full of talk about causes and effects, logical empiricists regarded the notion of causality as suspect and tried to eliminate
causal notions from the analysis of science. In this vein, Hempel (1965, p. 349) claimed that ‘causal explanation is, at least implicitly, deductive-nomological.’ By contrast, Scriven (1962) argued that a causal story can be explanatory in itself. For example, one can explain the tipping over of an ink bottle by telling a story of how somebody’s knee hits the table, causing the bottle to fall, without having to refer to laws. Hempel responded by claiming that Scriven’s story is merely an ‘explanatory sketch’ of a complete DN explanation. Scriven maintained, however, that a causal story needs no further articulation in order to be truly explanatory.

Many counter-examples to the covering law model suggest that causality is crucial for explanation. For example, one can predict an impending storm from a barometer reading combined with the law that whenever the barometer drops, a storm will occur. But surely this argument does not yield an explanation of the storm. Instead, a satisfactory explanation can only be given on the basis of a common cause, namely, the occurrence of very low air pressure. The barometer example suggests that, contra Hempel’s symmetry thesis, there is an crucial difference between prediction and explanation, and that causality is essential to explanation. Another famous example concerns a flagpole and its shadow. On the DN model, the length of the shadow of a flagpole can be explained by citing the height of the flagpole, relevant circumstances such as the position of the sun, plus the general laws of geometrical optics. However, one may equally well deduce the height of the flagpole from the length of the shadow. Both arguments conform to the criteria for DN explanation, but only the first one qualifies as an explanation, at least according to most people’s intuitions. Thus, there seems to be an asymmetry in explanation that is not captured by the DN model. Invoking causality provides an answer: causes explain effects, effects cannot explain causes.

3.1 The Causal-Mechanical Model
The most influential model of causal explanation was developed by Salmon. He abandoned the Hempelian view that explanations are arguments, and adopted an ‘ontic’ view, which asserts that ‘events are explained by showing how they fit into the physical patterns found in the world’ (Salmon 1998, p. 64). According to Salmon, these patterns are causal: ‘causal processes, causal interactions, and causal laws provide the mechanisms by which the world works; to understand why certain things happen, we need to see how they are produced by these mechanisms’ (Salmon 1984, p. 132). To substantiate this idea, Salmon (1984, pp. 135–83) presents a detailed theory of causality, in which the basic entities are causal processes. A process is causal if it can transmit a ‘mark’: a persistent change in its structure. Causal processes are the means by which causal influence is transmitted. If two or more causal processes intersect and
the result is a persistent change of (at least one of) the processes, this is called a causal interaction. In other words, causal interactions generate and modify causal structure. Explaining a phenomenon consists in describing the causal processes and interactions that have produced it (Salmon 1998, p. 60).

Salmon’s model has been widely discussed in the literature. One of the issues that received much attention is the question of whether an account of causation requires counterfactuals. Salmon’s theory of causality refers to counterfactuals, because his criterion for mark transmission contains a claim about what would have happened if the process had not been marked (Salmon 1984, p. 148). Salmon was unhappy with this, as it ran counter to his empiricist principles. He therefore accepted an alternative way to characterize causal processes and interactions proposed by Dowe (2000). Dowe invokes the notion of conserved quantities (obeying physical conservation laws) and defines causal processes as world lines of objects that possess a conserved quantity, and causal interactions as intersections of world lines that involve exchanges of conserved quantities.

Salmon’s causal-mechanical model is clearly inspired by the physical sciences. It can be questioned, however, whether it is a satisfactory model of all explanatory practices. On the one hand, it turns out that the model fails at the most fundamental level of reality: quantum-mechanical phenomena such as the Einstein-Podolsky-Rosen correlations cannot be explained in causal-mechanical fashion, because quantum-mechanical ontology cannot be described in terms of causal processes and interactions à la Salmon and Dowe (De Regt and Dieks 2005, p. 145). Salmon (1984, p. 133) admitted that his model fails in quantum mechanics, but hoped that the notion of mechanism could be generalized so as to apply to quantum mechanics. On the other hand, because the model is closely tied to physical science, it does not fit all explanatory practices of the life sciences, let alone of the social sciences. In recent years, alternative models have been proposed that are better suited for these purposes: first, models based on a different conception of mechanism; second, models based on a different analysis of causality. We will now turn to the latter and discuss the former in Section 4.2.

3.2 Alternative Models of Causal Explanation
A currently popular theory of causal explanation, proposed by Woodward (2003), is based on a conception of causation in which intervention and manipulation are central. According to Woodward, causal relation ‘X causes Y’ holds if and only if the value Y would change through intervention X. This allows, for example, for determining the effect of a drug by means of double-blind testing procedures. One group of patients is given the drug, while another (comparable) control group is given a placebo under similar circumstances
(as far as possible). Later measurements are taken to determine if the drug has had the desired effect, and the patient groups are compared to see whether or not the differences are statistically significant. In such experiments, the experimenter intervenes by manipulating the cause (treated/untreated) and observing whether an effect results (recovery/non-recovery). Woodward asserts that a good (causal) explanation exhibits patterns of counterfactual dependence that describe the outcomes of interventions. Explanations thereby allow us to answer ‘what-if-things-had-been-different questions’. For example, if a particular drug is effective, the outcome can be explained by the fact that the patient took the drug. Explaining a positive outcome for a patient as a result of the administrated medication is the same as responding to the question ‘what would have happened if the patient had not taken the drug?’ with the answer: he would (probably) not have recovered. A double-blind experiment devised to support this explanatory claim can be seen as an attempt to establish such counterfactual dependency relations. An advantage of Woodward’s causal theory lies in its general applicability. Since it does not make special assumptions about causal mechanisms or need laws of nature, it can apply not only to physics, but also to biomedical and social sciences (where mechanisms and laws are harder to find). Thus, it seems that Woodward’s theory is able to accommodate a wide range of explanatory practices across the various sciences.

Another recent model that intends to cover scientific explanation in various domains is the ‘kairetic account’ of Strevens. While agreeing with Salmon, Woodward and others that phenomena are explained by describing how they are causally produced, Strevens (2008, pp. 27–35) does not commit himself to a particular metaphysical theory of causation, but has an ‘ecumenical’ view: (almost) all existing theories of causation are compatible with the kairetic account. Of all causal influences, only some are relevant to a phenomenon’s occurrence. Take, for example, a car accident in winter; both an icy road and the gravitational effect of Mars are part of the situation and influence the event, but only the former is causally relevant. This is because the icy road makes a difference to the occurrence of the accident, whereas the gravitational effect does not. The core of the kairetic theory is an optimizing procedure used to identify the difference makers for events-to-be-explained (Strevens, 2008, pp. 86–110). This procedure involves eliminating irrelevant influences and abstracting the remaining influences in such a way that irrelevant details are removed (e.g. the exact structure of the ice may not be relevant, but the fact that the icy layer was, say, at least 1mm thick is relevant). The optimizing procedure leads to a ‘standalone explanation’, which is complete and sufficient for scientific understanding of the event. There can be many standalone explanations of one particular event, which can all be satisfactory and sufficient for understanding. However, Strevens (2008, pp. 123–37) argues that some can be
better than others and that, in particular, ‘deepening’ explanations, which give a lower-level account of why higher-level causal laws obtain, are superior from an explanatory point of view.

4. Functional Explanation

Many proffered explanations in the life sciences, psychology and the social sciences refer to functions of entities. These sciences are concerned with complex organized systems, the components of which contribute to the working of the system (organisms, human minds, societies and so forth). A functional explanation typically accounts for the role or presence of a component item by citing its function in the system. Functional explanations are often associated with teleology, the idea that there is goal-directedness in nature, because a functional description refers (implicitly or explicitly) to the goal of the system. Teleology was central to Aristotelian philosophy, which regarded ‘final causes’ as most important, because they appealed to the purpose of things for which they were designed. After the Scientific Revolution, however, teleology came to be regarded as suspect in all natural sciences, as it did not fit into the mechanical philosophy that had replaced the Aristotelian system. (In the human sciences, there is still room for teleology, because humans do have goals and intentions; see Section 5). Thus, the challenge for contemporary philosophers of science is to analyse functional explanation in such a way that it is compatible with modern science.

4.1 Early Views of Functional Explanation

Traditionally, philosophers assumed that functional explanations try to account for the presence of an item by showing how it contributes to the preservation or development of a system in which it occurs. The standard example: humans have a heart because it has the function of pumping blood, and pumping blood is required to stay alive. However, such explanations cannot be reformulated as DN explanations: from the fact that a functional item contributes to the preservation of the system, one cannot deduce that the item must exist, because there are always alternative ways to fulfil the function. For example, blood may be circulated by means of an artificial heart. This led Hempel (1965, p. 313) to conclude that functional explanations are illegitimate. Instead, valid scientific explanations of the behaviour of organized systems and their components will take the form of subsumption under laws, and will not differ essentially from explanations in the physical sciences. In the case of complex biological and social systems, these laws are typically laws of self-regulation, which specify feedback mechanisms. Nagel (1961,
Explanation

pp. 403–4), by contrast, argued that legitimate functional explanations do exist in biology, if items are de facto indispensable for growth and reproduction.

The demands placed on functional explanations by Hempel and Nagel are quite strong: they claim that their goal is to explain the presence of functional items, and that they should do so deductively. No wonder, critics have argued, that most actual functional analyses failed on their criteria; these criteria are simply too demanding. Thus, Cummins (1975) suggests that the aim of functional analyses is to show how the item contributes to the functioning of the system as a whole. Cummins claims that functions are dispositions (or capacities), and as such, they require explanation. Rather than explaining functions by subsuming them under general laws, Cummins argues that the function of a system should be explained by analysing it into a number of other dispositions, the organized manifestation of which results in a manifestation the system’s function. This implies that we can only speak of the function of an item against the (implicit or explicit) background of a chosen analytical explanation. The choice for a particular explanation depends on the context. Accordingly, an item may have different functions in different explanatory contexts. For example, while the heart is usually ascribed the function of a blood pump (in the context of explaining the circulatory system), the fact that it makes a noise may be functional, for example, in the context of medical diagnosis. This contradicts the traditional view, which qualifies pumping as the ‘real’ function, while the noise it makes is a mere side-effect.

To avoid trivialization, Cummins (1975, p. 764) lists criteria for evaluating when analytical explanations are truly interesting: the capacities in terms of which the function is analysed should be less sophisticated and different in type, and the organization of the resulting system should be relatively complex. In his later work, Cummins (2000) has applied his account to explanations in psychology and cognitive science, which are typically directed to capacities, such as the capacity to see depth, to learn and speak a language, and so forth. A further articulation and development of Cummins’ approach has been presented by Craver (2007, pp. 107–62), who argues that, at least in the life sciences and neuroscience, analytical explanations of functions are typically mechanistic explanations (see Craver and Kaplan, this volume). It is this idea, which offers a prospect of integrating functional explanation and causal-mechanical explanation, to which we will now turn.

4.2 The ‘New Mechanist’ Approach to Functional Explanation

Salmon’s causal-mechanical account of explanation, discussed in Section 3.1, is modelled on explanatory practices in the physical sciences. In recent years, philosophers have developed mechanistic approaches to functional explanation that are inspired by the contemporary practices of the life sciences.
The core of these approaches is an analysis of mechanisms in the spirit of Cummins: a mechanism is an organized whole that, by virtue of the interaction of its parts, produces specific behaviour or performs a particular function. Most ‘new mechanists’ place themselves in the tradition of causal-mechanical explanation and regard their proposals as complementing, rather than replacing, Salmon’s account. In their influential paper ‘Thinking about mechanisms’, Machamer, Darden and Craver (2000) characterize mechanisms as entities and activities organized so that they produce the phenomenon-to-be-explained. Phenomena are explained by describing the mechanisms that produce them. Their account is further developed in subsequent publications; see, for example, Craver (2007) and Darden (2008). Slightly different characterizations of mechanisms have been proposed by Glennan (2002), and Bechtel and Abrahamsen (2005), but all agree on the essential elements: biological mechanisms exhibit complex organization and, by virtue of the organized interaction of their parts, produce specific functional behaviour. A drawback of most mechanistic accounts is that they are descriptive and lack normative power. Craver (2007, p. 161) has attempted to remedy this deficiency by providing normative requirements for satisfactory mechanistic explanations.

Mechanistic explanations are typically not purely linguistic, but also use diagrammatic representations (Bechtel and Abrahamsen 2005, pp. 427–30; Machamer et al. 2000, pp. 8–9). Scientists prefer diagrams because they can directly convey the spatial organization of mechanisms. Diagrammatic representations of complex mechanisms have clear pragmatic advantages; they are far more tractable than linguistic representations. Moreover, reasoning about diagrams can be facilitated by simulation tools, such as scale models or computer models.

Mechanistic explanations are typically not constrained to a single level, but span different levels: the system is described in terms of its lower-level parts. But this does not imply that explanation is a reductive affair. Rather, as Craver (2007, pp. 256–71) argues, mechanistic explanations involve ‘inter-level integration’ of hierarchically organized mechanisms into one coherent mechanism, which results in a description of nested networks of mechanisms that explain phenomena at different levels. For example, a mechanistic description of the circulatory system in terms of the heart, blood vessels, the kidneys, the lungs and so forth explains its activity of delivering oxygen and nutrients to the body, while its parts (e.g. heart) can be mechanistically described in order to explain their respective activities (e.g. blood-pumping).

Remarkably, whereas much traditional philosophy of biology focuses almost exclusively on evolution, the new mechanists pay scant attention to evolution, and one might indeed ask whether their approach fits evolution at all (Darden 2008, p. 967). On the other hand, it has been suggested that
the mechanistic approach can be applied to domains outside biology, such as chemistry and cognitive psychology (Bechtel 2008; Darden 2008; Ramsey 2008).

5. Explanation in the Human Sciences

The models discussed in the previous sections are mainly inspired by explanatory practices in the natural sciences. Do they also apply to the human sciences: sciences that study humans or human affairs, for example, social sciences, psychology, historical science? Or do these sciences require alternative modes of explanation? To be sure, social scientists and psychologists employ nomological, functional and causal explanations, but there are reasons to think that these do not suffice. Furthermore, in historical science, such explanations often seem impossible or irrelevant. The main argument is that human agents have intentions, purposes and interests, and that adequate explanation of human actions should take these into account. In contrast to objects of natural science, the objects of the human sciences possess an ‘inside’ aspect that is not accessible via sensory observation and experiment (Collingwood 1994, p. 213). This ‘inside’ consists of the reasons and motives of humans for their actions. Such internal reasons stand in contrast to external causes of behaviour. Human sciences must also study the ‘inside’ of phenomena, by understanding the minds of the agents involved, and this requires a distinctive method.

Some have argued that this understanding can be achieved via empathy: historians, for example, try to discover and understand the motives and thoughts of historical agents by putting themselves in the place of the agents when they study historical evidence (Dray 2000, p. 224). The idea of empathic understanding and explanation of human action has recently been defended by Stueber (2006). Alternatively, it has been suggested that the essence of historical explanation lies in narration: thus, historians may explain an event by telling a story, for example, of events that led up to the event-to-be-explained. A narrative typically transcends the mere recounting of a sequence of events; it has a holistic aspect and brings the events together into a synthetic unity by invoking ‘colligatory concepts’, for example, the Renaissance or the Industrial Revolution (Dray 2000, p. 234; Mink 1987, pp. 81–4).

It may happen that a particular human practice can be explained in different ways, according to different explanatory strategies. For example, in explaining the phenomenon of gift giving, some anthropologists may focus on ‘inside’ aspects, the intentions involved, and try to account for gift giving along these lines. But others may seek a functional explanation of it, emphasizing the role of gift-giving practices in maintaining social structure. Moreover,
sociobiologists may try to underpin such a functional account by a causal explanation. It might be asked whether these different explanations are competing or complementary. This is the topic of the next section.

6. The Plurality of Models

We have discussed four distinct categories of explanation, and we have seen that even within these categories there is a variety of models of explanation. Many of these models do not have universal pretentions, but offer a possible way to explain phenomena in a restricted domain. This contrasts with the views of Hempel and other early philosophers of explanation, who wanted to formulate a model of explanation that would apply to science as a whole. Acknowledging that this ambition is misguided and developing various models of explanation for specific disciplines or domains is a big step forward. But now the question arises whether a single philosophical framework can be developed in which these different models can be embedded as types of explanation. This seems desirable, for one would at least want to know why the various models are all species of the same genus called explanation. In this section I will discuss two ways to account for the plurality of models of explanation. The first emphasizes the pragmatic nature of explanation, implying that explanation can have different forms in different contexts. The second is also contextual, but focuses on understanding, suggesting that understanding is a universal aim of science that can be achieved by contextually varying modes of explanation. The two approaches are not necessarily incompatible.

6.1 The Pragmatics of Explanation

Scientific explanation has a pragmatic dimension. In the real world, explanations are requested by some people and provided by others. What kind of explanation is suitable, acceptable, or sufficient depends on the people involved and on the situation. For example, a layperson may be puzzled by a given explanation of the outcome of a particular experiment, say, while a scientifically educated person may be enlightened by the same explanation, and an expert may not be satisfied with it. Hempel (1965, pp. 425–8) claimed that philosophers of science should abstract from such pragmatic aspects, and only a few early philosophers (most notably Scriven) paid attention to the pragmatic dimension of explanation. This changed in the 1980s, when Bas van Fraassen, Peter Achinstein and others presented full-fledged pragmatic theories of explanation. A pragmatic account of explanation may offer a prospect of unifying the various models of explanation discussed above, because it is built on the idea that explanations are given and received by
particular people in particular contexts for particular purposes. Different
contexts, people and purposes may require different types of explanation.

The most influential pragmatic theory of explanation is that of van
Fraassen (1980), developed as part of his constructive empiricism, which
asserts that empirical adequacy is the only epistemic aim of science. Other
aims of scientists, notably explanation, are pragmatic. According to van
Fraassen, explanations are answers to questions of the form ‘Why P?’ Taken in
isolation, this question does not determine the kind of answer that is asked
for. For example, the question ‘Why did Adam eat the apple?’ may be inter-
preted in at least three different ways, depending on whether ‘Adam’, ‘the
apple’, or ‘eat’ is emphasized. In these three interpretations, different kinds of
answers are expected. This reveals that an explanation-seeking question
always (at least implicitly) refers to a contrast-class \( X \) \( (P_1, P_2, \ldots, P_n) \) of alter-
native possibilities: ‘Why \( P_i \) [rather than another member of \( X \)]?’ Often, this
contrast-class is not made explicit, because it is clear from the context what it
is. Sometimes an explanation-seeking question does have an answer in one
context and not in another. Take, for example, the case discussed earlier: ‘Why
did Mr X contract paresis?’ has an answer if the contrast-class is the set of all
inhabitants of the town in which Mr X lives, but has no answer if the contrast-
class is the set of other syphilitics.

On the basis of these ideas, van Fraassen (1980, pp. 134–53) develops
a theory of explanation-seeking why-questions and their answers. A crucial
question concerns the evaluation of answers: what counts as a satisfactory
answer to the question ‘Why \( P_i \)?’ First of all, an answer, \( A \), must be scienti-
cally relevant; for example, likely to be true and favouring \( P_i \) over its alter-
natives. However, according to van Fraassen, not all scientifically relevant factors
are explanatorily relevant to a particular explanation request: explanations
single out ‘salient factors’ among the scientifically relevant factors. No general
criteria for explanatory relevance exist; it is the context that determines what
is explanatorily relevant. An explanation is ‘salient to a given person because
of his orientation, his interests, and various other peculiarities in the way he
approaches or comes to know the problem – contextual factors’ (van Fraassen
1980, p. 125). This also applies to asymmetry cases, such as the flagpole and
the shadow, since the context determines the relevance relation: in particular
contexts the length of the flagpole can be explained by the length of the shadow
(van Fraassen 1980, pp. 130–4).

Van Fraassen’s pragmatic theory of explanation does not replace the specific
models discussed in Sections 2–5, but incorporates them as possible answers to
explanation-seeking questions. What kind of answer is explanatorily relevant
depends on contextual factors, such as the interests and background knowl-
edge of the questioner. In a particular situation, a causal explanation may be
appropriate; in another situation, a nomological or functional explanation may
be called for. Alternative pragmatic models of explanation have been developed by Achinstein (1983) and Sintonen (1999); see Faye (2007) for a review of pragmatic models of explanation and further discussion of the way in which the pragmatic theory incorporates other models.

6.2 Understanding: A Common Goal of Explanation

Another way of accounting for the plurality of forms of explanation in scientific practice is by invoking the notion of understanding. A common feature of all types of explanation discussed in Sections 2–5 is that they provide understanding. Accordingly, there is a variety of explanatory strategies to reach the single aim of scientific understanding. Scientists choose, from a host of available strategies, whatever suits them best in a particular case to achieve this aim. The history and practice of science shows variation in preferred ways to achieve understanding.

What is scientific understanding, and how is it produced by the various types of explanation? These questions have been addressed by De Regt and Dieks (2005), who argue that the understanding achieved through explanation is not an extrascientific, pragmatic virtue (as van Fraassen claims), but forms part of the epistemic aims of science. According to De Regt and Dieks, explanatory understanding of phenomena requires intelligible theories, where intelligibility is related to the ability of scientists to use the theory. More precisely, intelligibility is defined as the value that scientists attribute to the theoretical virtues that facilitate the use of a theory for the construction of models (De Regt 2009, p. 31). Intelligibility is a necessary condition for constructing explanations and predictions on the basis of a theory. It is not an intrinsic property of theories, but a context-dependent value, and accordingly, a theory may be intelligible to one scientist, but not to another. This accounts for the variety of ways in which understanding is achieved in scientific practice. But such contextual variation does not entail that intelligibility is merely a matter of taste and cannot be tested. De Regt (2009, p. 33) suggests the following test for intelligibility: ‘A scientific theory T is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations.’ Causal reasoning, for example, can provide understanding because it enhances our ability to predict how the systems that we study will behave under particular conditions. However, understanding may also be achieved with the help of different conceptual tools, for example, visualization or analogical reasoning.

The theory of De Regt and Dieks depends crucially on a distinction between understanding phenomena (an epistemic aim of science) and understanding theories (intelligibility). While the latter is pragmatic and context-dependent,
it is a necessary condition for achieving the epistemic aim. Thus, the epistemic and the pragmatic dimensions of science are inextricably intertwined (De Regt 2009, p. 28). Like van Fraassen (1980, pp. 87–9), De Regt and Dieks argue that there is an essential pragmatic element to explanation, but they disagree with his sharp distinction between the epistemic and the pragmatic and his claim that explanation belongs to the pragmatic dimension only.

In sum, one does not have to deplore the present situation, in which there are many different accounts of explanation. This is not just a sign of disagreement or lack of consensus (though in some cases it is), but is also a sign of progress and deeper insight. While the early philosophers of explanation wanted a universal theory of scientific explanation, it has now become clear that this is an unrealistic and undesirable ideal. Scientific practice is varied, and explanations can have very different forms. The various models of explanation that have been developed are detailed accounts of how scientists explain phenomena in different disciplines and contexts. A general philosophical theory of explanation should respect this diversity, and the two theories discussed in this section are examples of how this may be done.

7. Wider Philosophical Implications

Explanation has been one of the main themes in philosophy of science in the last 50 years. But debates about explanation have not been isolated: there are many relations and interactions with other discussions in the philosophy of science. In this section, I will briefly review some important points of contact between explanation and other central issues in contemporary philosophy of science. Of the general issues dealt with in Part IIA of this Companion, four in particular appear to have direct links to explanation: realism, representation and models, reduction, and laws and causation. I will discuss each of them in turn.

7.1 Realism

Scientific realists hold that successful scientific theories are (approximately) true descriptions of the world, also at the theoretical, unobservable level. Realist philosophers, such as Boyd (1984), have suggested that the explanatory power of theories vindicates scientific realism. Their argument is based on inference to the best explanation (IBE), which is the idea that, in the face of given evidence and a set of competing hypotheses, we should infer to the truth of the hypothesis that best explains the available evidence (Lipton 2004). Applied to competing scientific theories, IBE states that one should infer the truth of the theory that is the best explanation of the empirical data. There
has been much debate about whether IBE is a reliable form of inference, and specifically whether it can support scientific realism. Van Fraassen (1980, pp. 19–22) argues against IBE, and even Lipton, who defends IBE as an inferential strategy, finds only very limited use for it in arguments for realism (2004, chapter 11). An important issue for the proponents of IBE is to determine criteria for best explanations. Lipton (2004, pp. 60–1) observes that we cannot equate ‘best’ with ‘likeliest’, because that would render IBE circular. Therefore, he suggests that it should be interpreted as ‘loveliest’, that is, as ‘the explanation that would, if true, provide the deepest understanding.’ But what are the criteria for loveliness? While Lipton (2004, p. 57) claims that the DN model cannot be used to flesh out IBE (because it would reduce loveliness to likeliness), it seems that the various alternative models of explanation discussed above are, in fact, highly suitable for this purpose.

7.2 Representation and Models
Contemporary philosophy of science acknowledges the various roles that models play in scientific practice. First and foremost, models are representations (of phenomena or theories), but models are also used for explanatory purposes. A theory of explanation in which models occupy central stage is Cartwright’s simulacrum account, which asserts that ‘to explain a phenomenon is to construct a model that fits the phenomenon into a theory’ (Cartwright 1983, p. 17). According to Cartwright, Hempel’s model fails to account for explanation in practice, because the laws used in explanations are typically false: theoretical laws are fictions that do not apply to real phenomena, but only to idealized models of them. To explain real phenomena, we need to represent them in terms of models. If one has succeeded in constructing a model (‘simulacrum’) and shown how the model can be derived from the theoretical laws, the phenomenon has been explained. Cartwright is not clear about how models provide explanations; she merely states they do so. Morrison (1999, p. 63) is more explicit and suggests that models are explanatory because they exhibit structural dependencies: ‘The model shows us how particular bits of the system are integrated and fit together in such a way that the system’s behaviour can be explained.’ Morrison emphasizes, like Cartwright, that application of theoretical laws in specific circumstances requires modelling. For example, one cannot explain the behaviour of a real pendulum with only the laws of Newtonian mechanics; one needs a model of the pendulum that provides details about the structural dependencies: the model ‘contextualizes the laws in a concrete way’ (Morrison, 1999, p. 64). But there is no unique way in which the explanatory power of a model relates to its representational features: this depends on the context and on the purposes for which the model is used.
7.3 Reduction
There is a direct and intimate link between reduction and explanation. The classical model of reduction developed by Nagel (1961, pp. 336–66) can be regarded as a species of deductive-nomological explanation, since it consists in a deductive explanation of particular theories and laws in terms of other theories and laws. Nagel’s model encountered serious problems and is widely discredited today. The alternatives are either anti-reductionism or some weaker form of reductionism.

Many of the alternatives to DN explanation have specific implications for the issue of reduction as well. For example, a direct link exists between unification and reduction. Reductionist scientists such as Steven Weinberg (1993) suggest that the ‘final theory’ of physics will be a ‘Theory of Everything’ that unifies all phenomena by reducing them to fundamental particle physics. A sophisticated way to flesh out this intuitive idea of reductive explanation is suggested by Kitcher (1989, pp. 447–8): reduction (both inter- and intra-level) consists in the extension of the range of phenomena explained by the same argument pattern. Of the causal models of explanation, Salmon’s theory is fully compatible with a reductionist view in which the fundamental causal structure of the world forms the basis of all explanations. In addition to etiological explanations (in which antecedent causes explain effects), Salmon (1984, p. 270) allows for ‘constitutive explanations’ that reductively explain phenomena on the basis of underlying causal mechanisms. The same holds for Strevens (2008, pp. 81–3, 470–2). On Woodward’s ‘manipulationist’ view, by contrast, explanation and reduction are independent: causal explanation does not entail reductionism, and reduction is not a (privileged) form of explanation.

Functional explanation, if regarded as a separate species of explanation, is associated with anti-reductionism, since it provides an autonomous mode of explanation used in higher-level sciences such as biology, cognitive science and social science; see Looren de Jong (2003). Mechanistic accounts of functional explanation have reductionist elements; however, they do not equate explanation with reduction, but rather with inter-level integration. The models of explanation in the human sciences discussed in Section 5 are essentially anti-reductionist, emphasizing the distinctive character of explanations in the human sciences and the impossibility of explaining human-scientific phenomena exhaustively on the basis of lower-level natural sciences.

7.4 Laws and Causation
The themes of laws and causation are related to explanation in various ways. First of all, the notion of law is a central (but problematic) element of Hempel’s covering law model. However, laws are typically elusive in higher-level
natural sciences (Beatty 1995; Kleinhans et al. 2005), and even more problematic in the human sciences (Dray 2000). Later philosophers of explanation have made attempts to solve these problems by proposing radically different accounts of laws. Kitcher (1989, p. 447) characterizes laws as universal premises in explanatory patterns, thereby reversing the relation between explanation and laws: the latter derives from the former, instead of the other way around. Woodward (2003, pp. 239–314) introduces the weaker idea of ‘invariance’ as an alternative for law. Invariance comes in degrees, and it applies to many generalizations that are used in scientific practice but do not qualify as laws (because they are not exceptionless, for example). Woodward’s account is especially congenial to the life sciences and social sciences.

Debates about the nature of causation are, of course, specifically relevant to the causal models of explanation discussed in Section 3. A prerequisite for developing a causal theory of explanation is having a satisfactory analysis of causation. Salmon devoted much effort to formulating an account of causal processes and interactions that would support his causal-mechanical theory. Central to Salmon’s analysis is a distinction between causal processes and pseudo-processes, on the basis of the criterion of mark transmission (Salmon 1984, pp. 139–57). Salmon later abandoned this account of causation and opted instead for a ‘conserved quantity’ theory (see Section 3.1). A quite different analysis of causation is Woodward’s manipulability theory, according to which manipulation and intervention are central to causation (see Section 3.2). Woodward’s theory is motivated by the intuition that the use of the concept of causation is rooted in the practical values of manipulation and control. Moreover, it relates causation to experimentation and is applicable in a variety of scientific disciplines and contexts.

8. Directions for Future Research

The philosophy of scientific explanation is, and will continue to be, a flourishing research field. While important progress has been made in the last decade, much work remains to be done. The new mechanist approach is currently very popular, but faces a number of challenges. For example, there is still no consensus about the precise characterization of mechanisms, about their ontological status and about the normative power of mechanistic analyses. Moreover, while due attention has been paid to mechanistic explanation in molecular biology and neuroscience, application to other domains and disciplines is in its infancy. The mechanist approach to functional explanation nicely complements Salmon’s causal-mechanical model. An important open question is whether it is also compatible with alternative causal models, such as Woodward’s or Strevens’. More generally, an interesting question concerns
the scope of causal models. Answers should be found via application of these models to case studies of actual explanatory practices in various scientific disciplines. Furthermore, the future of causal models will depend on whether consensus will emerge regarding the analysis of causation.

Meanwhile, explanation by unification has received less attention in the last decade. The objections that have been raised against it have not yet been adequately answered by its proponents. Also, they should demonstrate that unification is, in fact, part of the explanatory practice of scientists, a thesis that has recently been denied (Morrison 2000, p. 63). Finally, as I have argued above, the current diversity of models of explanation is not a problem, but is, in fact, a sign of progress. In addition, however, to further developing detailed models of scientific explanation for different domains and contexts, we need a general philosophical framework that accounts for this diversity. The pragmatic approaches discussed in Section 6 can prove to be well suited for this purpose.

Acknowledgements

I wish to thank Petry Kievit for helpful suggestions and discussion. This chapter was completed during a research leave at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS).

Notes

1 See Chapter 8, ‘Reduction, Multiple Realizability and Levels of Reality’, on the notion of ‘level’.
2 Mechanistic explanation in neuroscience is covered in detail in Chapter 14, ‘Towards a Mechanistic Philosophy of Neuroscience’.
3 This idea goes back to the nineteenth century, when Wilhelm Dilthey argued that the human sciences (Geisteswissenschaften) aim at understanding (Verstehen), while natural science aims at explanation (Erklären).
4 Narrative explanation is sometimes used in natural science as well, especially in cosmology and geology (Kleinhans et al. 2005).
5 For the social sciences, classic attempts to formulate distinctive models of explanation are Winch (1958), who emphasized the need for inside knowledge of rules in social practices, and Geertz (1973), who argued that human actions can be understood only by ‘thick description’, which includes the context of the behaviour.
6 See Chapter 5, ‘Scientific Realism with a Humean Face’ on scientific realism in general.
7 See Psillos (1999, pp. 78–97) for an overview of the debate and an ‘explanationist’ defence of realism. Hitchcock (1992), by contrast, argues that (the theory of) explanation is irrelevant to the realism debate.
8 See the chapter ‘Scientific Models and Representation’ for further discussion on these and related issues.
This applies to reduction of higher-level theories and laws in terms of lower-level theories and laws (inter-level reduction; for example, thermodynamics to statistical mechanics) and to reduction of theories at the same level (intra-level reduction, for example, Newtonian mechanics to relativity theory).

The problems of Nagel’s model are reviewed in Chapter 8, ‘Reduction, Multiple Realizability and Levels of Reality’. Dizadji-Bahmani et al. (2010) defend the Nagelian account.

See Chapter 6, ‘Causation and the Sciences’ for discussion of these themes.

References


1. Concepts of Evidence

In mainstream epistemology, evidence is commonly considered to be that which justifies belief. As Kelley (2006) puts it, ‘Evidence, whatever else it is, is the kind of thing which can make a difference to what one is justified in believing or (what is often, but not always, taken to be the same thing) what it is reasonable for one to believe.’ In philosophy of science, it is widely accepted that a prediction made by a theory is evidence for that theory when the prediction is observed to be true. As an example, Einstein’s general theory of relativity (his theory of gravitation) predicted that the sun’s gravity would bend starlight by twice the amount predicted by Newton’s theory of gravitation; such observations are therefore evidence for Einstein’s theory and against Newton’s theory.¹ But everyday examples are like this, too; that the flag on my mailbox is down is evidence that the mail carrier has already delivered the mail, for if the mail carrier had already delivered the mail then he would have put the flag down, and indeed, the flag is down. It’s not conclusive evidence, because some prankster could have put the flag down, but it is evidence.

Predictions become evidence for the hypothesis only after they are observed to be true. However, predictions might not follow deductively from the hypothesis, but may only be made probable by the hypothesis. This probability is known as the likelihood of the hypothesis relative to the evidence. Remember that the likelihood of a hypothesis is the probability of the evidence given the theory, not the probability of the theory given the evidence. In the special case in which the hypothesis logically entails the evidence, the likelihood of the hypothesis relative to the evidence reaches the highest possible value of 1.

We are mainly interested in cases in which evidence supports, but does not prove, the hypothesis. It might seem natural that the degree of inconclusiveness in this evidential relation is measured by the probability of the hypothesis given the evidence. Bayes’ theorem tells us why this should not be so (see Hawthorne, this volume). For it tells us that the probability of a hypothesis given the
evidence depends in part on the prior probability of the hypothesis, and the prior probability is not part of the (empirical) evidence. Bayes’ theorem states that

\[ P(H|E) = \frac{P(H)P(E|H)}{P(E)} \]

where the evidence is denoted by \( E \), \( P(H|E) \) is the posterior probability of the hypothesis, \( H \), \( P(H) \) is its prior probability, \( P(E|H) \) is the likelihood of \( H \) relative to \( E \), and \( P(E) \) is a prior probability of the evidence. We need to distinguish between the degree to which the hypothesis is believed, which is \( P(H|E) \), and the degree to which the hypothesis is supported by the evidence \( E \), which is the likelihood \( P(E|H) \). Yet, it is not clear that this is how the distinction should be expressed, for there are cases in which the evidence for a hypothesis seems very strong, but the likelihood is very small.

For example, consider the following case. Suppose that we randomly select a coin from a box containing two coins, one of which is fair, and the other is biased \( \frac{2}{3} \) towards heads. The evidence is that we have tossed the coin 2,000 times and 1,003 of the tosses landed heads, and the hypothesis is that the coin is fair. The posterior probability \( P(H|E) \) is very high (close to 1), but the likelihood, \( P(E|H) \), is extremely low (about \( 8.7 \times 10^{-603} \)). So, how can we say that the evidence for \( H \) is strong? Well, the point is that while \( 8.7 \times 10^{-603} \) is a low number, the likelihood of the alternative hypothesis, that the coin is \( \frac{2}{3} \) biased towards heads, has a likelihood that is even lower. In fact, it is lower by a factor of \( 1.776 \times 10^{50} \). As Sober (1990) and many others have pointed out, an example like this is not reason to think that evidence involves prior probabilities. The lesson of the example is that we must think of the evidence relation as being comparative. That is, the evidence better supports the hypothesis that the coin is fair, compared to the hypothesis that the coin is biased \( \frac{2}{3} \) towards heads.

To make this even clearer, consider the same case in which the coin is randomly selected from a box of \( 1.776 \times 10^{53} \) coins, all of which are biased towards heads by \( \frac{2}{3} \) except for just one coin that is fair. Now consider the evidence \( E \) to be the same as before – that 1,003 tosses out of 2,000 tosses land heads. Does the evidence support the hypothesis that the coin is fair? Yes, it overwhelmingly supports that hypothesis compared to the alternative hypothesis that the coin is biased towards heads by \( \frac{2}{3} \). Should we believe that the coin is fair? No, definitely not, because the posterior probabilities favour the hypothesis that the coin is \( \frac{2}{3} \) biased towards heads by a factor of 1,000. We may wish to say that the total evidence supports the hypothesis that the coin is biased, if we include in the total evidence the fact that there is a miniscule proportion of fair coins in the box, but we do not want to say that \( E \) better supports the hypothesis that the coin is biased.
Table 1  Test for a rare disease.

<table>
<thead>
<tr>
<th></th>
<th>Have disease</th>
<th>Don’t have disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>100</td>
<td>100,000</td>
</tr>
<tr>
<td>Test negative</td>
<td>1</td>
<td>10,000,000</td>
</tr>
</tbody>
</table>

These considerations are important in cases in which there are reliable tests for a disease that is extremely rare. If $E$ is the outcome that your test is positive, and $P(E|H) = 99\%$, then in this sense, the test result provides strong evidence that you have the disease, as opposed to not having the disease. But it is not reason to believe that you have the disease, as the numbers in Table 1 clearly demonstrate. So long as the probability of a false positive is non-negligible (say $P(E|not-H) = 1\%$), then there will be many more false positives than true positives. So, the probability that someone is a true positive given that they are positive will be low.

In summary, if we are comparing two hypotheses, then $P(E)$ drops out of consideration in the following way. Assuming that two hypotheses, $H_1$ and $H_2$, have the same evidence $E$,

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1)}{P(H_2)} \times \frac{P(E|H_1)}{P(E|H_2)}.$$  

In words, the ratio of the posterior probabilities is equal to the ratio of the prior probabilities times the ratio of the likelihoods. It is clear from this equation that the evidence $E$ only enters into the likelihood ratio. It is for this reason that it is commonly assumed that evidence affects our degrees of belief only via the likelihoods.

Bayesians could express their theory of evidence in terms of what is commonly called incremental confirmation. That is, $E$ is evidence for $H$ if and only if $P(H|E) > P(H)$.

$E$ is positive evidence for $H$ if and only if $P(H|E) > P(H)$.

It may appear that this evidence relation is not comparative, and that it has nothing to do with likelihoods. But this is wrong. First, note that it is a simple consequence of Bayes’ theorem that $P(H|E) > P(H)$ if and only if $P(E|H) > P(E)$, provided that the probabilities are non-zero and the conditional probabilities are well defined. Second, suppose that there are only two hypotheses, $H_1$ and $H_2$, that have non-zero prior probabilities. Then
The Continuum Companion to the Philosophy of Science

\[ P(E) = P(E|H_1)P(H_1) + P(E|H_2)P(H_2). \]

In this case, \( P(H_1|E) > P(H_1) \) if and only if \( P(E|H_1) > P(E|H_2) \). In the more general case, \( E \) will increase the probability of a hypothesis if and only if it has a greater likelihood than a weighted average of the likelihoods of the alternative hypotheses. Thus, the Bayesian theory of incremental confirmation is best understood as a theory of evidence.

Likelihoodist and Bayesian theories of statistical inference are premised on this view of evidence, which I will refer to as the likelihood theory of evidence (see Hacking 1965, Edwards 1987 and Royall 1997 for a defence of likelihoodism, and Howson and Urbach 2005 for an introduction to Bayesianism). Roughly, it is the view that the evidential support that observations provide for a hypothesis depends only on the likelihood of the hypothesis and the likelihoods of its rivals. This set of likelihoods is often called the likelihood function, or the likelihood profile. It is relatively uncontroversial that likelihoods do capture a great deal of what’s relevant to the relation between theory and evidence. The more controversial thesis is that nothing else is relevant to the relation between theory and evidence. The likelihood theory of evidence is therefore only a theory, and it has been challenged in the literature (Forster 2006, Norton forthcoming). A summary of the main reasons why a theory of scientific evidence must consider more than likelihoods is given in this chapter.

Sometimes it is unclear whether the observations themselves (the events or acts) are the evidence, or whether a statement of what is shown to be true by those observations constitute the evidence for a theory or hypothesis. Because we usually think of predictions as statements or propositions that are logically entailed by a hypothesis, and a prediction observed to be true is evidence, it is useful to think of evidence in terms of true statements. Whenever I use the term ‘evidence’ one can substitute ‘statement of evidence’.

Finally, it should be clear that this chapter focuses exclusively on observational or empirical evidence. Statements of evidence are statements that are established as being true on the basis of the observations. Typically, in science, the paradigmatic examples of empirical evidence will be established by the outcome of an experiment: 75% of the patients lowered their LDL cholesterol levels by 20 points after daily doses of 1,800 mg of red yeast rice for 3 months, for example. The idea is that evidence is something that we can take as given; it is the foundation upon which our more theoretical knowledge is built. Nevertheless, there are many philosophical challenges to the assumption that there is a theory-neutral observation language in terms of which evidence is presented (e.g. Kuhn 1970). For example, to say that correct predictions of lunar eclipses are evidence for the theory that eclipses are caused by the earth occluding the light from the sun seems to be biased by the fact that the meaning of the
word ‘eclipse’ is laden with the theory to be justified. But it is also clear that the evidence could be redescribed in a vocabulary that is at least neutral \textit{with respect to the theory under consideration}, even if it is not neutral with respect to all possible theories.

An alternative view that meshes with this neutrality holds that evidence should not go beyond what we know of our subjective mental states, which imply nothing about the world outside our heads. Williamson (2000) refers to this as the \textit{phenomenal conception of evidence}. The following version of the evil demon argument could be considered to be an argument for this conception of evidence. A team of evil scientists abduct you during the night while you are sleeping, surgically remove your brain and place it in a vat with saline solution wired up so that when you awake, everything appears to you to be completely normal. Your brain send impulses to the muscles in your legs, and it appears to you that your legs respond in the expected way; the door appears to looming towards you and it looks and feels as if your hand grasps the door handle. Based on this evidence, you believe that you’re opening the bedroom door, but you are doing no such thing. You are a brain in a vat. The story is constructed in such a way that you have exactly the same beliefs as you do when you really do get up in the morning and believe that you are opening the door. But do you have the same evidence in both cases? One intuition says that if you are justified in your belief in the good case, then you are justified in your belief in the bad case as well, and that’s \textit{because} your evidence is exactly the same. That is the phenomenal conception of evidence.

One person’s \textit{modus ponens} is another person’s \textit{modus tollens}; for example, Williamson (2000) thinks that the beliefs are not justified in the bad case, and therefore rejects the phenomenal concept of evidence. After all, the phenomenal conception of evidence is inconsistent with the widely accepted idea that evidence can be shared by a community of individuals. I see this point as being related to William Whewell’s (1858, 1989) idea that the line between Fact and Theory moves over time; ‘Fact’ is Whewell’s word for evidence. For Whewell, human knowledge is formed from an inseparable union of subjective and objective elements – the mind, on the one side, provides the subjective elements, and the world, on the other side, provides the objective elements. Evidence lies on the objective side of the antithesis, being delivered to us by the world, while theories are supplied by the mind. But the \textit{dividing line} between them changes over time. For example, at the time of Copernicus, it was merely a theory that the Sun and the planets moved in a three-dimensional spatial configuration, albeit a widely accepted theory; the observational evidence at the time consisted of the two-dimensional positions of the planets relative to the fixed stars. But as soon as the Earth’s motion was known, it was possible to calculate the three-dimensional position of the planets relative to
the Sun by triangulating the observations of a celestial body at different positions in the Earth’s orbit around the Sun, provided only that the motion of the earth in the intervening period is known. This is exactly what Kepler did (Hanson 1973, Forster 1988, in press). Thus, Kepler’s evidence for the elliptical orbit of Mars consisted of a set of discrete three-dimensional positions of Mars relative to the Sun.

An unanswered question is: When does theory become fact? Or, in other words, when do theoretical claims become sufficiently well established that they can serve as evidence for deeper theories? You won’t answer this question by analysing the meaning of the word ‘evidence’. It would require a more detailed examination of the relationship between theory and evidence.

Suppose, for the sake of argument, that Whewell is right; the division between theory and evidence can change over time. That allows that at one time, evidence consists only of subjective experiences, while at a later time, it includes facts that are about objects in the outside world, facts about readings on measuring instruments and the images on a photographic plate. That certainly denies the phenomenal conception of evidence, but it does so on grounds that are independent of the evil demon example or the brain-in-a-vat example. In particular, consider the lives of two twins, one living an ordinary life, and one in the vat, who are having exactly the same phenomenal experiences. How can it be that evidence for the vat-twin ever comes to include ‘facts’ about the world beyond the phenomenal experiences of that twin? Suppose that the vat-twin’s evidence was purely phenomenal prior to the abduction, but after becoming a brain-in-a-vat, he or she pursues a scientific career and achieves some Copernican-like advances in knowledge. If the twin’s evidence becomes non-phenomenal, it clearly does not refer to the same external world as in the case of the normal twin.

It is not my purpose to examine all of these questions here. My purpose is to point out that there is a broad range of issues that deserve careful consideration. Moreover, I think that these issues should only be tackled in conjunction with another set of issues that have not been touched upon very much in the mainstream philosophical literature. We have just raised some questions about what evidence is, but we should also raise more questions about what evidence is used for. The mainstream view is that evidence is that which justifies belief. Even that view is controversial in the statistical inference literature. Royall (1997) is very often quoted as distinguishing carefully among three questions that you might want to address when evaluating the testimony of the observations:

(1) What should you do?
(2) What should you believe?
(3) What do the observations tell you about the hypotheses you’re considering?
Question (1) is often formulated in terms of the acceptence or rejection of hypotheses, and (2) is the one that Bayesians address. We have already seen that belief involves prior probabilities as well as likelihoods. Question (3) is different from (2) and also from (1); (3) is the question that Royall wants to address in terms of a likelihood theory of evidence. Having touched upon (2), if only briefly, it is question (3) that I shall focus on in this chapter.

Even when evidence is used to justify a belief, it doesn’t have to be a belief in the truth of a theory. Instead, or in addition, the evidence might justify belief in the predictive accuracy of a theory. This point is relevant to the evil demon example, for the example is constructed so that the phenomenal experiences of the two twins are the same; therefore, the predictive accuracy of the beliefs of the abducted twin is the same as the predictive accuracy of the beliefs of the normal twin, provided only that the predictions refer to phenomenal experiences. Yes, the vat-twin has false beliefs about the external world, but does not have false beliefs about the internal states she would have if she tried to get up and open the door. In this respect, the vat-twin and the normal twin are in the same epistemic situation. Certainly, predictive accuracy is a less highly esteemed achievement than full-blown truth, but it is surely far from trivial. In fact, if the phenomenal conception of evidence is correct, the evil twin example may prove that it is the only epistemic goal that humans are able to achieve on the basis of the evidence available to them.

Again, it is not my intention to try to address all these questions here. My thesis is twofold: (a) Evidence for predictive accuracy is an important transitional step towards obtaining evidence for the truth of a theory, and (b) in light of that, theories of evidence should consider both goals. A common fault with traditional discussions is that they tend to consider examples that do not mirror the structure of real scientific examples. The problem with Bayesianism is that it does not consider evidence for the predictive accuracy of a scientific theory. The remainder of this chapter will develop a series of examples designed to show that, and to explain why this is an important omission for any theory of scientific evidence.

2. Unification and Evidence

2.1 The Beam Balance Example

Suppose we hang an object, labelled $a$, on one side of a beam balance (Figure 1), and find how far a second object $b$ must be hung on the other side to balance the first. The distance that $a$ is hung from the point at which the beam is supported (called the fulcrum) is labelled $x$, while the distance to the right of the fulcrum at which $b$ balances $a$ is labelled $y$. If $b$ is moved to the left of this point, then the beam tips until the object $a$ rests on the ground;
if $b$ is moved to the right, the beam tips the other way. In the first instance, $x$ is 1 centimetre (cm), and the beam balances when $y$ is 1 cm. This pair of numbers is a datum, or a data point. We repeat this procedure two more times with different values of $x$,

![Figure 1](image.png)

and tabulate the resulting data in Table 2.

### Table 2

Suppose we observe that the beam balances when $y$ is 1 cm if we set $x$ to 1 cm, 2 cm if we set $x = 2$ cm, and 3 cm if we set $x = 3$ cm.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cm</td>
<td>1 cm</td>
</tr>
<tr>
<td>2 cm</td>
<td>2 cm</td>
</tr>
<tr>
<td>3 cm</td>
<td>3 cm</td>
</tr>
</tbody>
</table>

We now ask: If $a$ were hung at a distance of 4 cm to the left of the fulcrum, at what distance would $b$ balance; viz. predict $y$ from $x$. Once we notice that for all three data, $x$ is equal to $y$, it appears obvious that we should predict that $y$ will be 4 cm when $x$ is 4 cm. However obvious it may be that we can make predictions from observations alone, it is widely accepted that observations do not speak for themselves. It is only in conjunction with a hypothesis that observations make predictions.

Once we know what quantity we want to predict (represented by the dependent variable $y$) and what information we think is relevant in the context to make the prediction (the independent variable $x$), one needs a formula, or model (an equation or equations with at least one adjustable parameter). In our example, it is $y = \beta x$, where $\beta \geq 0$. The hypothesis, or model, is represented by a family of curves in the $x$-$y$ plane, in this case, the family of straight lines with non-negative slope that pass through the origin. A statistical
procedure, such as the method of least squares, can now be used to estimate the value of the free parameter $\beta$. Its estimated value is the slope of the line in the family that best fits the data in the sense defined by the method of least squares. This picks out a unique curve from the family, sometimes called the fitted model, which can be used to make precise predictions.

2.2. Empirical Success and Predictive Accuracy

Suppose we divide the total data into two subsets; call them the training data and the test data. Define the degree to which a model succeeds in predicting the test data from the training data in the following way. First, find the curve that best fits the training data according to some definition of fit, such as the sum of the squared residues, or the likelihood (for further details, see Forster 2007). Now see how well the curve that best fits the training data succeeds in fitting the test data. The number obtained is a measure of empirical success. Clearly, there are a number of ways of measuring empirical success, depending on how the total data is divided into training sets and test sets, and how goodness-of-fit is defined, and also on how such numbers might be averaged together to produce new measures of empirical success. All these numbers are measures of empirical success because they are defined solely in terms of the empirical data in hand. They are not defined in terms of predictions that have not yet been checked, or in terms of degrees of simplicity or unification or any other non-empirical virtue.

Predictive accuracy, on the other hand, is used here to denote a property of a hypothesis that goes beyond the current evidence. In terms of the previous example, the empirical success of the model $y = \beta x$ consists in its successful prediction of any two data in Table 2 from the third datum. Its predictive accuracy, in contrast, includes its ability to predict that $y$ will be 4 cm if $x$ is 4 cm, even though that observation has not yet been made. Clearly, hypotheses do not wear their predictive accuracy on their sleeves. In this respect, predictive accuracy is like truth. The empirical success of the model is evidence for assertions about the predictive accuracy of the model, where the strength of evidence depends on the particular kind of predictive accuracy under consideration.

If we define the sense in which predictions are approximately true in terms of some statistical measure of goodness-of-fit, then we can define the degree of predictive accuracy as the average goodness-of-fit achieved by the model fitted to the training data in some hypothetical set of new data generated by whatever mechanism generated the original set of data. Just as there are many measures of empirical success, there are also many kinds of predictive accuracy, depending on which hypothetical set of new instances we wish to consider and what we use as the training data. Irrespective of whether our beliefs
are also true, choosing models with high degrees of predictive accuracy is an important way in which evidence is used in science and in everyday life.

2.3 Instrumentalist and Realist Goals

There are two main goals of scientific inference that philosophers have discussed. One is based on the instrumentalist view that the goal of science is to maximize predictive accuracy. Another is based on the realist view that the goal of model selection and theory choice is to represent the reality (the truth) underlying the observed phenomena. Both of these goals are legitimate and important in their own right. They are different goals, but they are importantly related. First, any model that correctly represents the underlying reality is true, and therefore all its logical consequences are true, and so all its predictions are true. But does this mean that true hypotheses are predictively accurate? Surprisingly, the answer is NO, as is clearly seen by considering a logical truth that makes no empirical predictions, and therefore has no empirical success and no degree of predictive accuracy. Truth, to the extent that it is a worthy epistemic goal, has to be informative truth, and in this respect instrumentalist goals are important, even to a realist.

2.4 Unification

In an application of the beam balance model to a single pair of objects \{a, b\}, the following two equations differ only in how the adjustable parameter is represented:

\[
y = \frac{m(a)}{m(b)} x, \quad y = ax.
\]

The reason that the equations are empirically equivalent in the given context is that they parameterize exactly the same family of curves; namely all the straight lines with positive slope that pass through the origin. The difference between them is merely linguistic. But if we think of the model more broadly, for example, as applying to three pairs of objects, \{a, b\}, \{b, c\} and \{a, c\}, then this changes. UNI stands for UNIfied model, and DUM stands for DisUnified Model. Now there are three equations in each model.

**UNI:** \(y_1 = \frac{m(a)}{m(b)} x_1, \quad y_2 = \frac{m(b)}{m(c)} x_2, \quad \text{and} \quad y_3 = \frac{m(a)}{m(c)} x_3.\)

**DUM:** \(y_1 = \alpha x_1, \quad y_1 = \beta x_2, \quad \text{and} \quad y_3 = \gamma x_3.\)
Each equation is a sub-model in a broader model. The surprising fact is that
the broader models are no longer empirically equivalent, even though their
sub-models are empirically equivalent. So, in the case of UNI, there is a sense
in which the model is greater than the sum of its sub-models.

This is not because UNI is derivable from Newton’s theory, while DUM is
not. In fact, there is a clear sense in which UNI entails DUM because UNI is
‘nested’ in DUM; UNI is a special case of DUM in which the constraint \( \gamma = \alpha \beta \)
holds necessarily. This is because the third mass ratio, \( m(a)/m(c) \) is equal to the
product of the other two mass ratios according to the mathematical identity:

\[
\text{CONSTRAINT: } \frac{m(a)}{m(c)} = \frac{m(a)}{m(b)} \frac{m(b)}{m(c)}.
\]

Since UNI entails DUM, and Newton’s theory plus certain auxiliary assump-
tions entail UNI, then by the transitivity of entailment, Newton’s theory plus
the same auxiliary assumptions entail DUM. Both models are consequences of
Newton’s theory.

UNI is logically stronger than DUM, but not merely in the sense in which
X&Y is stronger than Y, where X is some irrelevant metaphysical assertion that
has no empirical consequences. UNI is \emph{empirically} stronger because it entails a
relational fact about the data that DUM does not. UNI makes testable
predictions that DUM does not make. To draw this out, suppose that both models are
true, and there is very little noise in the data or that the data set is large enough
that the noise tends to cancel out, so that the statistical estimation of the
parameters \( \alpha, \beta \) and \( \gamma \) is fairly accurate. Let \( \hat{\alpha}, \hat{\beta} \) and \( \hat{\gamma} \) denote these estimated
values. Then UNI predicts that \( \hat{\gamma} - \hat{\alpha} \hat{\beta} = 0 \), within the bounds of approximation
predicted by the statistical assumptions, and this prediction is something that
counts as \emph{observational} evidence. DUM does not make this prediction. If the
prediction is true, then there is evidence for UNI that is not evidence for DUM,
even though there is no evidence against DUM. If UNI is accepted in light of
the evidence, this is not a rejection of DUM, because UNI entails DUM. Yet,
there is a sense in which the observation that \( \hat{\gamma} - \hat{\alpha} \hat{\beta} \approx 0 \) is evidence for UNI
without being evidence for DUM.

Is it possible for Bayesianism to capture this idea? Certainly, any evidence
that is deductively entailed by a hypothesis gives a likelihood of one,
which cannot be lower than the likelihood of any alternative hypotheses.
Therefore, the probability of any hypothesis cannot be lowered by deductive
evidence. However, it’s not clear whether the probability of DUM will be
raised or lowered by ‘irrelevant evidence’, or will stay the same. That will
depend on what its rivals are, and it will also depend on how the likelihood of
DUM is determined, since any model is the disjunction of more specific
hypotheses, and the likelihood of a disjunction is an average of the likelihoods of the disjuncts. Also, the fact that \( \gamma - \hat{\alpha} \hat{\beta} = 0 \) is only one part of a larger body evidence, and it’s not clear how or whether a Bayesian should separate one piece of evidence from the total evidence. This will come up again in a later example.

2.5 The Variety of Evidence
In order to show that UNI is better supported by the evidence than DUM, we need values of \( \hat{\alpha}, \hat{\beta} \) and \( \hat{\gamma} \) estimated from the data. This requires that the evidence is sufficiently varied or diverse. There may well be other reasons why the variety of evidence is important in science, but this particular reason has not received much emphasis in the philosophical literature, despite the fact that the illustrative example used by Hempel (1966, chapter 4) involves three applications of Snell’s law, which have exactly the same logical structure as the beam balance example used here.

2.6 Evidence for Component Causes
Cartwright (1983) presents a sceptical argument against the existence of component causes. The intuitive argument can be presented in the following way (Forster 1988). The coffee cup sitting on the table is motionless because it is subject to two component forces, which cancel each other out. One is the weight of the cup acting downwards, and the other is an equal and opposite force of the table on the cup acting upwards. But what we see is consistent with the magnitude of the component forces being of any magnitude, provided that they cancel each other out. The principle that a cause is measured by its effect therefore dictates that only the resultant force can be measured by its observed effect. There is no empirical justification for believing in the existence of component causes.

A detailed reply to this argument is found in Forster (1988), but a version of the beam balance example illustrates the essential idea. Suppose that a unit mass \( u \) is hung on the right side of the beam balance at a distance \( y \), while \( a \) and \( b \) are hung together on the right at a distance \( y \), such that the beam balances. The simplest Newtonian model is: \( m(u)y = (m(a) + m(b))x \), where \( m(u) = 1 \). We can also simplify the example by assuming that \( a \) and \( b \) are hung at a unit distance from the fulcrum, so that \( x = 1 \). Then the equation becomes \( y = m(a) + m(b) \). Within this narrow context, Cartwright’s point is correct: There is no way of estimating the values of \( m(a) \) and \( m(b) \) separately; we can only estimate their sum. But if we repeat the same experiment with \( b \) and \( c \), and then with \( a \) and \( c \), the measurements of the resultants provide an estimate of the component masses, for it provides three equations in three unknowns. So, if the
context is widened, to include 11 objects, which can be arranged in 55 different pairs, we have, in effect, five independent measurements for each mass. Not only are the component causes measured by their effects, but they are independently measured by the effects many times over. Cartwright’s argument is importantly wrong.

Cartwright (1983) appears to believe that such replies rely on a linearity assumption, or in this case, on the additivity of masses. But the counterexample only assumes that there is a sufficient number of equations to solve for the unknowns, which ensures that there are independent measurements.

2.7 Realism
It is now clear that unification can increase predictive power, and when the predictions are borne out by the data, they are evidence for the model. However, it is also clear that this predictive advantage could be replicated by disunified models if we add constraints to them. Thus, refer to DUM plus the constraint \( \gamma - \alpha \beta = 0 \) as DUM+. Now, it is easy to see that DUM+ makes all the same predictions, yet there is a sense in which DUM+ does not explain why it should be true that \( \dot{\gamma} - \dot{\alpha} \dot{\beta} = 0 \). DUM+ entails the prediction that \( \dot{\gamma} - \dot{\alpha} \dot{\beta} = 0 \), but does not explain it. It merely asserts it as a brute fact. UNI, on the other hand, explains it, or at least Newton’s theory explains it by explaining how the constraint arises from the fact that each coefficient is a ratio of mass values. That is, the existence of mass properties and the mathematical identity

\[
\frac{m(a)}{m(c)} = \frac{m(a)m(b)}{m(b)m(c)}
\]

explains the contingent fact that \( \dot{\gamma} - \dot{\alpha} \dot{\beta} = 0 \). Clearly, the falsity of the prediction could not refute the mathematical identity, so it is therefore only the existence of masses that is tested, and therefore that proposition is the one supported by the observations. It is hard to see how Bayesianism could begin to capture such ideas, since UNI and DUM+ are identical in their likelihoods.

2.8 The Theoreticians’ Dilemma
In an article with this title reprinted in Hempel (1965), Hempel argues that there is, in at least some instances, a third model that is predictively equivalent to UNI and DUM+. This is a model that eschews the use of theoretical parameters altogether. DUM+ does not do this because there is a sense in which \( \alpha, \beta \) and \( \gamma \) could be understood as representing some property of pairs of objects. No adjustable parameter is explicitly defined in terms of observational
quantities. But given the statistical method of estimation that fixes their values, there is a sense in which these parameters could be replaced by their statistical estimates, $\hat{a}$, $\hat{b}$ and $\hat{\gamma}$, which are expressed solely in terms of observational quantities (the $x$'s and the $y$'s). They are complicated functions, but it is possible in principle. Let us call the third model CRAIG. The substitution does not preserve the meaning of the original model, but it would replicate its empirical success. At first sight, the existence of CRAIG is an embarrassment to a realist because CRAIG cannot represent the world behind the observed phenomena – for it contains no theoretical terms. But it need not be an embarrassment to the kind of realist who is committed only to explaining the empirical success of CRAIG, and who sees nothing wrong with using exactly the same explanation used to explain the empirical success of UNI. After all, CRAIG has been constructed in order to produce the same empirical success. Such examples are only telling against the kind of realism that holds that every theory or hypothesis must explain its own empirical success. Neither DUM+ nor CRAIG can explain their own empirical success, but why should that worry a realist? A sophisticated realist is not committed to simply ‘reading off’ a realist interpretation from the best scientific explanation. She is free to read in an interpretation (see Psillos, this volume). So, if inference to the best explanation refers to the best explanation of the phenomena currently held by scientists, then this may tie the realist’s hands unnecessarily. Nor is the realist obliged to interpret every part of a theory in a realist way. Realists should be selective about what is interpreted to be real. (Friedman 1981, 1983; Forster 1988; Humphreys 2004).

One worry for a realist may be that the notion of explanation being appealed to is not very easy to capture (see De Regt, this volume). Hempel’s deductive-nomological theory of scientific explanation, also known as the covering law model of explanation (‘Studies in the logic of explanation’ in Hempel 1965), can claim that the deductive entailments offered by DUM+ and CRAIG are not genuine explanations because they do not appeal to genuine laws of nature. But then the notion of a law of nature is hard to capture, and it’s also not clear that UNI does appeal to a law of nature. Nevertheless, there are strong intuitions about explanation, inference, evidence and realism that are not easy to dismiss.

The same intuitions convince us of the existence of an everyday world of three-dimensional objects. Why is it that we take for granted that the everyday world consists of three-dimensional objects, when all the visual evidence we have for that conclusion rests on the visual system’s analysis of two-dimensional projections onto our retinas? A plausible answer is that the third dimension, depth, is independently measured by various visual cues, and all the measurements agree. Such a coincidence provides evidence for existence of three-dimensional objects in the same way that the observations in the beam
balance example support claims about the existence of mass. It seems that we should either deny the realist conclusion in neither case or in both cases. There does not appear to be any solid middle ground.

3. Truth Versus Predictive Accuracy

I shall now present a concrete example to show that the realist goal of truth and the instrumentalist goal of predictive accuracy are not only different, but also sometimes incompatible – optimizing one does not optimize the other. First, imagine that we do beam balance experiments with the 55 possible pairs of objects drawn from a set of 11 objects. Suppose we consider the simpler situation in which they are hung together in pairs on the right at a unit distance in order to balance a unit mass hung at a variable distance \( y \) on the left. Our beam balance model consists of 55 equations, where each equation has the form \( y_1 = m(a) + m(b) \). Since each of the numbers \( y_1, y_2, \ldots, y_{55} \) is known, there are 55 equations for 11 unknown mass values. This overdetermination of the unknowns means that there are five independent measurements of each mass, and if any one of these disagree, then the model is refuted by the data.

DUM, on the other hand, also has 55 equations, but its equations are of the form \( y_1 = m(a^*b) \), where \( a^*b \) is the name of the composite object consisting of \( a \) paired with \( b \). DUM does not assume the constraint \( m(a^*b) = m(a) + m(b) \), so it has 55 unknowns. There is no overdetermination of its parameters, even though they are ‘measured by their effects’, so the data provides no evidence for the existence of these properties. (Cartwright’s (1983) conclusion is turned on its head in this example – the only evidence for the ‘resultant’ masses \( m(a^*b) \) is via the evidence for the component masses \( m(a) \) and \( m(b) \).)

Now change the example slightly by supposing that in the first experiment, in which \( a \) and \( b \) are hung together, there is a heavy lead hanger used that increases the value of \( y_1 \) to above the sum of the masses of \( a \) and \( b \). UNI is not only false, but the evidence shows that it is false, so its posterior probability plummets to zero. DUM, on the other hand, is able to accommodate the data without any problem. So, in this situation, UNI is refuted, but DUM survives. If we are only interested in the probability of truth, then there is nothing wrong with this conclusion. In fact, it is the correct conclusion. The evidence shows that there is no probability that UNI is true, whereas it is possible that DUM is true.

If we are interested in the accuracy of ‘next-instance’ predictions, then DUM is also the correct choice. For, assuming that the lead hanger will still be present in any repetitions of the experiment with \( a \) and \( b \), then the predictions that DUM makes will be exactly right. Standard model selection criteria, such as AIC (Forster and Sober 1994), will correctly favour DUM in such a situation,
if their goal is to maximize predictive accuracy in this sense, or if their goal is
to maximize the probability of truth.

But is this the end of the story about scientific evidence? Are the goals of
truth and next-instance prediction the only goals that need to be considered?
Is there evidence for something else, which is also important in science?
Whewell (1858, 1989) famously made a distinction between prediction of facts
of the same kind and the prediction of facts of a different kind. That distinc-
tion applies here. UNI is succeeding in predicting facts in one experiment
on the basis of facts in another experiment. In particular, if we exclude the
‘erroneous’ experiment in which the lead hanger is present, then all 11 masses
have four or more independent measurements that all agree. Independent
measurements can alternatively be seen as predictions of each other. This kind
of prediction is different from next-instance prediction. Even though the total
evidence demonstrates that UNI is false, the partial agreement of independent
measurements is evidence that UNI is capturing something that is approxi-
mately or partially true, or pointing towards the truth in some way. DUM, on
the other hand, has not shown any success in predictions whatsoever. It has
merely accommodated the data. Surely, we do not want to end the story by
saying that the observations decrease the probability of UNI and increase the
probability of DUM. There is more to be said about evidence than that.

This example makes it plausible that even in the example in which both
UNI and DUM accommodate the data perfectly (neither hypothesis is refuted),
and both have their probabilities increased (because rival hypotheses are
refuted), we should be suspicious of saying that the observations provide
evidence for DUM. For none of its predictions have been tested! UNI, on the
other hand, has gained support not merely for the reliability of its predictive
content, but also for its claim that objects really have mass properties.

4. Conclusion

I have argued for two conclusions. One is that a Bayesian notion of evidential
support involves only likelihoods, despite appearances to the contrary.
Having excluded prior probabilities from consideration, I then went on to
argue that we need to consider more than likelihoods if we wish to under-
stand the evidence for the existence of mass properties, for example, or even
for the existence of three-dimensional objects. The examples given were spe-
cifically designed to clarify realist intuitions about evidence. I maintain that
any good philosophical theory of evidence should at least address these intu-
tions, in some way or another. Either we must argue that these intuitions are
misleading or wrong, or we should abandon a strictly Bayesian or Likelihood-
ist theory of evidence. Personally, I would prefer to explore the second option,
but in either case there is more to be said about evidence and the way it features in explanation and the debate over scientific realism.

Notes

1 Anyone interested in an excellent introduction to the evidence for Einstein's theory of relativity should read Will (1986).
2 It may depend on other empirical evidence, but eventually, it must depend on an a priori probability.
3 The same point applies to van Fraassen's notion of empirical adequacy, which refers to observations that will be made in the future as well as those made in the past. The difference between predictive accuracy and empirical adequacy is that van Fraassen's notion requires that predictions are exactly true, whereas predictive accuracy is well defined in cases in which predictions are false, but approximately true. Approximate truth is notoriously hard to define, as a general notion, but statisticians routinely use notions of approximate goodness-of-fit, and that is all the predictive accuracy requires; see Forster and Sober 1994.

References

Forster, M. R. and Sober, E. (1994), ‘How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions’, British Journal for the Philosophy of Science, 45, 1–35.


11 Bayesian Confirmation Theory

James Hawthorne

1. Introduction

Scientific theories and hypotheses make claims that go well beyond what we can immediately observe. How can we come to know whether such claims are true? The obvious approach is to see what a hypothesis says about the observationally accessible parts of the world. If it gets that wrong, then it must be false; if it gets that right, then it may have some claim to being true. Any sensible attempt to construct a logic that captures how we may come to reasonably believe the falsehood or truth of scientific hypotheses must be built on this idea. Philosophers refer to such logics as logics of confirmation or as confirmation theories.

Among philosophers and logicians, the most influential contemporary approach to the logic of the hypothesis confirmation is Bayesian Confirmation Theory. This approach employs probability functions to represent two distinct things: what a hypothesis says about how likely it is that specific evidential events will occur, and how strongly the hypothesis is confirmed or refuted by such evidence. The hypothesis-based probabilities of evidence claims are called likelihoods. When the evidence is more likely according to one hypothesis than according to an alternative, that increases the probabilistic degree of confirmation of the former hypothesis over the later. Any probabilistic confirmation theory that employs the same probability functions to represent both the likelihoods hypotheses confer on evidence claims and the degrees to which hypotheses are confirmed by these evidence claims will be a Bayesian confirmation theory, because a simple theorem of probability theory, called Bayes’ Theorem, expresses precisely how these likelihoods contribute to the confirmational probabilities of hypotheses.

This chapter describes the essential features of Bayesian confirmation theory. Section 2 presents the probabilistic axioms for confirmation functions. Section 3 describes how this logic is applied via Bayes’ Theorem to represent the evidential support of hypotheses. Section 4 draws on a Bayesian Convergence Theorem to show why this logic may be expected to refute false hypotheses.
and support true ones. Section 5 generalizes the treatment of Bayesian likelihoods described in Section 3. Section 6 concludes by briefly commenting on what confirmation functions are conceptually.

2. The Axioms for Confirmation Functions

A confirmation function is a binary function, \( P_\alpha[A \mid B] \), on sentences of a language capable of expressing scientific hypotheses and theories. Logicians make this idea precise by taking the language and its deductive logic to be that of predicate logic (including the identity relation) because that language is known to have the expressive power needed to represent the deductive logic of any scientific theory. Such a language possesses a non-logical vocabulary consisting of names (and variables) and predicate and relation terms, and a logical vocabulary consisting of the standard logical terms: ‘\(~\)’ for ‘not’, ‘\(\cdot\)’ for ‘and’, ‘\(\lor\)’ for ‘inclusive or’, ‘\(\supset\)’ for truth-functional ‘if-then’, ‘\(\equiv\)’ for ‘if and only if’, ‘\(\forall\)’ for ‘all’, ‘\(\exists\)’ for ‘some’ and ‘\(=\)’ for the relation ‘is the same thing as’. This language permits the expression of any scientific theory, including set theory and all the rest of mathematics employed by the sciences.

The axioms for confirmation functions are essentially semantic rules that constrain each possible confirmation function to respect the meanings of the logical terms (not, and, or, etc.), much as the axioms for truth-value assignments in the semantics for deductive logic constrain each possible truth-value assignment to respect the meanings of the logical terms. These rules don’t determine which confirmation functions are correct (just as the semantic rules for truth-value assignments don’t determine which way of assigning truth-values to sentences captures the actual truths). The correctness of various measures of confirmation may depend on additional considerations, including what the non-logical terms and sentences of the language mean.

Here are the axioms for the confirmation functions, treated as semantic rules on an object language \( L \) that’s powerful enough to express any scientific theory.

Let \( L \) be a language whose deductive logic is predicate logic with identity – where ‘\( C \vdash B \)’ abbreviates ‘\( C \) logically entails \( B \)’, and ‘\( \vdash B \)’ abbreviates ‘\( B \) is a tautology’. A confirmation function is any function \( P_\alpha \) from pairs of sentences of \( L \) to real numbers between 0 and 1 that satisfies:

1. \( P_\alpha[D \mid E] < 1 \) for some \( D, E \);

2. if \( B \vdash A \), then \( P_\alpha[A \mid B] = 1 \);

for all \( A, B, C \),
3. If $|=(B=C)$, then $P_\alpha[A|B] = P_\alpha[A|C]$;

4. If $C|=-(B\land A)$, then either $P_\alpha[(A\lor B)|C] = P_\alpha[A|C] + P_\alpha[B|C]$ or, for every $D$, $P_\alpha[D|C] = 1$;

5. $P_\alpha[(A\cdot B)|C] = P_\alpha[A|(B\cdot C)] \times P_\alpha[B|C]$.

Each function satisfying these rules is a possible confirmation function. The subscript ‘$\alpha$’ reminds us that many alternative functions $\{P_\beta, P_\gamma, \ldots\}$ obey these rules. All the usual theorems of probability theory follow from these axioms.

Some Bayesian logicians have explored the idea that, like deductive logic, a logic of confirmation might be made to depend only on the logical structures of sentences. It’s now widely agreed that this project cannot be carried out in a plausible way. The logic of confirmation must also draw on the meanings of the sentences involved. Thus, one should also associate with each confirmation function $P_\alpha$ an assignment of meanings to non-logical terms, and thereby to sentences. This suggests two additional axioms (or rules).

6. If $A$ is analytically true (given the meanings that $P_\alpha$ associates with the language) or an axiom of set theory or pure mathematics employed by the sciences, then $P_\alpha[A|B] = 1$ for each sentence $B$.

It follows that if a sentence $C$ analytically entails $A$, then $P_\alpha[A|C] = 1$.

When a contingent sentence $E$ is considered certainly true, it’s common practice to employ a confirmation function that assigns it probability 1 on every premise $B$. This saves the trouble of writing $E$ as an explicit premise, because when $P_\beta[E|C] = 1$, $P_\beta[H\land E|C] = P_\beta[H|C]$. But writing $P_\beta[H|C] = r$ instead of $P_\beta[H|E\cdot C] = r$ when one is certain of $E$ hides the confirmational dependence of $H$ on $E$. This is a bad thing to do in a logic of confirmation, where the whole idea is to provide a measure of the extent to which premise sentences indicate the likely truth of conclusion sentences. Omitting $E$ as an explicit premise makes the logic enthymematic. In deductive logic, one wouldn’t write ‘$(E\supset A)\models A$’ just because one is already certain that $E$. One shouldn’t do this in a logic of confirmation either. The logic should represent $E$ as maximally confirmed by every possible premise (including $\sim E$) only in cases in which $E$ is logically or analytically true or an axiom of pure mathematics. This motivates the Axiom of Regularity.

7. If $A$ is neither logical nor analytically true, nor a consequence of set theory or some other piece of pure mathematics employed by the sciences, then $P_\alpha[A|\sim A] < 1$. 

199
Taken together, axioms 6 and 7 imply that a confirmation function assigns probability 1 on every possible premise to precisely the sentences that are non-contingently true according to its associated meaning assignment.1

Perhaps additional axioms should further constrain confirmation functions. In particular, when hypotheses describe chance situations, a rule like David Lewis’s (1980) Principal Principle seems appropriate. Consider a hypothesis that says systems in state Y have objective chance (or propensity) \( r \) to acquire an attribute X: \( \text{Ch}(X,Y) = r \). When \( c \) is a system in state Y (i.e. \( c \in Y \)), this should give rise to a direct inference likelihood: \( P_{\alpha}[c \in X | \text{Ch}(X,Y) = r \cdot c \in Y] = r \).

One might add an axiom requiring confirmation functions to satisfy this principle. A general axiom of this kind should also specify precisely what sorts of information \( B \) may interfere with these likelihoods; that is, when \( P_{\alpha}[c \in X | \text{Ch}(X,Y) = r \cdot c \in Y \cdot B] \) should continue to equal \( r \), and for what \( B \) may degree \( r \) no longer hold. Spelling out such a direct inference likelihood axiom in full generality turns out to be quite difficult, so we’ll not pursue it now. But let’s keep in mind the idea that chancy claims in scientific theories should often lead to objective values for likelihoods, on which all confirmation functions should agree.

That’s the axiomatic basis for the logic of confirmation functions. However, this characterization leaves two important questions untouched:

(1) What, conceptually, is a confirmational probability function?; and
(2) Why should we consider a confirmation function to be a good way of measuring evidential support?

These issues can only be adequately addressed after we see how the logic applies evidence to the confirmation of hypotheses (for a contrasting account, see Forster, this volume). However, the subjectivist reading of Bayesian confirmation functions has become so prominent in the literature that I will say something about it before proceeding.

The subjectivist interpretation takes \( P_{\alpha}[A | B] \) to express the degree of belief (or confidence) an agent \( \alpha \) would have in \( A \) were she to become certain that \( B \), but possesses no other information that’s relevant to \( A \). Although this kind of belief-strength interpretation may be appropriate for probabilities in decision theory, it faces very significant difficulties as a way of understanding confirmation functions. I’ll discuss some of these difficulties later, but please forgo this reading for now, since it can be very misleading. Instead think of a confirmation function as a kind of truth-indicating index. Later I’ll bolster this idea with an account of how evidence can bring confirmation functions to point towards the truth-values of hypotheses. Because of this feature, confirmation functions should influence one’s belief-strengths regarding the truth of hypotheses, although they are not themselves measures of belief-strength.
3. The Bayesian Logic of Evidential Support

Let’s see how the logic of confirmation functions represents evidential support for scientific hypotheses. Let \(<H_1, H_2, \ldots, H_m, \ldots>\) be an exhaustive list of alternative hypotheses about some specific subject. The list may contain a simple pair of alternatives – for example, \(<\text{Joe is infected by HIV, Joe is not infected by HIV}>\) – or it may be a long list of competitors (e.g. of alternative ailments that may be responsible for the patient’s symptoms). The competitors may make claims about some single event (e.g. about what disease(s) afflicts Joe), or they may be grand theories (e.g. about what laws of nature govern the universe). The list of alternative hypotheses or theories may, in principle, be infinitely long. The idea of testing infinitely many alternatives may seem extraordinary, but nothing about the logic itself forbids it. The alternative hypotheses need not be entertained all at once. They may be constructed and assessed over millennia.

Practically speaking, in order for the list of competing hypotheses to be exhaustive, it may need to contain a catch-all hypothesis \(H_k\) that says none of the other hypotheses is true (e.g. ‘the patient has an unrecognized disease’). When only a finite number \(m\) of explicit alternative hypotheses is under consideration, the catch-all alternative \(H_k\) will be equivalent to the sentence that denies each explicit alternative. Thus, if the list of alternatives, including the catch-all, is \(<H_1, H_2, \ldots, H_m, H_k>\), then \(H_k\) is equivalent to \((\sim H_1 \land \ldots \land \sim H_m)\).

The evidence employed to test hypotheses consists of experiments or observations that each hypothesis says something about. In the older hypothetico-deductive account of confirmation, each hypothesis \(H_i\) speaks about observations by deductively entailing evidence claims. However, hypotheses cannot usually accomplish this on their own. They usually draw on statements \(C\) that describe the conditions under which evidential outcome \(E\) occurs. In addition, hypotheses often rely on background knowledge and auxiliary hypotheses \(B\) (e.g. about how measuring devices function) to connect them via experimental circumstances \(C\) to evidential outcomes \(E\). So the deductive logical relationship through which a hypothesis speaks about evidence takes the form: \(H_i : B \land C \models E\). If the observation condition and evidential outcome \((C \land E)\) occurs, this may provide good evidence for \(H_i\) provided \(B\) holds up. On the other hand, if \(C\) holds but \(E\) is observed to be false, then deductive logic alone gives us \(B \land C \models \sim E \models \sim H_i\) and hypothesis \(H_i\) is falsified by \(B \land C \models \sim E\). Thus a hypothesis \(H_i\) usually fails to entail evidential claims on its own, but only speaks about evidence deductively with the assistance of background and auxiliary claims together with descriptions of the experimental or observational circumstances. Similarly, when hypotheses speak about evidence probabilistically via likelihoods, conditions \(C\) and background \(B\) play a comparable enabling role.
In Bayesian confirmation theory, the degree to which a hypothesis \( H_i \) is confirmed on evidence \( C-E \), relative to background \( B \), is represented by the posterior probability of \( H_i \), \( P_a[H_i|B-C-E] \). Bayes’ Theorem shows how this posterior probability depends on two kinds of probabilistic factors. It depends on the prior probability of \( H_i \), \( P_a[H_i|B] \), and on the likelihood of evidential outcome, \( E \), according to \( H_i \) together with \( B \) and \( C \), \( P_a[E|H_i;B-C] \). Let’s consider the nature of each. Then we’ll see how they come together in the logic of hypothesis evaluation.

**Likelihoods.** Likelihoods express what hypotheses say about observationally accessible parts of the world. If a hypothesis, together with auxiliaries and observation conditions, deductively entails an evidence claim, axiom 2 guarantees that every confirmation function \( P_a \) assigns the likelihood value 1—that is, if \( H_i;B \cdot C \models E \), then \( P_a[E|H_i;B-C] = 1 \). Similarly, if \( H_i;B-C \not\models E \) and \( H_i;B-C \) is contingent, the axioms yield \( P_a[E|H_i;B-C] = 0 \). However, quite often the hypothesis \( H_i \) will only imply the evidence to some probabilistic degree. For instance, \( H_i \) may itself be an explicitly statistical or chance hypothesis, or \( H_i \) may be a non-statistical hypothesis that’s probabilistically related to the evidence by statistical auxiliary hypotheses that reside within background \( B \). In either case the likelihoods may be the kind of direct inference likelihoods described near the end of section 1. That is, when \( H_i;B \models Ch(X,Y) = r \) (the chances of acquiring \( X \) for systems having \( Y \) is \( r \)) and \( C \) is of form \( c \in Y \) and \( E \) is of form \( c \in X \), we should have \( P_a[E|H_i;B-C] = r \) (provided \( H_i;B \) doesn’t also entail some relevant defeater of this direct inference).\(^3\) Such likelihoods should be completely objective in that all confirmation functions should agree on their values, just as all confirmation functions agree on likelihoods when evidence is logically entailed. Functions \( P_a \) that satisfy the confirmation function axioms but get such direct inference likelihoods wrong should be discarded as illegitimate.

Not all scientific likelihoods are warranted deductively or by explicitly stated chance claims. Nevertheless, the likelihoods that relate hypotheses to evidence in scientific contexts will often have widely recognized objective or intersubjectively agreed values. For likelihoods represent the empirical content of hypotheses: what hypotheses say about the observationally accessible parts of the world. So the empirical objectivity of a science relies on a high degree of agreement among scientists on their values.

Consider what a science would be like if scientists disagreed widely about the values of likelihoods for important hypotheses? Whereas expert \( \alpha \) takes \( H_1 \) to say \( E \) is much more likely than does \( H_2 \) (\( P_a[E|H_1;B-C] \gg P_a[E|H_2;B-C] \)), her colleague \( \beta \) sees it in just the opposite way (\( P_a[E|H_1;B-C] \ll P_a[E|H_2;B-C] \)). Thus, whereas \( \alpha \) considers \( C-E \) (given \( B \)) to be powerful evidence for \( H_1 \) over \( H_2 \), \( \beta \) takes the very same evidence to forcefully support \( H_2 \) over \( H_1 \). If this kind of disagreement occurs often or for important hypotheses in a scientific
discipline, the empirical objectivity of that discipline would be a shambles. Each scientist understands the empirical import of these hypotheses so differently that each \( H_j \) as understood by \( \alpha \) is an empirically different hypothesis than \( H_j \) as understood by \( \beta \). Thus, the empirical objectivity of the sciences requires that experts understand significant hypotheses in similar enough ways that the values of their likelihoods are closely aligned.

For now, let’s suppose that each hypothesis \( H_j \) in the list of alternatives has precise, objective or intersubjectively agreed values for its likelihoods (relative to appropriate background and auxiliaries). We’ll mark this agreement by dropping the subscript ‘\( \alpha \)’, ‘\( \beta \)’ and so forth from expressions that represent likelihoods, because all confirmation functions under consideration agree on them. Nevertheless, there are perfectly legitimate scientific contexts where precise agreement on the values of likelihoods isn’t realistic; so later, in Section 5, we’ll see how the present supposition of precise agreement may be relaxed. But for now, the main ideas will be more easily explained if we focus on cases in which all confirmation functions precisely agree on the values of likelihoods.

Scientific hypotheses are usually tested by a stream of evidence: \( C_1,E_1, C_2,E_2, \ldots, C_n,E_n \). Let’s use the expression ‘\( C^n \)’ to represent the conjunction (\( C_1,C_2,\ldots,C_n \)) of descriptions of the first \( n \) observation conditions, and use ‘\( E^n \)’ to represent the conjunction (\( E_1,E_2,\ldots,E_n \)) of descriptions of their outcomes. So a likelihood for a stream of \( n \) observations and their outcomes will take the form \( P[E^n|H_i,B,C^n] = r' \). Furthermore, the evidence should be representable as probabilistically independent components relative to a given hypothesis \( H_i,B \):

\[
P[E^n|H_i,B,C^n] = P[E_1|H_i,B,C_1] \times P[E_2|H_i,B,C_2] \times \ldots \times P[E_n|H_i,B,C_n].
\]

Prior and Posterior Probabilities. The degree to which a hypothesis is confirmed on the evidence, \( P_\alpha[H_i|B,C^n,E^n] \), is called the posterior probability of the hypothesis – it represents probabilistic degree of confirmation posterior to taking account of the evidence. Bayes’ Theorem will show that posterior probabilities depend on two kinds of factors: likelihoods, \( P[E^n|H_i,B,C^n] \), and prior probabilities \( P_\alpha[H_i|B] \). Prior probabilities represent the degree to which a hypothesis \( H_i \) is supported by non-evidential plausibility considerations, prior to taking the evidence into account. The notion of priority for prior probabilities isn’t temporal – it might make better sense to call them non-evidential probabilities. Though non-evidential, the plausibility considerations that inform values for priors may not be purely a priori. They may include both conceptual and broadly empirical considerations not captured by the likelihoods.

Because plausibility assessments are usually less objective than likelihoods, critics sometimes brand priors as merely subjective, and take their role in the evaluation of hypotheses to be highly problematic. But plausibility assessments often play a crucial role in the sciences, especially when evidence is
insufficient to distinguish among some alternative hypotheses. Furthermore, the epithet ‘merely subjective’ is unwarranted. Plausibility assessments are often backed by extensive arguments that draw on forceful conceptual and empirical considerations not captured by likelihoods. That’s the epistemic role of the thought experiment, for example.

Indeed, we often have good reasons, besides the evidence, to strongly reject some logically possible alternatives as just too implausible, or as, at least, much less plausible than better conceived candidates. In evaluating hypotheses, we often bring such considerations to bear, at least implicitly. For, given any hypothesis, logicians can always cook up numerous alternatives that agree with it on all the evidence available thus far. Any reasonable scientist will reject most such inventions immediately, because they look ad hoc, contrived or plain foolish. Such reasons for rejection appeal to neither purely logical characteristics of these hypotheses, nor to evidence. All such reasons are ‘mere’ plausibility assessments, not part of the evidential likelihoods.

Prior plausibilities are ‘subjective’ in the sense that scientists may disagree on the relative merits of plausibility arguments, and so disagree on the values for priors. Furthermore, the plausibility of a hypothesis is usually somewhat vague or imprecise. So it’s reasonable to represent priors by an interval of values, a plausibility range, rather than by specific numbers. We’ll see more about how that works a bit later. The main point is that plausibility assessments in the sciences are far from mere subjective whims. They play an important role in the epistemology of the sciences. So it’s a virtue of Bayesian confirmation theory that it provides a place for such assessments to figure into the logic of hypothesis evaluation.

Forms of Bayes’ Theorem. Let’s now examine several forms of Bayes’ Theorem, each derivable from our axioms. Here is the simplest:

\[
P_a[H_i|B\cdot C^n\cdot E^n] = \frac{P[E^n|H_i\cdot B\cdot C^n] \cdot P_a[H_i|B] \cdot P_a[C^n|H_i\cdot B]}{P_a[E^n|B\cdot C^n] \cdot P_a[C^n|B]}
\]

Here the posterior probability of a hypothesis is seen to depend on the likelihood it assigns the evidence, its prior probability, and the simple probability of the evidence, \(P_a[E^n|B\cdot C^n]\). If an outcome \(E_k\) occurs with likelihood \(P[E_k|H_i\cdot B\cdot C_k]\) = 0, then the cumulative likelihood \(P[E^n|H_i\cdot B\cdot C^n]\) = 0 as well. As a result the
Bayesian Confirmation Theory

posterior degree of confirmation of \( H_i \) crashes to 0; \( H_i \) is falsified by the evidence.

This version of Bayes’ Theorem includes the terms \( P_a[H_i|C^*B] \) and \( P_a[C^*B] \), which express how likely it is that the conditions for the experiments or observations actually hold. These factors are often suppressed in presentations of Bayes’ Theorem, perhaps by hiding conditions \( C^* \) in the background \( B \). However, that approach makes \( B \) continually change as new evidence is accumulated. So it’s preferable to make these factors explicit, and deal with them directly. That’s easy, because in realistic cases the ratio \( \frac{P_a[H_i|C^*B]}{P_a[C^*B]} \) should be 1, or nearly 1, because the truth of the hypothesis should not be relevant to whether the observation conditions hold.

The simple probability of the evidence represents a weighted average of likelihoods across all the alternative hypotheses:

\[
P_a[H_j|B,C_n] = \sum_j P_a[H_j\mid B,C_n] \frac{P_a[H_j]}{P_a[H_i]} \quad \text{for} \quad P_a[C^*B] = P_a[C^*].
\]

This factor is hard to assess if one isn’t aware of all hypotheses worthy of consideration. So in most cases, another form of Bayes’ Theorem is more useful – a form that compares one pair of hypotheses at a time.

(2) Ratio Form of Bayes’ Theorem:

\[
\frac{P_a[H_j\mid B,C_n,E^n]}{P_a[H_i\mid B,C_n,E^n]} = \frac{P[E^n\mid H_j,B,C_n]}{P[E^n\mid H_i,B,C_n]} \times \frac{P_a[H_j\mid B]}{P_a[H_i\mid B]} \times \frac{P_a[C^n\mid H_j]}{P_a[C^n\mid H_i]}
\]

\[
= \frac{P[E^n\mid H_j,B,C_n]}{P[E^n\mid H_i,B,C_n]} \times \frac{P_a[H_j\mid B]}{P_a[H_i\mid B]} \quad \text{for} \quad P_a[C^*B] = P_a[C^*].
\]

The second line follows when neither hypothesis makes the occurrence of the observation conditions more likely than the other: \( P_a[C^*B] = P_a[C^*H_i] \). This should hold for most real applications, so let’s suppose it holds throughout the remaining discussion.\(^7\)

This form of Bayes’ Theorem is the most useful for many scientific applications, where only a few alternative hypotheses are considered at a time. It shows that likelihood ratios carry the full import of the evidence. Evidence influences the evaluation of hypotheses in no other way. Although this version has not received much attention in the philosophical literature, it’s so central to a realistic Bayesian Confirmation Theory that I’ll discuss it in detail.

Notice that the ratio form of the theorem easily accommodates situations in which we don’t have precise values for prior probabilities. For one thing, it only depends on our ability to assess how much more or less plausible \( H_j \) is than \( H_i \) – the ratio \( P_a[H_j\mid B]/P_a[H_i\mid B] \). Such relative plasibilities are much easier to judge than are specific numerical values for individual hypotheses. This results
in assessments of ratios of posterior confirmational probabilities – for example, $P_a[H_i|B\cdot C\cdot E]/P_a[H_i|B\cdot C\cdot E] = 1/10$ says ‘on the evidence, $H_i$ is ten times more plausible than $H_i'$. Although such posterior ratios don’t supply values for the individual posterior probabilities, they place an important constraint on the posterior confirmation of $H_i$ since logically $P_a[H_i|B\cdot C\cdot E] \leq P_a[H_i|B\cdot C\cdot E]/P_a[H_i|B\cdot C\cdot E]$.

Furthermore, this form of Bayes’ Theorem tolerates a good deal of vagueness or imprecision in assessments of the ratios of prior plausibilities. In practice, one need only assess bounds for the prior plausibility ratios to achieve meaningful results. Given a prior ratio in a specific interval, $q \leq P_a[H_j|B]/P_a[H_i|B] \leq r$, a likelihood ratio $P[E^n|H_i\cdot B\cdot C^n]/P[E^n|H_j\cdot B\cdot C^n] = s_n$ produces a posterior confirmation ratio in the interval $s_n \cdot q \leq P_a[H_j|B\cdot C^n\cdot E^n]/P_a[H_i|B\cdot C^n\cdot E^n] \leq s_n \cdot r$. As the likelihood ratio value $s_n$ approaches 0, the interval value for the posterior ratio gets smaller, and its upper bound $s_n \cdot r$ approaches 0; so the absolute degree of confirmation of $H_j$ $P_a[H_j|B\cdot C^n\cdot E^n]$, also must approach 0. This is really useful, because it can be shown that when $H_j\cdot B\cdot C^n$ is true and $H_i$ is empirically distinct from $H_j$, the values of likelihood ratios $P[E^n|H_i\cdot B\cdot C^n]/P[E^n|H_j\cdot B\cdot C^n]$ will very likely approach 0 as the amount of evidence increases. (I’ll discuss this Likelihood Ratio Convergence result below.) When that happens, the upper bound on the posterior probability ratio also approaches 0, driving the posterior probability of $H_i$ to approach 0, effectively refuting $H_i$. Thus, false competitors of a true hypothesis are eliminated.

Relative to each hypothesis, evidential events should be probabilistically independent of one another (or at least parsible into independent clusters). So the likelihood ratio for the total evidence decomposes into a product of likelihood ratios for each observation:

$$\frac{P[E^n|H_i\cdot B\cdot C^n]}{P[E^n|H_j\cdot B\cdot C^n]} = \frac{P[E_i|H_i\cdot B\cdot C^n_i]}{P[E_i|H_j\cdot B\cdot C^n_i]} \times \frac{P[E_{n\cdot i}|H_i\cdot B\cdot C^n_{n\cdot i}]}{P[E_{n\cdot i}|H_j\cdot B\cdot C^n_{n\cdot i}]} \times \frac{P[E_n|H_i\cdot B\cdot C^n]}{P[E_n|H_j\cdot B\cdot C^n]}$$

It follows from (2) that a previous confirmation ratio (based on the previous evidence) is updated on new evidence via multiplication by the likelihood ratio for the new evidence:

(3) Ratio Bayesian Updating Formula:

$$\frac{P_a[H_j|B\cdot C^n\cdot E^n]}{P_a[H_i|B\cdot C^n\cdot E^n]} = \frac{P[E_n|H_j\cdot B\cdot C^n]}{P[E_n|H_i\cdot B\cdot C^n]} \times \frac{P_a[H_j|B\cdot C^n\cdot E^n]}{P_a[H_i|B\cdot C^n\cdot E^n]}$$

$$= \frac{P[E_i|H_j\cdot B\cdot C^n_i]}{P[E_i|H_i\cdot B\cdot C^n_i]} \times \frac{P[E_{n\cdot i}|H_j\cdot B\cdot C^n_{n\cdot i}]}{P[E_{n\cdot i}|H_i\cdot B\cdot C^n_{n\cdot i}]} \times \frac{P_a[H_j|B]}{P_a[H_i|B]}$$
The second line of (3) shows how the contribution of any individual piece of evidence may be reassessed (even tossed out) if it comes into doubt. Similarly, prior probability ratios (or intervals for them) may be reassessed and changed to reflect additional plausibility considerations.8

From a Bayesian perspective, when scientists report on their experimental findings in research journals, they should indicate the impact of the evidence on various hypotheses by reporting the values of the likelihood ratios, \( P[E|H_i; B; C]/P[E|H_i; B; C] \), for the evidence \( C\cdot E \) obtained from their research.9 Although they may say little or nothing about the (ratios of) prior plausibilities, some conception of the plausibility of the hypotheses must be in play, at least implicitly, because if no one in the relevant scientific community takes hypothesis \( H_j \) at all seriously (i.e. if the relevant scientific community takes \( P_\alpha[H_j|B] \) to be almost 0 to begin with), then no one will care about an experimental study that ‘finds strong new evidence against \( H_j \)' by establishing some result \( (C\cdot E) \) that makes the likelihood ratio \( P[E|H_j; B; C]/P[E|H_i; B; C] \) extremely small. No respectable scientific journal would bother to publish such results. If prior plausibility played no role, such results would deserve as much consideration as any.

Bayesian confirmation is a version of eliminative induction. Suppose \( H_i \) is a true hypothesis, and consider what happens to each of its false competitors, \( H_j \), if enough evidence becomes available to drive each of the likelihood ratios \( P[E^n|H_i; B; C]/P[E^n|H_j; B; C] \) toward 0. Equation (2) shows that each alternative \( H_j \) is effectively refuted because \( P_\alpha[H_j|B; C-E] \leq P_\alpha[H_i|B; C-E]/P_\alpha[H_i|B; C-E] \) approaches 0. As this happens the posterior probability of \( H_i \) must approach 1, as the next two forms of Bayes’ Theorem show.

The odds against \( A \) given \( B \) is, by definition, \( \Omega_\alpha[\neg A|B] = P_\alpha[\neg A|B]/P_\alpha[A|B] \). If we sum the ratio versions of Bayes’ Theorem over all alternatives to hypothesis \( H_i \) (including the catch-all \( H_{i'} \) if needed), we get an Odds Form of Bayes’ Theorem:

(4) Odds Form of Bayes’ Theorem

\[
\Omega_\alpha[\neg H_i|B; C^n; E^n] = \sum_{j\neq i} \frac{P[H_j|B; C^n; E^n]}{P[H_j|B; C^n; E^n]} \cdot \frac{P[H_j|B; C^n; E^n]}{P[H_i|B; C^n; E^n]} = \sum_{j\neq i} \frac{P[E^n|H_i; B; C^n]}{P[E^n|H_i; B; C^n]} \cdot \frac{P_\alpha[H_i|B]}{P_\alpha[H_j|B]} \cdot \frac{P_\alpha[H_j|B]}{P_\alpha[H_i|B]}
\]

If the catch-all alternative isn’t needed, just drop the expression after the ‘+’ sign. We represent the term for the catch-all hypothesis separately because...
the likelihood of evidence relative to it will not generally enjoy the kind of objectivity possessed by likelihoods for specific hypotheses. We indicate this by leaving the subscript ‘α’ on catch-all likelihoods.

Although the catch-all hypothesis lacks objective likelihoods, the influence of the whole catch-all term should diminish as additional specific hypotheses become articulated. When a new hypothesis \( H_{m+1} \) is made explicit, the old catch-all \( H_\alpha \) is replaced by a new one, \( H_\alpha' \), of form \( \neg H_1 \ldots \neg H_m \neg H_{m+1} \). The prior probability for the new catch-all hypothesis is peeled off the prior of the old catch-all: \( P_\alpha [H_\alpha' \mid B] = P_\alpha [H_\alpha \mid B] - P_\alpha [H_{m+1} \mid B] \). So the influence of the catch-all term should diminish towards 0 as new alternative hypotheses are developed. The relationship between the odds against \( H_i \) and its posterior probability is this:

(5) Bayes’ Theorem: From Posterior Odds to Posterior Probability

\[
P_\alpha[H_i \mid B \cdot C^\alpha \cdot E^\alpha] = \frac{1}{1 + \Omega_\alpha[-H_i \mid B \cdot C^\alpha \cdot E^\alpha]}.
\]

For scientific contexts in which not all significant alternative hypotheses can be surveyed, the formulas for posterior odds and posterior probabilities provided by equations (4) and (5) are only of conceptual interest. They tell us about the nature of the logic, but may not permit us to compute actual posterior probabilities of hypotheses that remain unrefuted by likelihood ratios. In practice, the best we can usually do in such contexts is compare pairs of hypotheses, and find evidence enough to drive one of each pair to near extinction via extreme likelihood ratios. Thus, Bayesian confirmation is fundamentally a variety of eliminative induction, where the hypothesis that remains unrefuted is our best candidate for the truth.

If we are fortunate enough to develop the true alternative, then each of its evidentially distinct rivals may be laid low by evidence via the likelihood ratios. As that happens, the true hypothesis will climb to the top of the list of alternatives and remain there: its posterior plausibility \( P_\alpha[H_i \mid B \cdot C^\alpha \cdot E^\alpha] \) will become many times larger than the posterior plausibilities of alternatives. In principle, its posterior probability heads towards 1, but in practice we merely recognize such a hypothesis as very strongly confirmed – superior to all alternatives considered thus far. Thus, this Bayesian logic is a formal representation of how the evaluation of hypotheses in the theoretical sciences actually operates.
4. The Likelihood Ratio Convergence Theorem

When $H_i \cdot B$ is true, the series of likelihood ratios $P[E^n|H_j \cdot B \cdot C^n]/P[E^n|H_i \cdot B \cdot C^n]$ will very probably favour $H_i$ over empirically distinct alternatives $H_j$ by heading towards 0 as the evidence accumulates (as $n$ increases). A Bayesian Convergence Theorem establishes this fact. Before stating the theorem, I’ll first explain some notation.

For observation sequence $C^n$, consider each of the possible outcomes sequences $E^n$. Some would result in likelihood ratios for $H_j$ over $H_i$ that are less than $\varepsilon$, for some chosen small increment $\varepsilon > 0$ (e.g. you might choose $\varepsilon = 1/1000$). For specific $\varepsilon$, the set of all possible such outcome sequences is expressed by \(\{E^n: P[E^n|H_j \cdot B \cdot C^n]/P[E^n|H_i \cdot B \cdot C^n] < \varepsilon\}\). This will be some particular finite set of sentences. Now, consider the disjunction of all sentences in that set; the resulting disjunctive sentence asserts that one of the outcome sequences, described by one of the sentences in the set, is true. We indicate this disjunctive sentence by placing the ‘or’ symbol ‘\(\lor\)’ in front of the expression for the set: \(\lor(E^n: P[E^n|H_j \cdot B \cdot C^n]/P[E^n|H_i \cdot B \cdot C^n] < \varepsilon)\). How likely is it, if $H_j \cdot B \cdot C^n$ is true, that this disjunctive sentence will be true? That is, how likely is it, if $H_j \cdot B \cdot C^n$ is true, that ‘one of the outcome sequences $E^n$ will occur that makes $P[E^n|H_j \cdot B \cdot C^n]/P[E^n|H_i \cdot B \cdot C^n] < \varepsilon$’? The Likelihood Ratio Convergence Theorem answers this question by providing a lower bound on how likely this is, and that lower bound approaches 1 as $n$ increases. The Theorem expresses this in terms of a likelihood: $P[\lor(E^n: P[E^n|H_j \cdot B \cdot C^n]/P[E^n|H_i \cdot B \cdot C^n] < \varepsilon)|H_i \cdot B \cdot C^n]$.

The full statement of the theorem comes in two parts. The first part encompasses cases where $H_j$ says some outcome is impossible that $H_i$ counts as possible; the second part encompasses evidential sequences where such extreme disagreement doesn’t happen.

Likelihood Ratio Convergence Theorem:

1. Suppose a subsequence $C_i^n$ of the whole evidence stream consists of observations where for each one, $C_i$, there is some possible outcome $E_k$ deemed possible by $H_j \cdot B$ to at least some small degree $\delta > 0$ but deemed impossible by $H_i \cdot B$ – that is, for each $C_i$, there is a possible $E_k$ such that $P[E_k|H_j \cdot B \cdot C_i] \geq \delta > 0$ but $P[E_k|H_i \cdot B \cdot C_i] = 0$. Then,

\[
P[\lor(E^n: P[E^n|H_j \cdot B \cdot C^n] = 0)|H_i \cdot B \cdot C^n] \geq 1 - (1-\delta)^m,
\]

which approaches 1 for large $m$.

2. Suppose the evidence stream $C^n$ consists of observations where for each one, $C_i$, each possible outcome $E_k$ deemed possible by $H_j \cdot B$ is also deemed possible by $H_i \cdot B$ – that is, for each $C_i$, if $P[E_k|H_j \cdot B \cdot C_i] > 0$, then
\[ P[E_k \mid H_i \cdot B \cdot C_k] > 0. \] And further suppose that for each \( E_k \) such that \( P[E_k \mid H_i \cdot B \cdot C_k] > 0 \), \( P[E_k \mid H_i \cdot B \cdot C_k] \geq \gamma P[E_k \mid H_i \cdot B \cdot C_k] \), for some small positive \( \gamma \leq 1/3 \) – that is, there is some positive \( \gamma \leq 1/3 \) such that no such possible \( E_k \) disfavours the competitor \( H_i \) so much as to make \( P[E_k \mid H_i \cdot B \cdot C_k] / P[E_k \mid H_i \cdot B \cdot C_k] < \gamma \). Then for any small positive \( \varepsilon < 1 \) you might choose (but large enough that for the number of observations \( n \) being contemplated, the value of \( (1/n) \sum_{k=1}^{n} \text{EQI}[C_k \mid H_i / H_j \mid B] > -(\log \varepsilon)/n \)),

\[
P[\forall \{E^n: P[E^n \mid H_i \cdot B \cdot C^n] / P[E^n \mid H_i \cdot B \cdot C^n] < \varepsilon \} \mid H_i \cdot B \cdot C^n] > 1 - \frac{(\log \gamma)^2}{n(1/n) \sum_{k=1}^{n} \text{EQI}[C_k \mid H_i / H_j \mid B] + (\log \varepsilon)/n^2}
\]

which approaches 1 for large \( n \), provided \( (1/n) \sum_{k=1}^{n} \text{EQI}[C_k \mid H_i / H_j \mid B] \) has a positive lower bound – that is, provided the sequence of observation \( C^n \) has an average expected quality of information (average EQI) for empirically distinct \( H_j \) given \( H_i \) that doesn’t get arbitrarily near 0 as the evidence sequence increases.12 (The base of the log doesn’t matter, but let’s take it to be 2; then for \( \varepsilon = 1/2^\gamma \), \( \log \varepsilon = -k \); and for \( \gamma = 1/2^\nu \), \( \log \gamma^2 = \nu^2 \).)

The term on the right-hand side of the inequality is a worst case lower bound. The actual value of \( P[\forall \{E^n: P[E^n \mid H_i \cdot B \cdot C^n] / P[E^n \mid H_i \cdot B \cdot C^n] < \varepsilon \} \mid H_i \cdot B \cdot C^n] \) will, in all but the most extreme cases, be much larger than this bound. That is, given two specific hypotheses \( H_i \) and \( H_j \) (and their associated likelihoods for possible outcomes), one can actually compute the precise value of \( P[\forall \{E^n: P[E^n \mid H_i \cdot B \cdot C^n] / P[E^n \mid H_i \cdot B \cdot C^n] < \varepsilon \} \mid H_i \cdot B \cdot C^n] \). For most hypotheses and most types of possible evidence, the value of this likelihood is much larger than the lower bound given by this worst case theorem.

The term \( (1/n) \sum_{k=1}^{n} \text{EQI}[C_k \mid H_i / H_j \mid B] \) is an information theoretic measure of how good, on average, the range of all possible outcomes of \( C_k \) are at distinguishing between \( H_i \) and \( H_j \), if \( H_i \cdot B \) is true. The formula for each \( \text{EQI}[C_k \mid H_i / H_j \mid B] \) is

\[
\text{EQI}[C_k \mid H_i / H_j \mid B] = \sum_{O_{ts} \cdot C_k} \log(P[O_{ts} \mid H_i \cdot B \cdot C_k] / P[O_{ts} \mid H_j \cdot B \cdot C_k]) P[O_{ts} \mid H_i \cdot B \cdot C_k]
\]

where the sum ranges over all the possible outcomes \( O_{ts} \) of observation \( C_k \), that \( H_i \) takes to be possible (i.e, for which \( P[O_{ts} \mid H_i \cdot B \cdot C_k] > 0 \)).13

Thus, the Likelihood Ratio Convergence Theorem establishes that if \( H_i \) (together with \( B \cdot C^n \)) is true, as the sequence of observations \( C^n \) increases, it becomes highly likely (as near 1 as you like) that its outcomes will provide likelihood ratios as close to 0 as you wish.14 This theorem is not subject to the usual criticisms of Bayesian convergence results. The theorem (and its proof) does not rely on prior probabilities in any way. It doesn’t suppose that the
evidence is ‘identically distributed’ – it applies to any pair of empirically distinct hypotheses. It’s a weak law of large numbers result that gives explicit lower bounds on the rate of convergence, so there’s no need to wait for the infinite long run. It’s a convergence to truth result (not merely convergence to agreement). It doesn’t depend on countable additivity.

Furthermore, because this theorem doesn’t depend on prior probabilities, it’s not undermined by their being reassessed and changed as new conceptual and broadly empirical considerations are introduced. Provided that the series of reassessments of prior plausibilities doesn’t push the prior of the true hypothesis ever nearer to zero, the Likelihood Ratio Convergence Theorem implies (via equation (2)) that the evidence will very probably bring the posterior probabilities of its empirically distinct rivals to approach 0 via decreasing likelihood ratios; and as this happens, the posterior probability of the true hypothesis will head towards 1 (via equations (4) and (5)).

5. When Likelihoods Are Not Precise

For some important contexts, it’s unreasonable to expect likelihoods to possess precise, agreed values, but the evidence remains capable of sorting among hypotheses in a reasonably objective way. Here’s how that works.

Consider the following continental drift hypothesis: the land masses of Africa and South America were once joined, then split and have drifted apart over the eons. Let’s compare it to an alternative contraction hypothesis: the continents have fixed positions acquired when the earth first formed, cooled and contracted into its present configuration. On each of these hypotheses, how likely is it that

(1) the shape of the east coast of South America should match the shape of the west coast of Africa as closely as it in fact does?
(2) the geology of the two coasts should match up so well?
(3) the plant and animal species on these distant continents should be as closely related as they are?

One may not be able to determine anything like precise numerical values for such likelihoods. But experts may readily agree that each of these observations is much more likely on the drift hypothesis than on the contraction hypothesis, and they jointly constitute very strong evidence in favour of drift over contraction. On a Bayesian analysis, this is due to the fact that even though these likelihoods do not have precise values, it’s obvious to experts that the ratio of the likelihoods is pretty extreme, strongly favouring drift over contraction.
(according to the Ratio Form of Bayes’ Theorem), unless contraction is taken to be much more plausible than the drift on other grounds.\textsuperscript{17}

I argued earlier that disagreement on likelihoods among members of a scientific community would be disastrous to the scientific enterprise were it to result in disparate assessments of which hypotheses are favoured by evidence. However, precise values for likelihoods are not crucial to the way evidence sorts among hypotheses. Rather, ratios of likelihoods do all the heavy lifting. So, when two confirmation functions $P_\alpha$ and $P_\beta$ disagree on the values of likelihoods, they’ll agree well enough on the refutation and support for hypotheses if they yield directionally agreeing likelihood ratios.

**Directional Agreement Condition** for Likelihoods Ratios: The likelihood ratios for a pair of confirmation functions $P_\alpha$ and $P_\beta$ directionally agree on the possible outcomes of observations relevant to a pair of hypotheses just in case for each possible outcome $E_k$ of the conditions $C_k$ in the evidence stream,

$$P_\alpha[E_k|H_j;B;C_k]/P_\alpha[E_k|H_i;B;C_k] < 1 \text{ just when } P_\beta[E_k|H_j;B;C_k]/P_\beta[E_k|H_i;B;C_k] < 1,$$

and

$$P_\alpha[E_k|H_j;B;C_k]/P_\alpha[E_k|H_i;B;C_k] > 1 \text{ just when } P_\beta[E_k|H_j;B;C_k]/P_\beta[E_k|H_i;B;C_k] > 1,$$

each of these likelihood ratios is either close to 1 for both functions or for neither.

When this condition holds, the evidence supports $H_i$ over $H_j$ according to $P_\alpha$ just when it does so for $P_\beta$. Furthermore, although the rate at which the likelihood ratios increase or decrease as evidence accumulates may differ for these confirmation functions, the total impact of the cumulative evidence will affect the refutation and support of hypotheses in the same way. Indeed, the **Likelihood Ratio Convergence Theorem** still applies. The proof of the theorem doesn’t depend on likelihoods being objective. It applies to each confirmation function $P_\alpha$ individually. Thus, when a family of confirmation functions satisfy the Directional Agreement Condition and enough empirically distinguishing observations are forthcoming, each will very probably yield likelihood ratios for empirically distinct false competitors of a true hypothesis that become extremely small. Directional Agreement guarantees that if such convergence towards 0 happens for one of the agreeing confirmation functions, it must happen for them all. As that happens, the posterior confirmation value of the true hypothesis must rise towards 1 according to each confirmation function in the family.

### 6. What Is Confirmational Probability?

Now that we understand how confirmational probabilities may come to indicate the falsehood or truth of hypotheses, perhaps it’s not so important to...
try to interpret them. But, let’s briefly consider some prominent views about how to understand what these functions are.

6.1 The Syntactic Logical Interpretation
Early on, some Bayesian logicists attempted to explicate confirmation functions that depend only on the syntactic structures of sentence, in the same way that deductive logical entailment depends only on syntactic structures. Most logicians now take this project to be fatally flawed. On this view, hypotheses with the same syntactic structure should have the same prior probability values. But given any hypothesis, logicians can easily construct an infinite number of alternatives with the same syntactic structure. Most such alternatives would be quickly rejected by scientists as ad hoc and ridiculously implausible, but such assessments are based on semantic content, not logical structure. So semantic content should matter. Moreover, how are we supposed to implement this syntactic approach in real cases? Are we to compare the syntactic structures of the various interpretations of quantum theory to see which has the higher prior probability? The defenders of the syntactic-structural view owe us credible reasons to base non-evidential plausibility assessments on syntactic structure alone.

6.2 Subjective Degrees of Belief
Think of $\alpha$, $\beta$, $\gamma$ and so forth as logically ideal agents, each having his or her own degrees of belief in various propositions. We may represent each agent’s belief-strengths in terms of a belief-strength function, $P_\alpha$, $P_\beta$, $P_\gamma$, and so forth, defined on statements. Taking unconditional probability as basic, read $P_\alpha[A] = r$ as saying ‘the strength of $\alpha$’s belief that $A$ is $r$’; and read $P_\alpha[A|B] = r$ as saying ‘the strength of $\alpha$’s belief that $A$, were she to become newly certain of $B$ (and nothing more than $B$) would be $r$.’ This is the widely subscribed Subjectivist or Personalist interpretation of confirmational probabilities. Subjectivists intend this to be the same notion of probability employed in Bayesian decision theory.

Versions of the problem of old evidence present major difficulties for this view. They show that belief-function likelihoods cannot maintain the kind of objective values that confirmation-function likelihoods should have. This problem is much worse than usually realized.

Suppose ‘$E$’ says ‘the coin lands heads on the next toss’, $H$ says ‘the coin is fair’ and $C$ says ‘the coin is tossed in the usual way on the next toss’, so that the confirmational likelihood should be $P[E|H\cap C] = 1/2$. However, if the agent is already certain that $E$, then her belief function likelihood should be $P[E|H_j\cap C] = 1$ for every hypothesis $H_j$, which undermines the role of the
likelihood in testing any hypothesis. Furthermore, even when the agent isn’t certain of $E$, but becomes certain of trivial disjunctions involving $E$ – for example, ‘either the outcome of the next toss will be heads, or Jim won’t like the outcome of the next toss’, ($E \vee F$) – it can be shown that belief-function likelihoods become radically altered from their objective values.

The problem is that an agent’s belief-function likelihood has to represent her belief strength in the evidence statement when the hypothesis is added to everything else the agent already holds. But other beliefs and partial beliefs (even hunches) the agent has will almost always severely interfere with the objective values the likelihoods should have for confirmational purposes. Thus, a Bayesian account of confirmation and belief will require confirmation functions that are distinct from belief functions, and some account of how degrees-of-confirmation are supposed to inform an agent’s degrees-of-belief.

Furthermore, the subjectivists’ ideal agents are taken to be logically omniscient. They assign belief-strength 1 to all logical truths. How is this unobtainable norm for confirmation/belief functions supposed to be relevant to the functioning of real people? However, if we don’t reduce confirmation functions to ideal agents’ belief functions, this ‘logical omniscience problem’ has no hold over confirmation functions. Real people use the logic of confirmation functions in the same sort of way they might use any logic to inform their real (non-ideal) belief strengths.

6.3 Another Logical View

Rather than ask what confirmation functions are, it’s more fruitful to ask what they do. Under appropriate circumstances they’re truth-indicating indices. If, among the alternative hypotheses proposed to account for a given subject-matter, we are fortunate enough to think of a hypothesis that happens to be true, and if we find enough ways to empirically test it against rivals, then all that’s needed for confirmational success is persistent testing and not too much bad luck with how the evidence actually turns out. For, according to the Likelihood Ratio Convergence Theorem, the true hypothesis itself says, via its likelihoods, that a long enough (but finite) stream of observations is very likely to produce outcomes that will drive the likelihood ratios of empirically distinct false competitors to approach 0. As this happens, the confirmation index of these competitors, as measured by their posterior probabilities, also approaches 0, and the confirmation index of the true hypothesis (or its disjunction with empirically equivalent rivals) will approach 1.

This result does not imply that whatever hypothesis has index near 1 at a given moment is likely to be the true alternative. Rather, it suggests the pragmatic strategy of continually testing hypotheses, and taking whichever of them has an index nearest to 1 (if there is one) as the best current candidate for being true.
The convergence theorem implies that maintaining this strategy and continually testing is very likely to eventually promote the true hypothesis (or its disjunction with empirically indistinguishable rivals) to the status of best current candidate, and it will remain there. So if we align our belief strengths for hypotheses with their approximate confirmation indices, eventually we should (very probably) come to strongly believe the true hypothesis. But this eliminative strategy only promises to work if we continue to look for rivals and continue to test the best alternative candidates against them. This strategy shouldn’t seem novel or surprising. It’s merely a rigorously justified version of scientific common sense.

When the empirical evidence is meagre or unable to distinguish between a pair of hypotheses, the confirmation index must rely on whatever our most probative non-evidential considerations tell us. We often have good reasons, besides the observable evidence, to strongly discount some logically possible alternatives as just too implausible, or at least as significantly less plausible than some better conceived rivals. We always bring some such considerations to bear, at least implicitly. It is a virtue of Bayesian confirmation theory that it provides a place for such assessments to figure into the logic of hypothesis evaluation.

Notes

1. $P_\alpha[A \mid C] = 1$ for all $C$ just when $P_\alpha[A \mid \neg A] = 1$. So $P_\alpha[A \mid C] = 1$ for all $C$ implies $A$ is either logical or analytically true, or a consequence of set theory or some other piece of pure mathematics employed by the sciences. Axiom 6 yields the converse implication.

2. $B$ may itself contain hypotheses that are subject to confirmation via the same kind of treatment described for hypotheses $H_i$ and $H_j$ below (though their confirmation may be relative to some simpler auxiliaries $B^*$; perhaps even tautological $B^*$).

3. See David Lewis’s (1980) argument for the objectivity of likelihoods based on chance statements. His Principal Principle is a direct inference principle governing such likelihoods. Lewis maintains that objective chance is a purely theoretical concept, and that the Principal Principle captures ‘all we know about chance’.

4. The only exception is the catch-all hypothesis $H_K$, which seldom yields objective likelihoods.

5. If the evidence were not parsible into independent parts in this way, then hypothesis $H_i \cdot B$ would always have to consult a large number of past evidential results, $(C^\omega \cdot E^\omega)$, in order to say how likely the various outcomes $E$ of the next experiment $C$ are – since $P[E \mid H_i \cdot B \cdot C \cdot (C^\omega \cdot E^\omega)]$ would differ from $P[E \mid H_i \cdot B \cdot C]$. In other words, vague prior plausibilities may be represented by a set of confirmation functions that jointly cover the plausibility ranges for hypotheses.

6. This supposition also avoids inessential complexity. Nothing I’ll say below changes much when ratios $P_\alpha[C^\omega \mid H_i \cdot B]/P_\alpha[C^\omega \mid H_j \cdot B]$ don’t get exceptionally far from 1. If they did, the experimental conditions themselves would count as significant evidence.

7. Technically, changing the value of the (interval covering the) prior plausibility ratio means switching to a different confirmation function (or different set of functions with priors that span the new interval).
Here Bayesian Confirmation Theory agrees with the view about how statistical hypotheses should be tested called Likelihoodism. See Edwards 1972; Royall 1997; and Forster, this volume.

For a proof see Hawthorne 2008, supplements 4–7.

This provision can fail only if new observations $C_n$ can only produce ever weaker evidence (whose likelihood ratio values have to be ever closer to 1 for all possible outcomes of $C_n$) as more evidence is obtained (as $n$ is increased).

EQI$[C_n | H_i / H_j | B]$ is the expected value of the logs of the likelihood ratios. Each EQI$[C_n | H_i / H_j | B]$ is greater than 0 if some $O_{ki}$ has $P[O_{ki} | H_i ; B - C_j] > P[O_{kj} | H_j ; B - C_j]$; otherwise EQI$[C_n | H_i / H_j | B] = 0$.

A short evidence sequence may suffice if the average expected quality of information is large.

This claim depends on $H_i$ being empirically distinct from each alternative – that is, that the $C_n$ have possible outcomes $E_i$ such that $P[E_i | H_i ; B - C_j] > P[E_i | H_j ; B - C_j]$. If the true hypothesis has empirically equivalent rivals, then convergence implies posterior probability of their disjunction goes to 1. Among the equivalent rivals, $P_p[| H_i | B - C_j - E_i ] = P_p[| H_j | B ]$. So the true hypothesis can obtain a posterior probability near 1 (after evidence drives the posteriors of empirically distinct rivals near 0) just in case plausibility considerations result in its prior plausibility being much higher than the sum of those of its empirically equivalent rivals.

Technically, an imprecise likelihood is represented by a set of confirmation functions with likelihood values that span the interval of the imprecision.

Historically, geologists largely dismissed the evidence described above until the 1960s. The strength of this evidence didn’t suffice to overcome non-evidential (though broadly empirical) considerations that made the drift hypothesis seem much less plausible than the traditional contraction view. Chiefly, there appeared to be no plausible mechanisms that could move the continents through the ocean floor. Such objections were overcome when a plausible enough ‘convection mechanism’ was articulated and evidence favouring it over the ‘embedded in bedrock’ model was acquired.

Carnap (1950) is the most widely known proponent of this idea.

Ramsey (1926), de Finetti (1937) and Savage (1954) are well-known proponents of this view. Howson and Urbach (1993) provide a comprehensive treatment.

Glymour (1980) first raised this problem. For details of the versions mentioned here see Hawthorne (2005).

References


216
Bayesian Confirmation Theory


This page intentionally left blank
B. Philosophy of Particular Sciences
This page intentionally left blank
Philosophy of Physics

Nick Huggett

1. Introduction

Explaining the philosophical dimensions of physics naturally requires first knowing some physics. But given the intended audience of this Companion, I don’t want to assume that the reader has the necessary knowledge. I have attempted to solve the difficulty by offering somewhat unusual presentations of three ‘pillars’ of contemporary physics, which emphasize a philosophical understanding of the material; each is followed by a brief survey of some more philosophical questions. The choice of philosophical topics is selective, aimed to help illuminate the basic physics and to cover recent issues, and is based on my interests. The guiding thought is that it is more useful for the reader to better understand a few topics, than to have a hazy understanding of more. I have not assumed an extensive mathematical background, but I have assumed a willingness to learn mathematical ideas and reason in mathematical terms – after all, how can one be a philosopher of science without learning the language of science?

2. The First Pillar: General Relativity and Spacetime Theories

2.1 The Theory

2.1.1 Metric Geometry

General relativity (GR) involves generalizing familiar – Euclidean – spatial geometry, whose defining feature is the Pythagorean theorem: if a, b and c (the hypotenuse) are the sides of a right-angled triangle then \( a^2 + b^2 = c^2 \). This theorem manifestly presupposes that angles and lengths are well defined; the function from intersecting lines to angles and from lines to their lengths is called the ‘metric’, and generalizing it takes us to GR.

We also need to consider ‘coordinates’: continuous maps from regions of d-dimensional space (or spacetime) to d-tuples, \( x = \langle x_1, x_2, \ldots, x_d \rangle \), of real numbers. At this level of description, coordinates are nothing but arbitrary labels for points of space; but some coordinates are more suited to a given geometry than others. In the familiar Cartesian coordinates (in which axes are straight
lines at right angles, and the coordinate along an axis equals the distance from
the origin) the distance between two points is given by \( \Delta x = \sqrt{(\Delta x_1^2 + \ldots + \Delta x_d^2)} \)
(in two dimensions this follows immediately from the Pythagorean theorem).
Since this quantity equals the length of the shortest curve joining the points,
the formula represents the metric in the given coordinates.

There are many Cartesian coordinates, related by translations of the origin
and rotations – the ‘symmetries’ relating them. Two points are important to
stress: first, all such coordinates agree on the value of \( \sqrt{(\Delta x_1^2 + \ldots + \Delta x_d^2)} \)
between any two points; second, the symmetries of the coordinates reflect the
symmetries of the space – no geometric properties are changed if a figure is
translated or rotated.

Moving from pure geometry to physics requires making contact with the
properties of material systems. As a warm-up, consider a fairy tale (idealizing
the situation prior to GR): imagine a world in which the objects of physical
enquiry are pieces of string and various three-dimensional bodies. Suppose
that the more immediately apparent properties of these objects include judge-
ments of how closely pieces of string follow the same path, and how many
times longer one body is than another in its vicinity. Then we might imagine
beings in the world discovering prescientifically: when any two pieces of
string are pulled tight between two points, they tend towards the same path;
some bodies are ‘stiff’ – when ‘forces’ are applied to either of a pair of stiff
bodies, the ratio of their lengths is little or unchanged.

Then imagine such facts being developed into a well-confirmed, formal
scientific law, L: there are coordinates in which, between any two points, any
taut piece of string will be \( \sqrt{(\Delta x_1^2 + \ldots + \Delta x_d^2)} \) times a standard length of any
stiff body. This theory represents the physical properties of the bodies in ques-
tion: it gives the coordinates of paths taut strings will lie on, and it implies the
constancy of relative length of stiff bodies.

Now, because L features the Pythagorean formula, we can alternatively
express it by saying that space is Euclidean, taut strings lie along straight lines
and stiff bodies are of constant length: we can express it as a theory of geom-
etric objects. (The coordinates referred to in L are then, of course, Cartesian.) We
should be clear, however, that such notions as ‘straight line’ and ‘length’ have
no empirical significance, except through their connections to strings and
bodies and the physics governing them; indeed, they get physical significance
because physics picks out a natural way of naming points. So how should the
beings take their discovery that space is Euclidean? Read naively, the second
formulation assigns space an intrinsic, matter-independent structure, but in
the first formulation it just plays a role in representing the physics of bodies.
More on this question later.

It’s not hard to imagine that the Euclidean metric is not the only possible
length-function: the same curves will generally be assigned different lengths
by Euclidean and non-Euclidean metrics. It’s not even hard to picture in two dimensions: for instance, imagine taking a region of Euclidean space and deforming it so it sat perfectly even with the surface of a sphere; however you did it, curves would have to be stretched or squashed, changing their lengths. It's harder to visualize in more dimensions, but still non-Euclidean geometries are ‘curved’ with respect to ‘flat’ Euclidean geometry. Thus the notion of a straight line is also generalized to (broadly) the shortest curve between two points, termed a ‘geodesic’ to avoid confusion. Because they are curved arbitrarily, in non-Euclidean geometries there are typically no coordinates in which the metric takes a constant form like the Pythagorean function, so typically there are no special coordinates. (Note that Poincaré 1905, chapter 5, argued that a theory empirically equivalent to L can be formulated in any geometry.)

2.1.2 Spacetime Geometry

The first generalization behind GR is from flat to curved, the second is from space to spacetime. Points of spacetime are possible spatiotemporal locations of happenings, often called ‘events’. In coordinate terms, \( x_0 \) represents the time, and the other coordinates, \( x \), the position of an event. In terms of physical systems whose properties are formalized in theories with geometric content, systems that give geometry physical significance, the generalization means moving from static properties to processes: from length comparisons to the propagation of waves, say. Now we should consider real fundamental physics, not fairy tales: especially the quantum theory of matter and force in ‘special relativity’ (SR). This theory too, however, can be written in the form of equations involving coordinates, though now of space and time.

So what are the coordinates in question? Approximately, they are those attached to laboratories; but not exactly, since the earth is accelerating, changing velocity as it orbits the sun. So how does one pick coordinates in which the laws hold? The problem with attaching them to this or that system is that we end up using equations that are false, not because they do not properly describe physical processes, but because we use them in the wrong coordinates. But there’s a simple solution; the coordinates in question are just those in which the equations correctly describe the physics – so-called ‘inertial coordinates’!

That may sound circular, but it isn’t; the equations describe things correctly if and only if there are coordinates in which they hold. We are effectively formulating the theory to say that there are coordinates in which the equations hold, exactly as we did in the fairy tale earlier. Things are only ‘circular’ in the sense that one must discover simultaneously the laws of motion and the coordinates in which they hold, but that is not impossible, just a description of the epistemic problem that physicists face and solve; for example, Newton’s
Principia demonstrates that his laws hold in some frame, identified within experimental error as that of the fixed stars. The complete set of laws is complex, so consider a fragment of the most accurate quantum theories of matter, which describes some general consequences of the theory – approximately, because they are couched in terms of ‘classical’ bodies and the like.

First, in inertial coordinates: the ‘law of inertia’, ‘inertial trajectories’ (those of force-free bodies) satisfy \( x(x_0) = v x_0 + b \) (where \( v \) is a constant velocity and \( b \) the initial position); the ‘light principle’, light travels paths satisfying \( x(x_0) = ce x_0 + b \) (where \( c \) is the speed of light, \( e \) any unit spatial vector, and \( b \) the point of emission). Inertial and light trajectories are the analogues of the fairy tale’s strings. Second, some systems are periodic in time, with constant relative periods as they move together with no external influences; such natural clocks are spacetime correlates of stiff bodies.

Then (in analogy with L) for any events, p and q: let \( l \) be any inertial trajectory through p, let \( t_1 \) be the clock time until p after the point on \( l \) at which light would have to be emitted in order to reach q, and let \( t_2 \) be the clock time after p until the point on \( l \) at which light would be received from q. Then in any inertial coordinates, \( t_1t_2 = (\Delta x/c)^2 - \Delta x_0^2 \) (where \( \Delta x^2 = \Delta x_1^2 + \ldots + \Delta x_{d-1}^2 \)). See Figure 1 (and Wald 1992, chapter 2).

![Figure 1](image-url)  
**Figure 1** In a system of inertial coordinates, (a) shows \( l \), an inertial trajectory (satisfying the law of inertia) of a clock passing through p and the light paths (satisfying the light postulate) connecting \( l \) to q. The times measured on the clock between the events indicated satisfy \( t_1t_2 = (\Delta x/c)^2 - \Delta x_0^2 \). (b) shows a similar set-up, except that p occurs before the point on \( l \) from which a light signal reaches q, so that \( t_1 < 0 \), and \( \Delta s^2 \equiv t_1t_2 < 0 \).
The situation is completely analogous to the Euclidean fairy tale: the properties of matter, forces and fields are described by a formal theory postulated to hold in some coordinates. Moreover, the fragment again suffices to give physical significance to a metric: any points p and q are assigned a physical ‘distance’ $\Delta s^2 \equiv t_1 t_2$, or in inertial coordinates $(\Delta x/c)^2 - \Delta x_0^2$.

Because it measures extension in space and time, we call $\Delta s$ the ‘interval’. If $\Delta s^2 > 0$ then the ‘distance’ is ‘spacelike’; if $\Delta s^2 = 0$, ‘lightlike’; and if $\Delta s^2 < 0$, ‘timelike’ (with $\Delta s = -\sqrt{-\Delta s^2}$). Thus a notable difference from spatial geometries is that in ‘pseudo-Riemannian’ spacetime geometries the metric can give curves non-positive lengths. This particular metric looks like the Pythagorean formula, so call it pseudo-Pythagorean; as Euclidean geometry is that for which straight lines have lengths given by the Pythagorean formula (in suitable coordinates), so ‘Minkowski’ geometry is that for which geodesics have intervals given by the pseudo-Pythagorean formula (in suitable coordinates). And hence, as in the fairy tale, the theory fragment can be expressed geometrically: spacetime is Minkowski, light travels lightlike geodesics, and free bodies travel along timelike geodesics, whose intervals are measured by clocks. Indeed the whole theory can be expressed in terms of Minkowski geometry. And so the same question comes up, for much the same reasons: is the geometric formulation just a convenient representation of the physics of matter and forces, or should it be taken literally, as describing some intrinsic feature of spacetime?

Next, it is easy to check that if the fragment holds in some coordinates, then it holds in any other related to it by the ‘Lorentz transformations’ (in one spatial dimension):

$$x_0' = (x_0 - vx_1/c^2)/\sqrt{(1 - v^2/c^2)}$$
$$x_1' = (x_1 - vx_0)/\sqrt{(1 - v^2/c^2)},$$

where v is the velocity of the primed inertial coordinates, $(x_0', x_1')$ relative to the unprimed ones. Indeed, the same is true of the full quantum theory of matter and force; it follows that nothing in the physics of the theories privileges any of those coordinates, which is the statement of the ‘special principle of relativity’. It’s now straightforward to demonstrate and understand the various surprising features of SR: relativity of simultaneity, length contraction, time dilation and so on.

For instance, consider a moving clock: transform to inertial coordinates in which it is at rest; the theory dictates that its ticks mark off that time coordinate, while it travels no distance between ticks; Lorentz-transforming back to the original coordinates, one calculates that in those coordinates more time than one tick’s worth passes per tick. Since the theory holds in any inertial coordinates, it follows that if we modelled the moving clock according to the
laws expressed in the original coordinates, then we would find that it ran slow relative to its rest state. But the relativity principle means we don’t have to know the details of its construction to reach that conclusion; the argument I just gave suffices!

Still, it is perfectly proper to explain its dilation in terms of its physics, in the original coordinates. (Just as it would be proper to explain the dilation of a clock at rest in the original coordinates, relative to the second set of coordinates.) The technique of using relativity to solve a problem by transforming to coordinates in which it is simplified is crucial in physics, and was pioneered by Christiaan Huygens (Blackwell 1977) in the context of Newtonian collision theory. It shows what the laws entail for physical systems in the original coordinates, given what they entail in the other.

Bringing everything together, in the fairy tale, that the theory is expressed in terms of the Pythagorean formula is necessary and sufficient for its having a geometric representation in Euclidean space. It means that in the geometric representation, the theory depends only on the Euclidean metric – since that is represented by the Pythagorean formula. It follows immediately that the theory will have the symmetries of Euclidean space: it’s easy to check that if the theory holds in one set of coordinates, it will hold in any that differ only by translations and rotations. This ‘Euclidean relativity’ is the analogue of the special principle, and indeed the analogy is complete; that a theory is expressed in terms of the pseudo-Pythagorean formula is necessary and sufficient for its having a representation in Minkowski spacetime (because the formula is the coordinate form of the metric), whose symmetries include the Lorentz transformations. These connections between symmetries, formulation in terms of the coordinate form of a geometric function (here the metric), and a representation in a geometry picked out by the function (hence with the same symmetries) generalize, so many theories of physical systems in space or time have geometrical representations.

2.1.3 General Relativity
Now we bring together both our generalizations of Euclid in a theory of curved spacetime, which includes gravity, unlike special relativistic theories. At the core is the ‘Einstein field equation’ (EFE), written schematically as:

$$G(g(x_\alpha, x)) = E(x_\alpha, x)$$

where $g$ represents the gravitational field and $E$ the distribution of (‘stress’-) energy in the form of matter, other forces or gravity itself. The EFE represents how gravity interacts with any given system of matter; $g$ is not directly equated to $E$, the $G$-function represents the more subtle formula.
So far things are much as before, with a theory describing the physics of systems (including gravity) in spacetime. But strikingly, g has the formal properties of a pseudo-Riemannian metric, assigning numbers to curves just as such a metric assigns them spacetime intervals. (Quickly: the g-field assigns a ‘g-interval’ to infinitesimal curves through any point; the number assigned to any curve is obtained by adding up – integrating – the g-intervals of the infinitesimal segments along it.) Moreover, unless $E=0$, g is not a flat metric, like the Minkowski one, but curved.

That’s interesting, but it doesn’t imply that those numbers are the intervals; abstractly, a field with the formal properties of a metric can live in a spacetime whose geometry is given by another metric. The real revolution of GR is that g plays the very same role played by the metrics in our earlier examples: according to the EFE it connects up to physical systems in the same way as they do, so it represents physical geometry. For instance, light travels the lightlike geodesics of g and clocks measure the interval along g’s timelike geodesics.

The motion of free bodies is especially interesting; since g plays the same role as the Minkowski metric, according to the law of inertia, inertial trajectories must be the timelike geodesics of g. However, the idea of ‘force-free’ motion must be revised, for gravity is already represented by the g-field, and is not some force that can either act or not – the metric never vanishes, never fails, for instance, to pick out geodesics. That is, ‘free’ now means ‘no non-gravitational force acts’, and the ‘law of inertia’ expresses the action of the gravitational field; bodies with no non-gravitational forces follow the geodesics of the gravitational metric. That the g-field plays the role of the metric in the EFE is expressed formally in the ‘strong equivalence principle’ (SEP). It follows immediately that all bodies fall at the same rate in the same gravitational field, since they follow the same geodesics regardless of mass; the equivalence principle is often expressed informally as this consequence.

If g describes a curved spacetime geometry, there will (generally) not be any ‘special’ coordinates: all coordinates are ‘equal’, there are no symmetries (or, in a sense, everything is a symmetry). Thus the representational story developed earlier no longer applies; in GR, the metric does not arise because the spacetime symmetries of the theory pick it out, it is one of the physical objects of the theory, and there is no question of its merely being representational.

**2.2 Issues**

This presentation of GR has taken quite a different tack from the usual one, emphasizing from the start the relation of geometry to physical fields, rather than introducing the theory in terms of abstract geometry. In part I have chosen this approach because multiple views facilitate understanding, but
there is also a philosophical point: if one thinks of GR primarily as a theory of geometry, then a ‘geometric’ interpretation seems almost inevitable, but by emphasizing the ‘dynamical’ role of geometry, another understanding seems equally plausible. Recent work on spacetime has focused on the competition between these two interpretations.

The work of Earman (1970) and Friedman (1983) (especially) made the geometric interpretation dominant; their work was in many ways a product of the resurgence of scientific realism of the time, and certainly responded to earlier, more positivistic approaches to spacetime physics. However, Brown (2005) (also DiSalle 2006, and Huggett 2009) has recently invigorated the dynamical approach, though not without opposition (e.g. Norton 2008, Janssen 2009). Some arguments rely on metaphysical considerations, or appeals to best explanation, but here I will just mention two physical considerations. On the one hand, the energy of the g-field is generally not well defined, leading Pooley (2006, p. 101) to argue that it cannot be identified as a normal physical field; what else, then, but geometry? On the other, one can have spacetime theories with two metrics, one of which represents the ‘real’ geometry while the other satisfies SEP; in such a theory the geometry manifested by matter is that of the latter. Brown (Section 9.5) argues that the physical significance of the g-field is always through the SEP and the behaviour of matter.

Aside from its bearing on the significance of physical geometry, GR has philosophically important cosmological implications. First, while there is no way to define a preferred present in terms of the geometry of Minkowski spacetime, it is possible in some models of GR. Indeed, in the standard model of a uniform (on large scales), expanding universe, it is possible to define a privileged ‘foliation’ (a partition of spacetime into three-dimensional spaces). However, there are also models in which there is no preferred foliation, and even some in which no foliation at all is possible (Gödel 2000); if the present corresponds to a preferred foliation, then it is not a general feature of GR, but only of contingent circumstances.

Second, there are models in which the EFEs hold in general, but which contain ‘singularities’ of some kind, and the equations no longer make formal sense: for instance, spacetime curvature becomes infinite. A natural response is to think that such models are unphysical. However, this response is undermined by theorems showing that singularities are the norm in universes resembling ours, and more particularly because the believed current state of the universe is one that implies a singularity a finite time in the past – the famous big bang (see Wald 1992, chapter 4). Once that singularity is accepted, then there is no obvious reason to deny others, especially those associated with black holes. These lead to a number of interesting problems: for instance, the breakdown of the EFEs amounts to a breakdown of determinism, perhaps hidden from us by a kind of ‘cosmic censorship’ (discussed by Earman 1995).
3. The Second Pillar: Quantum Mechanics

3.1 The Theory
When atoms emit light, they only do so at certain discrete frequencies, in discrete packets of energy – photons – of certain allowed amounts. It follows that electrons only have discrete allowed energy levels in an atom; a photon is emitted when an electron drops between levels, and its energy is the difference between them. This situation is puzzling, because in the similar-seeming case of a planet orbiting the sun, the system can have any energy, depending on the radius of the orbit. While I will make no attempt at responsible history here, this ‘spectrum problem’ was crucial in the development of quantum mechanics (QM); for example see Pais (1988, chapter 9).

The example is intended to motivate the following, but we will approach QM by first looking at an abstract, general formulation, and then come back and see how a concrete realization solves the spectrum problem. So first think about vectors in two dimensions, lines directed from the origin to any point: any two can be added to obtain another and any can have its length changed by multiplication by a real number. In mathematical terms, the system is a ‘vector space’. Moreover, there is a ‘dot product’ between any two vectors in the space, a measure of ‘how much’ they point in the same direction: as special cases, a vector ‘dotted’ with itself equals its length squared, and with a vector at right angles equals zero. In mathematical terms, there is thus an ‘inner product’, \( \langle \mathbf{v}, \mathbf{w} \rangle \), a kind of multiplication between vectors. Such a vector space (assuming certain conditions on the inner product) is a ‘Hilbert space’, the abstract framework of QM.

Here the two-dimensional picture will generally suffice, but QM generalizes: first, the relevant spaces are ‘complex’, meaning that vectors can be multiplied by complex numbers; second, they are multi-, even infinite-, dimensional. But even in these cases, certain features carry over. First, the inner product is ‘linear’: using Greek letters for vectors and roman for the numbers that multiply them, we have \( \langle c\psi, \phi \rangle = c \langle \psi, \phi \rangle \) and \( \langle \psi + \phi, \chi \rangle = \langle \psi, \chi \rangle + \langle \phi, \chi \rangle \) and so on.

Second, we can define ‘operators’, maps from any vector to another in the same space: so \( \psi \) maps to \( O\psi \). For example, from each vector to one twice as long, or to a unit vector along the x-axis, or to the vector at \( \theta \), and so on. Of special interest are linear operators, \( O \): \( O(c\psi + \phi) = c(O\psi) + O\phi \) (these can be written as matrices). Finally, some operators have ‘eigenvectors’. When an operator acts on one of its eigenvectors the result is the same vector, up to a factor, called an ‘eigenvalue’: if \( \psi \) is an eigenvector of \( O \), then \( O\psi = \xi \psi \). For example, consider the operator that reflects the two-dimensional space about the x-axis: most vectors are mapped into new directions, but vectors along the x-axis are unchanged (eigenvalue 1) while those along the y-axis are reversed (eigenvalue \(-1\)) – these are the eigenvectors.
Now for the physical application of this formalism.

1. States of systems are represented by unit ($\psi \cdot \psi = 1$) vectors in a (complex) Hilbert space.

2. Physical quantities, or ‘observables’, are represented by linear operators with real-valued eigenvalues (more precisely, by ‘Hermitian’ operators); according to the ‘eigenvector–eigenvalue link’, $\psi$ represents a state in which the observable represented by $O$ takes the value $o$ if $O\psi = o\psi$.

3. The ‘Born Rule’: assuming $O\psi = o\psi$, if the state of a system is $\phi$, then the probability that a measurement of $O$ will have outcome $o$ is $Pr(O = o) = (\psi \cdot \phi)^2$. (Since the space is complex, squaring means taking the modulus squared, multiplying $\phi \cdot \psi$ by its complex conjugate.)

4. If the result of the measurement of $O$ is $o$, then afterwards the system will be in state $\psi$ (hence the probability of the same result in an immediately repeated measurement is one).

The reader should be struck by this theory; things start familiarly, assigning physical states and quantities to objects in a mathematical space – much as positions are assigned points in a three-dimensional space, and velocities to vectors in that space. But by the time we reach the Born Rule, something strange has happened, for we are explicitly talking about predictions for measurement outcomes. We are used to physics describing the behaviour of basic entities, independent of measurement; indeed, we expect to have to model experiments in terms of the physics of the system and apparatus to derive predictions. What are we to make of a theory formulated in terms of ‘measurement’? We will return to this question later; for now we will take (1–4) simply as an algorithm for generating predictions.

The following is an example: a ‘particle’ in one dimension, the $x$-axis. Consider functions of position, $\psi(x)$ and $\phi(x)$, say. They can be multiplied by complex numbers and added together to get new functions; they satisfy the abstract concept of a vector, so families of them form vector spaces. Moreover, we can define an inner product:

$$\psi \cdot \phi = \int dx \, \psi^*(x)\phi(x).$$

An integral is the area under a function, so the inner product is the area under the product of the functions, a sensible measure of their overlap. (We use $\psi$’s complex conjugate $\psi^*(x)$ for formal reasons.) Thus we have an (infinite dimensional) Hilbert space. Unit ‘vectors’ satisfy $\int dx \, \psi^*(x)\psi(x) = 1$; by (1) these ‘state-vectors’ or ‘wave functions’ represent the states of a particle.
To illustrate the rest of the algorithm, we will consider two observables, position and momentum. By (2), position is represented by the operator, $X$, whose eigenvalues are possible positions, $x$; unsurprisingly, the corresponding eigenstate is (heuristically) $\delta(x)$, a wave function zero everywhere except $x$. (Glossing over technical details; such a ‘Dirac-$\delta$’, is not, strictly speaking, a function – hence $X$ is not, strictly speaking, an operator.) Then (3) allows us to calculate the probability that the position of a particle in state $\psi(x)$ is measured to be $q$:

$$\Pr(X=q) = (\psi(x) \cdot \delta(x-q))^2 = (\int dx \psi^*(x)\delta(x-q))^2 = \psi^2(q).$$

Heuristically, the area under $\delta(x-q)\psi(x)$ equals $\psi(q)$ because $\delta(q)$ is a function with unit area all located at $q$, which is multiplied by $\psi(q)$ at that point. Let’s label this result:

$$\Pr(q) = \psi^2(q).$$

Because the probabilities involve squaring, they are unchanged if a wave function is multiplied by any number whose square is unity: for example, $(-1 \cdot \psi(q))^2 = (-1)^2 \cdot \psi^2(q) = \psi^2(q)$. $-1$ is not the only number having this property: any complex number $k = a + ib$, such that $a^2 + b^2 = 1$, does. Hence, according to (3), $\psi(x)$ and $k \cdot \psi(x)$ predict the same frequencies in position measurements – such experiments cannot distinguish them. The same goes for all states and observables, so we say that $\psi(x)$ and $k \cdot \psi(x)$ represent the same physical state. Indeed, we go further. Vectors related by multiplication by any complex number form a ‘ray’, and we modify (1): rays represent physical states. (And (3): use a unit vector from the ray to calculate probabilities.)

An important step in the development of QM was De Broglie’s postulate that a particle with momentum $p$ is associated with a wave of wavelength $\lambda$ given by $\lambda = 2\pi\hbar/p$. Or, in our (later) framework, the eigenstates of momentum are regular, ‘plane’ waves, such as

$$\psi(x) = A \cdot \cos \frac{2\pi x}{\lambda}.$$
circumference, $2\pi r$, of the orbit, so its ends join up smoothly: $\lambda = 2\pi r/n$. By De Broglie’s formula, the momentum eigenstates thus have momentum $p = \hbar n/r$; hence the allowed kinetic energies are $E = \hbar^2 n^2 / 2mr$ ($n = 1, 2, \ldots$). This model reproduces the ‘quantization’ of energy that is demonstrated by the atom – and illustrates how vector-talk lies behind the relevant physics. (Stipulating a fixed orbital radius did most of the work here; a realistic calculation involving electromagnetic forces reproduces quantization; see Griffiths (2005, pp. 145–60).)

Another important example of the Hilbert space approach involves ‘spin’. Electrons have a property akin to – but not quite the same as – the angular momentum of a spinning top. Suppose a top has angular momentum $J$ about its axis of rotation, which is not necessarily vertical: the angular momentum in the vertical, say, depends on the projection of the axis into the vertical. Hence, it takes any values between $\pm J$, as the top is oriented continuously from pointing up to pointing down. However, electron spin is quantized: empirically, when an electron’s spin is measured in the vertical, the magnitude is always the same, and the only variation is in the orientation, up or down.

So the appropriate quantum vector space is two dimensional; one unit vector, $\uparrow$, represents spin up, and another at right angles represents spin down, $\downarrow$. Why at right angles? The Born Rule tells us that in state $\uparrow$ the probability for measuring spin up is one; so the probability for measuring spin down – that is, for the state being $\downarrow$ – must be zero. And only vectors at right angles have zero dot products, hence zero probabilities by the Born Rule.

What are we to think of the other state-vectors in the space, say $\psi = 1/\sqrt{2} (\uparrow + \downarrow)$, which is neither up nor down? ($1/\sqrt{2}$ normalizes the state.) Intuitively, an electron in state $\psi$ appears to have zero vertical spin, but we have seen that vertical spin measurements never have that result. We will say more about this below; here consider what (1–4) entail about a spin measurement. Using the linearity of the inner product, the Born Rule tells us

$$\Pr(\uparrow) = (\uparrow \cdot \psi)^2 = (\uparrow \cdot 1/\sqrt{2} (\uparrow + \downarrow))^2 = 1/2 (\uparrow \cdot \uparrow + \uparrow \cdot \downarrow)^2 = 1/2 (1 + 0)^2 = 1/2,$$

(remember, $\uparrow$ is a unit vector, at right angles to $\downarrow$). Similarly, $\Pr(\downarrow) = 1/2$.

Pick units in which ‘up’ equals $+1$ units of spin and ‘down’ equals $-1$ units of spin; the average, or ‘expected’ spin is $(1/2 + 1) + (1/2 - 1) = 0$ – that is the sense in which $\psi$ means zero vertical spin. Physically, the quantum algorithm tells us the frequencies of outcomes of measurements on an ‘ensemble’ of systems in some state; hence average values for quantities measured on such an ensemble. An average value calculated this way is called an ‘expectation value’ (according to 2–3, for any observable this is the sum of the products of the eigenvalues and corresponding probabilities).
3.2 Issues

So much for the quantum algorithm. Its extraordinary power and empirical success make it sensible to ask whether it is more than an algorithm, whether it is an accurate description of the world; and if not, what kind of physics lies behind it, producing the predicted outcomes. Certain kinds of positivism or instrumentalism are happy with nothing more than a predictive algorithm, and many physicists are content, it seems, to take QM that way. However, to do so is to make a serious change in the goals of physics, one highly questionable, historically: for instance, because Newton was not satisfied with an algorithm for predicting the positions of the planets, but sought to explain them, physics started on a path to the current understanding of the world. So we will press the question of interpretation.

The first problem with treating the algorithm as a description is the existence of states like \( \psi = \frac{1}{\sqrt{2}} (\uparrow + \downarrow) \), a ‘superposition’ of up and down. Given (1–2), an electron in such a state seems not to be in any state of (vertical) spin, since up and down are the only options. The standard move is to say that its spin is ‘indeterminate’ or ‘ill-defined’; there is no fact of the matter, and even to ask about it is to make some kind of category error. If one wants to talk about, let alone know, the spin, one has to contrive a situation in which the spin is measured, and (2–3) kick in. Such a view is generally called the ‘Copenhagen interpretation’ because of work done by Niels Bohr and disseminated by his students; whether it accurately captures Bohr’s ideas is debatable, so we will just call it the ‘standard’ interpretation (Howard 2004).

Of course, expectation values are well defined in any state, so one can think of these in a general sense as the ‘values’ taken by an observable, even in a superposition. But they have to be thought of statistically, not as actual possessed values, if the state is ‘indeterminate’ for an observable.

The idea of indeterminacy is so bizarre that one wonders why it might be entertained. (The bizarreness lies not in the indeterminism of the algorithm, but in its indeterminacy – it’s not that things are chancy, but that the ‘reality’ of basic properties changes over time.) The answer is the crucial quantum phenomenon of ‘interference’, illustrated by the ‘two-slit’ experiment.

If a low intensity beam of photons is shone through a narrow slit, then, on a screen on the other side, the photons will form a distribution whose shape is a ‘Gaussian bell curve’, centred opposite the slit. (The experiments are harder, but all these points hold for electrons, too.) That observed outcome is the prediction of the quantum algorithm; if \( x \) is the distance to the left or right of the slit, a detailed calculation shows (using (+)) that

\[
Pr(x) = |\psi|^2(x) \propto \exp(-x^2/a),
\]
The Continuum Companion to the Philosophy of Science

Figure 2  (a) Shows that the Gaussian formed on a screen by photons passing a single slit \( Pr(x) \) can be thought of as the fraction of photons arriving at a certain spot and as the predicted quantum probability. (b) Shows first, in grey, the pattern expected on an ignorance understanding of the experiment. The actual result, predicted by QM, showing interference, is shown (impressionistically) in black.

(a) Shows that the Gaussian formed on a screen by photons passing a single slit \( Pr(x) \) can be thought of as the fraction of photons arriving at a certain spot and as the predicted quantum probability. (b) Shows first, in grey, the pattern expected on an ignorance understanding of the experiment. The actual result, predicted by QM, showing interference, is shown (impressionistically) in black.

a Gaussian (see Figure 2). It is also the common sense expectation if one imagines photons as little particles, with the wave function describing a spread of trajectories through the slit. So far, then, we have no need to appeal to ‘indeterminacy’ in position, just ignorance about detailed motions, in a rather familiar way. Things change if there are two adjacent slits.

Assuming mere ignorance of photon trajectories, we expect each photon to go through one slit or the other, resulting now in a Gaussian opposite each slit. Suppose \( \psi_L \) and \( \psi_R \), are the wave functions that result when only the left and right, respectively, slits are open; for example, \( \psi_L^2(x) \) is a Gaussian centred opposite the left slit. On the ignorance view, photons going through the left will distribute according to \( \psi_L^2(x) \) and those going through the right according to \( \psi_R^2(x) \), so the outcome will be proportional to \( \psi_L^2(x) + \psi_R^2(x) \), a ‘double hump’. But that is neither what is observed, nor what is predicted by QM. According to QM, the final state vector is the sum (normalized to unity) of these wave functions, \( 1/\sqrt{2} (\psi_L(x) + \psi_R(x)) \), and by (+) the probability is

\[
Pr(x) = \frac{1}{1/\sqrt{2}2} (\psi_L^2(x) + \psi_R^2(x)) \]

\[
= 1/2 (\psi_L^2(x) + \psi_R^2(x) + \psi^*_L(x) \psi_R(x) + \psi_L(x) \psi^*_R(x)).
\]

The final terms are the ‘interference’ between the two wave functions, formally exactly the same kind as that familiar from water and sound waves. It is the source of the distribution pattern observed, which has a peak opposite the middle of the slits, then a series of decreasing peaks. If the photons were really just passing through one slit or the other in a random manner, then the mere presence of the other slit could have no effect on the photon, and it could not cause interference. The presence of interference means that we cannot
simply take the wave function to express ignorance about position; indeterminacy about position then seems to be an alternative.

Given its bizarre nature, that proposal has been the subject of searching critiques. First, Einstein, Podolsky and Rosen invoked ‘locality’ in an attempted refutation (Einstein et al. 1935). It turns out in QM that states of definite position have indefinite momentum, and vice versa (i.e. there are no eigenstates of both observables). That’s one way of putting Heisenberg’s ‘uncertainty principle’ (in the standard interpretation, anyway). Now, QM allows states of two particles with zero total momentum, but indefinite momentum for either particle. Suppose such a pair flies apart, until I investigate the left one. If I measure its position, I can use the quantum algorithm (roughly speaking) to determine the position of the other. If I measure its momentum, then I can determine the other’s momentum (equal and opposite by conservation). Moreover, such predictions are born out by experiment – the particle on the right always has the predicted properties. Locality, Einstein claimed, implies that neither my decision about what to measure, nor the outcome can have instantaneous effects on the right; so it must be the case that the particle on the right already had position and momentum, right from the start. The argument is thus meant to show that, given locality, QM itself shows that the wave function is not the whole story, that the algorithm is an ‘incomplete’ theory.

The idea is to find more variables (often called ‘hidden’, even when they are observable!) to complete the theory; variables sufficient to make all basic quantities take definite values. However, Bell and others showed that no such project was possible, on pain of violating locality; the verified predictions of QM are incompatible with a ‘local hidden variable’ theory (Bell 1964). That may appear a victory for the standard interpretation, but not so; as Einstein’s argument shows, that interpretation implies that the measurement on the left has the instantaneous effect of making definite a property on the right. (Some suggest that the ‘effect’ is merely in our subjective knowledge, but it is surely worth resisting importing subjectivity into physics.) Local hidden variables were the only way to make local sense of the predictions of QM; Bell ruled out that final hope – better, the experimental verification of non-local correlations did. Clearly then, an important foundational question concerns the compatibility of QM with relativity theory, which imposes constraints on how causal effects can propagate; how does that notion of ‘locality’ relate to that in play here? (Maudlin 1994, especially chapters 3–5).

The second kind of criticism of the standard interpretation concerns the ‘measurement problem’. (4) asserts that on measurement of a superposition, the state ‘jumps’ to an eigenstate: from a spread out wave, $\psi(x)$, to a definite position, $\delta(x)$; or from neither up nor down, to up, say. However, according to QM, at times other than measurement, the state vector evolves continuously (and deterministically) in Hilbert space (the ‘unitary’ or ‘Schrödinger’
dynamics). So only the activity of ‘measurement’ distinguishes the cases, determining which dynamics applies. Bell argued that such a concept has never been defined in the precise way necessary for a fundamental theory; we merely have a pragmatic notion suitable for an algorithm (Bell 2004). So there is a serious – perhaps irredeemable – lacuna in the algorithm if one wants to take it as a full description.

No wonder, then, that there are many proposals to modify or replace QM with a theory that agrees with the predictions of the algorithm. It must be acknowledged that none receive significant support from physicists, outside of specific disciplinary communities; for one thing, the pragmatic success of the quantum algorithm suffices for most purposes to which physicists have put it. But some very brief examples. First, suppose each photon goes through just one of the two slits; it is still the case that the wave function is different than when there is just one slit – it ‘goes through’ both. So perhaps the wave function ‘guides’ the particle, and the different form in the one and two slit cases means that particles are guided to different points on the screen. The De Broglie-Bohm ‘pilot wave’ theory works in this way. Second, one might address the measurement problem by adding jumps to the dynamics without reference to ‘measurement’; the Ghirardi-Rimini-Weber-Bell approach, for instance, postulates a small chance per particle per unit time of the wave function localizing in space. Finally, perhaps there are no jumps at all. For instance, perhaps the quantum probabilities describe frequencies over a large ensemble of co-existing realities – perhaps many worlds, if the frequencies are over arrangements in space, or perhaps many minds, if the frequencies are over mental states. For discussion and references, see Albert (1992, chapters 5–7).

4. The Third Pillar: Statistical Physics

4.1 Classical Statistical Mechanics

4.1.1 The Theory

The industrial revolution is a terrific example of how entirely practical matters can drive advances in theoretical physics. For it was solving engineering problems concerning the construction of maximally efficient steam engines that led to the development of thermodynamics (TD), which was subsequently understood within statistical mechanics (SM).

Although the behaviour of physical systems depends on their microscopic parts and the laws governing them, many robust (or ‘projectable’) properties involved in physical laws are independent of the finer details of such composition. When we consider such properties while ignoring a system’s microscopic details, we call it ‘macroscopic’; the term primarily refers to the perspective we take. Some macro-properties (e.g. momentum) are involved in mechanics, but
others fall under other laws: temperature is proportional to the product of pressure and volume, energy is conserved and so on. Broadly, TD is the theory of such properties and their laws. Most famously, there is the second law of thermodynamics (TDII), according to which entropy – another macro-property – can never decrease in an isolated system. For instance, a steam engine works by heating water until it expands as steam and pushes a piston. After, one would like to recycle the hot steam to produce more mechanical work. But less heat means less entropy, so TDII limits how much heat can be reused.

Many other processes produce or redistribute heat: a vase smashing to the ground heats the floor up a little, increasing entropy; an ice cube melts as heat is more evenly distributed between it and its surroundings, again increasing entropy. TDII is compatible with such processes, but prohibits their time reverses (and many, many others), giving a lawful explanation of the invariable orientation of time in many instances. There is a considerable literature (also considerable confusion) about what an ‘arrow of time’ might be, but, at least for present purposes, it can simply be taken to be a time asymmetric law: one that prohibits the time reverses of certain allowed processes. Thus TDII is an arrow of time.

That is not to say, however, that it is a ‘fundamental’ arrow, for TDII is not a ‘fundamental’ law. Rather, it is what Einstein described as a ‘principle’ law, one that states a profound empirical regularity; in his scheme, principle laws constrain ‘constructive’ laws, which explain the phenomena. For TD, the constructive theory (due to Boltzmann) appeals to microphysics. Suppose the system is composed of particles with the same possible states of positions and momenta; a ‘microstate’ of the system is an assignment of such states to the particles. Supposing Newtonian mechanics, the microdynamics is deterministic; the microstate fixes the future evolution. Macrostates ignore microdetail, so they are collections of microstates.

Consider the microstates within any macrostate. Postulate that the fraction which will evolve into a macrostate $\vartheta$ in the near future equals the number of microstates in $\vartheta$ divided by the total number of microstates, the fraction of the statespace occupied by $\vartheta$. That is, postulate that the system is equally likely to end up in any equally sized region; the dynamics do not favour any.

That’s in the ballpark of TDII: if all states are equally likely, then as time goes on the system will evolve into larger macrostates. The ‘current size’ should grow, like entropy, though only probabilistically.

We can do much better. First, ‘ordered’ microstates are much less common than ‘disordered’ ones. For instance, there are $c^n$ microstates in which each of $n$ particles is in one of $c$ states (each of the $n$ has $c$ ‘choices’). When, say, 100 molecules spread from 10 states to 20, the number of microstates increases by a stupendous factor of $\sim 10^9$, and the effect grows with the system. In short, if $\vartheta$ is more disordered than $\phi$, then $\vartheta$ is vastly bigger than $\phi$, and the current state is, correspondingly, vastly more likely to evolve to $\vartheta$ than $\phi$.
Second, higher entropy macrostates do correspond to greater ‘disorder’: when the smashing vase heats the floor up, its molecules have greater energy and so jiggle in more ways; when the ice cube melts, its molecules also have more vibrational states. Entropy-increasing processes indeed correspond to enormous increases in macrostate size – we can take the entropy of a macrostate as a measure of its size.

Thus the dynamical postulate does explain TDII: it is overwhelmingly likely that a system will evolve towards increasing size, disorder and entropy. (Until a state of maximum entropy – ‘equilibrium’ – is reached; then our reasoning implies that most states evolve into other equilibrium states.)

4.1.2 Issues
This explanation changes the content of TDII: the increase in entropy is not literally inevitable, rather overwhelmingly probable. Such a modification is readily comprehensible: TDII was extracted from phenomena, and hence by induction; understanding it in more fundamental terms reveals a lacuna in that induction.

But there are problems. First, in classical mechanics, a microstate and its time reverse correspond to the same macrostate: the microstate describes particles’ locations and velocities, but the macrostate doesn’t care whether a particle is moving from A to B or from B to A, only how fast. Moreover, the microlaws are time symmetric, and allow the time reverse of any allowed process. Hence within any macrostate, for every microstate leading to higher entropy in the future, there must be another that has just come from a higher entropy macrostate.

Therefore, the fraction of microstates leading to higher entropy must equal the fraction that came from higher entropy in the past; thus the probability that entropy will increase equals the probability that it was decreasing. By our earlier reasoning, the former is overwhelming, making it overwhelmingly likely that (if it is not at equilibrium) any system evolved into its current state by a steady drop in entropy!

So, our reasoning shows both that TDII will hold in the future and that its converse held in the past! That conclusion is irreconcilable with TDII: vases have not been ‘unsmashing’ until the present, and water has not been spontaneously emitting heat and freezing. It is at odds with the inductive evidence for TDII, and so undermines our reasons for believing it in the future! Indeed, an anti-entropic past is so different from our memories and records that the conclusion undermines any faith in them whatsoever – and hence undermines all inductive science, including the very microphysics that led us here.

To recover TDII in the past, we must postulate that entropy was lower at some specified earlier time, simply ruling out almost all the pasts compatible
with the present macrostate. That such a postulate is needed was realized early on; Albert was the first to clearly explain that it is the only way to avoid an epistemic meltdown, and hence has a ‘transcendental’ support as the only picture of the world in which we could have reliable knowledge of the past.

Our difficulties are not over, for when should the earlier low entropy be postulated? Pragmatically, we postulate it at the start of whatever past time period interests us: for the vase, some time before it smashed; for the ice cube, when it was placed in the room. More precisely, we can take it to be when the system was created, or more or less isolated from the rest of the world; this is Reichenbach’s ‘branch system’ approach. However, one can always ask how a branch got to be in its initial low entropy state. Since it is incredibly unlikely to have happened by chance, the only explanation is that it was a part of some larger branch, which itself was at low entropy at an even earlier time (as the ice cube was part of a larger ‘branch’ including a freezer and its power source). But now we have a regress. The obvious response is to postulate low entropy for the initial – ‘big bang’ – state of the universe as a whole. But there are technical problems with incorporating gravity into the notion of entropy, and even with defining a meaningful concept of entropy in the early stages of the universe (Earman 2006). An alternative is to postulate some specific initial macrostate for the universe as a whole – a state as yet only partially discovered by cosmology. Albert calls such a postulate the ‘past hypothesis’ (Albert 2000, chapter 4). The claim is that such a state is overwhelmingly likely to evolve into a low entropy state for the universe and its subsystems, at the time when such notions are well defined. Then our earlier reasoning shows that TDII holds. Loewer develops this idea further, claiming that statistical mechanics thereby amounts to a complete theory of objective probability: the conditional probability of A on B is the relative volume of states compatible with A, in the set of microstates compatible with B and the big bang macrostate (Loewer 2004; see Winsberg 2008 for objections).

The question of its explanation aside, TDII is highly suggestive regarding deep philosophical issues: could this temporal arrow explain others? That causes precede their effects, that we can influence the future, not the past, and that knowledge of the past differs from that of the future? Is there an explanatory path through these arrows, or are they independent? The literature is beyond citation here.

Finally, note that according to our current best physics, there is a fundamental arrow of time, in the form of the weak interaction between nuclear particles. This arrow was not brought to bear in our discussion, because the effect is too attenuated to explain the phenomena. Maudlin, however, takes it seriously (2007, §5).
4.2 Quantum Statistics

4.2.1 The Theory
Consider how collections of particles (indeed, any systems) are represented in QM (we tacitly assumed the following in modelling measurements and non-locality). First, suppose each is assigned a specific single-particle state: suppose there are \( n \) in states \( \alpha, \beta, \ldots \gamma \) (which may or may not be distinct). Then a suitable state for the whole is a product (using subscripts to associate states with particular particles):

\[
\Psi = \alpha_1 \cdot \beta_2 \cdot \ldots \cdot \gamma_n. \tag{*}
\]

(Technically, ‘\( \cdot \)’ is the ‘tensor product’.)

We can easily extend the formal apparatus of QM to such states.

(1) Specify that the inner product with another state, \( \Psi' = \alpha'_1 \cdot \beta'_2 \cdot \ldots \cdot \gamma'_n \), is obtained by multiplying the inner products (which are numbers) for the individual particles:

\[
\Psi \cdot \Psi' = (\alpha_1 \cdot \alpha'_1) \times (\beta_1 \cdot \beta'_1) \times \ldots \times (\gamma_1 \cdot \gamma'_1).
\]

(2) Operators include products of operators on the single-particle states, \( O = A_1 \cdot \beta_2 \cdot \ldots \cdot \gamma_n \):

\[
O \Psi = A_1 \alpha_1 \cdot \beta_2 \cdot \ldots \cdot \gamma_n.
\]

(3) Superposition means that sums of vectors and operators are also vectors and operators: for instance, \( \Psi + \Psi' \) is also an (unnormalized) vector. But as before, both inner product and operators are linear: for example,

\[
(\Psi + \Psi') \cdot \Phi = \Psi \cdot \Phi + \Psi' \cdot \Phi, \text{ and } (O + O')(\Psi + \Psi') = O\Psi + O\Psi' + O'\Psi + O'\Psi',
\]

and so forth.

Identical particles have both the same intrinsic properties and possible states. (Particles are ‘exactly’ non-identical only when they differ in strictly constant properties, but they can be ‘approximately’ non-identical if they have properties that are at least approximately different for the life of the system: for instance, particles separated in space. Then they can be formally treated as non-identical particles, and the following is ignored in practice.) Suppose the state space is \( m \)-dimensional; let \( \alpha, \beta, \ldots \gamma \) be a set of \( m \) orthonormal vectors. Each of the \( n \) particles can be in any of the \( m \) states, so we have...
n×m states like (*). By (1) and the orthonormality of α, β, . . . γ these are orthonormal; moreover, no other states are orthogonal to all of them. Thus, the ‘full’ state space is n×m-dimensional; not a surprise, since these states correspond naturally to the classical MB states.

But now we have a problem, for empirically, quantum particles do not obey MB statistics, as they should given indifference over such states. First, QM arose in response to the puzzle of modelling radiation from a perfectly non-reflective (‘black’) body. The solution requires that electromagnetic radiation be carried in discrete units, suggesting particles – ‘photons’ – with quantized energy levels (see Section 2.1.2). However, MB counting of the states of the photons does not quite work; instead, photons are ‘permutable’, meaning that

\[ \text{rearrangements of states among the particles are counted as the very same state.} \]

The new kind of counting that results is called ‘Bose-Einstein’ (BE). Second, electrons obey Fermi’s ‘exclusion principle’: no two can be in the same state. If they could, all the electrons in any atom would drop to the lowest energy level, losing their familiar chemical properties. We shall see a natural interpretation of exclusion if we again assume permutability; the result is ‘Fermi-Dirac’ (FD) state counting.

All known elementary particles are of these permutable kinds: ‘bosons’ or ‘fermions’, or collectively, ‘quanta’. Consider (*), if Ψ represents a state then, for instance, \( β_1 ⋄ α_2 ⋄ . . . ⋄ γ_n \) should represent a distinct one differing only in the states of the first and second particle, contrary to permutability. However, their superposition, \( α_1 ⋄ β_2 ⋄ . . . ⋄ γ_n + β_1 ⋄ α_2 ⋄ . . . ⋄ γ_n \), is also a vector, and clearly, if the first and second particles exchange states now, it is unchanged! To obtain a vector invariant under any permutation just requires performing the same trick with respect to all permutations: for any vector \( Φ \),

\[ \sum_{\text{permutations}} Φ \]

is unchanged or ‘symmetric’ under permutations. Thus, superposition allows a quantum representation of permutability as the superposition of all the non-permutable vectors corresponding to a single distribution – the subspace of symmetric vectors is suitable for representing bosons.

However, for instance, \( α_1 ⋄ α_2 ⋄ . . . ⋄ α_n \) is symmetric, but violates the exclusion principle, so symmetric states are inappropriate for fermions. The trick here is to remember that vectors in the same ray represent the same state, so permutability is also represented by ‘antisymmetric’ vectors, which change signs when particle pairs exchange states. But such states enforce exclusion: if two particles were in the same state, then permuting them could not change the vector at all. So antisymmetric states correspond to distributions satisfying the exclusion principle, and they represent fermions. (Antisymmetry is obtained by summing vectors related by even numbers of exchanges, and
subtracting those obtained by an odd number: for example, \( \chi = \psi_1 \diamond \phi_2 - \psi_1 \diamond \phi_2 \).

It is clear in this case that if \( \phi = \psi \) then \( \chi = 0 \): antisymmetry implies exclusion.

An (anti)symmetric ray is an ‘invariant representation’ of the permutations: a space closed under their action. Invariant representations that contain no smaller invariant representations are ‘irreducible’; since rays correspond to single states, they contain no subspaces of states – (anti)symmetric rays are irreducible. Abstractly, the (anti)symmetric subspaces are the spaces of irreducible representations of the given symmetry types. This abstract characterization generalizes: the full space contains irreducible representations with other symmetry types, and the question of whether these are realized by the states of fundamental particles, and if not, why not, has intrigued many. One point is that (anti)symmetric states are the only one-dimensional irreducible representations; in higher dimensional representations, permutations change the ray within the representation. But we have already accepted that it is not vectors that represent states, but rays; analogously, it is possible to take irreducible representations to represent states. The question becomes why particle states are never such ‘generalized rays’.

### 4.2.2 Issues

Fermi’s exclusion principle is highly suggestive of Leibnizian identity of indiscernibles (II). However, the permutability of quanta complicates the issue.

To pose the question physically, we consider quantum properties, those represented by Hermitian operators. For example, if there are two quanta, it seems that \( A_1 = A \diamond I \) represents the observable ‘property A on quantum 1’ and \( A_2 = I \diamond A \) the same but on quantum 2. Then we could prove a version of II if for any fermion state there was some A such that \( \langle A_1 \rangle \neq \langle A_2 \rangle \). However, it is straightforward to show that because such a state is a superposition of every permutation, there is never such an A. Similar results hold for relations; and for bosons; and for any number of quanta (and indeed to higher symmetries for pairs of particles invariant under exchanges). Thus, many commentators concluded that all quanta violate II, indeed in the strongest way – they are indiscernible in any state (French and Redhead 1988; Huggett 1999).

Inevitably, things are not so simple. First, Saunders (2006) argues that such results ignore ‘weak discernibility’: two individuals are weakly discernible if and only if there is some relation each bears to itself, but neither bears to the other (the characteristic features of ‘=’). Then, for instance, consider the anti-symmetric spin ‘singlet’ state, \( \uparrow_1 \downarrow_2 - \downarrow_1 \uparrow_2 \). Then the relation ‘is in the same spin state as’ weakly discerns; for instance, in each term of the superposition 1 has the same spin as itself and a different spin from 2. Muller and Saunders (2008) claim that this insight can be shown to make all fermions discernible, and indeed, Muller and Seevinck (2009) claim that all quanta in all states are discernible. However, while the basic point seems sound, its implementation
is still being debated. One issue is whether the operators, in terms of which the weakly discerning relation is defined, are truly observables; for one thing, for bosons, the operators given are not well defined on the symmetric subspace, since it is not invariant under their action. According to the standard interpretation, such operators are not observables at all; their measurement could collapse the system out of the space of possible states!

In fact, the use of such problematic observables is a threat to the coherence of philosophical work on II: for instance, $A_1$ and $A_2$ defined above have the same problem, and so do the operators invoked in the proofs that II fails. French and Redhead addressed this problem ‘transcendentally’, by claiming that any talk of II requires that quanta be individuals, that in some sense permuting them ‘could’ make a difference after all. They concluded that all vectors in the full state space actually do represent possible quanta states, but that quanta are dynamically restricted to (anti)symmetric vectors. Then, even operators that are only defined on the full space can represent properties, though unmeasureable ones. However, does II place such a requirement? I believe the answer is ‘no’, and that weak discernibility can be shown using observables defined on the (anti)symmetric states.

Alternatively, one could conclude that quanta are not individuals, properly speaking: the state-function labels do not individuate different things at all. This view has a natural expression in the ‘Fock representation’, in which a state is labelled merely by its distribution – the particle labels in the tensor product superpositions are abandoned. The appearance or not of particle labels is mere notation; whether they have metaphysical significance is the substantive question (Teller 1995, chapters 2–3). But note that even in this case, quanta come in discrete units, in the sense that there can be definite numbers of them in different definite states; that is what a distribution is. So quanta could, even on this view, be thought of as individuals in a weak sense: countable, and hence numerically diverse (Pooley 2006, pp. 109–14).

5. Conclusion

There is no road to properly understanding physics but through the technical formulations of its theories; the presentation here is at the simplest level at which a reliable outline can be given. However, it does offer both an overview of some of the fields studied, for those who wish to go further, and a sketch of the big ideas and problems, and their logic, for those needing an idea of what is at stake philosophically in the foundations of physics.

We have discussed three pillars of physics: spacetime, quantum mechanics and statistical physics. Arguably, a fourth pillar is field theory: theories of systems described by functions onto the points of spacetime, such as the
electric field, \( E(x_0,x) \) (or g-field). Pursuing this subject leads naturally to the ‘Lagrangian’ formulation of physical laws: a quantity known as the action takes its minimum value for allowed evolutions. That would allow one to see a profound connection between symmetries and conservation laws, and discuss non-spatial ‘gauge’ symmetries. We could also investigate quantum field theories: these theories generalize many particle QM, by allowing the number of quanta to be indeterminate (Teller 1995). And one could go even further, and explore what happens when one attempts to quantize the g-field of GR, and develop a quantum theory of gravity and spacetime.

But these, and many others, are, regrettably, all topics for other places.

Acknowledgements

I would like to thank John Van Dyke for considerable assistance with this essay.

References


1. Introduction

Philosophical questions about biology have been addressed by philosophers and scientists for centuries. Yet as a genuine discipline within philosophy, philosophy of biology started to emerge in the 1970s (Byron 2007). One motivation for this was the fact that much of traditional philosophy of science – growing out of logical positivism – focused on physics as the exemplar of science. Thereby, past philosophy of science simply assumed that accounts of confirmation, theory structure, laws and explanation would apply to biology as well, creating biased or inadequate views about the nature of science. But rather than directly addressing issues in general philosophy of science in the context of biology, philosophers of biology have, for the most part, engaged in questions that originate from within biology, pertaining to concepts from a specific biological field or to phenomena from a particular biological domain. Some of these questions have been raised by biologists, resulting in debate and fruitful interaction among scientists and philosophers. This interdisciplinary practice of contemporary philosophy of biology is illustrated by the International Society for the History, Philosophy, and the Social Studies of Biology, and by philosophical journals to which many scientists contribute. Most questions in philosophy of biology are epistemological questions concerning, for example, the character of particular biological explanations, models and concepts, and these are sometimes combined with issues about scientific method and practice. Yet these issues also tap into metaphysical considerations, at least insofar as they hinge on facts about the biological world.

Originally, most discussions in philosophy of biology centred on issues in evolutionary biology and systematics. For instance, the units of selection debate ponders which kind of entity is the fundamental player driving evolutionary change – the individual, the group of individuals, the gene or even the species. Arising out of disputes among evolutionary biologists, the epistemological question here is how genuine explanations in terms of natural selection actually work. The related metaphysical issue is at what level of biological organization evolutionary causation by selection takes place, and philosophers have offered several clarifications to these biological questions. Given
the many statistical models of evolutionary theory, philosophical interpretations of the nature and role of probabilistic factors (e.g. genetic drift) are put forward. The adaptationism debate concerns the circumstances in which phenotypic traits of species are best explained with exclusive reference to natural selection. In this context, attempts by sociobiologists and evolutionary psychologists at explaining the evolution of human behavioural and cognitive traits have been criticized by philosophers as falling short of the standards for acceptable evolutionary explanations. Philosophical studies of the meaning of ‘biological function’ discuss what the ascription of functions to biological traits involves (occasionally, but not always, a reference to a trait’s selection history), and whether functional and teleological explanations are distinct from causal explanations.

In addition to the traditional philosophical questions pertaining to evolution (which are still live issues), nowadays molecular and experimental biology are of increasing philosophical focus. This philosophical trend is largely due to the rise of genetics and molecular biology, with such biological advances prompting the philosophical debate as to whether classical genetics, and possibly other biological disciplines, can be reduced to molecular biology. It has also been argued that the notion of ‘genetic information’ is empirically misguided or of no explanatory relevance in biology. Apart from the role of genes and non-genetic factors in explanations of the development of organismal traits, developmental biology has recently become a biological field subject to philosophical interest, as a link to evolutionary issues has been created by the recent emergence of evolutionary developmental biology. With the scientific advance and societal importance of the life sciences and modern biomedical sciences, contemporary philosophers of biology also address issues in philosophy of medicine. Whereas views about the biological world have traditionally been based on higher animals (and plants), very recent scientific findings about microorganisms (largely driven by their now recognized medical relevance) may soon lead to challenges for philosophical views of biological individuality, social organization, the notion of species and the idea of a universal tree of life.

It is beyond the scope of any single article to provide a review of the major issues and positions in philosophy of biology (for overviews, see Hull 1974; Sober 1993; Griffiths and Sterelny 1999; Garvey 2007; Hull and Ruse 2007). Instead, the aim of this chapter is to lay out what implications biology has for some issues in general philosophy of science, including natural kinds, conceptual change, discovery and confirmation, explanation and reduction, and naturalism. Some of these offer additions and corrections to general metaphysical, semantic and epistemological views, and illustrate fruitful ways of conducting philosophical investigations that go beyond the practice of most of general philosophy (as the section on ‘Naturalism’ discusses).
2. Natural Kinds

A few decades ago, the notion of natural kinds gained prominence in philosophy, in particular in the context of the causal theory of reference (Putnam 1975) and rigid designation across possible worlds (Kripke 1980). Ever since, it has enjoyed widespread acceptance, particularly in metaphysics. However, while philosophers not acquainted with the philosophy of biology still take species and higher taxa (e.g. vertebrates) to be prime examples of natural kinds, in the 1970s the biologist Michael Ghiselin (1974) and the philosopher David Hull (1978) argued that species are not natural kinds. Instead, a particular species is an individual, a whole having the organisms of that species as its parts. This position has been extremely influential among biologists and philosophers of biology. Recently, following Boyd (1999), some philosophers have put forward a revised notion of natural kinds that is claimed to apply to species. As a result, in contemporary philosophy of biology there are diverging views about which of the various biological things are natural kinds, what notion of ‘natural kind’ is appropriate for biological kinds, and how biological kinds relate to kinds studied by other sciences.

Whereas in a nominal kind several objects are grouped together by mere human convention, the idea of a natural kind is that there are groupings of objects that conform to the objective structure of the world (Bird and Tobin 2009). The traditional notion of natural kinds was closely tied to kinds in physics and chemistry. It assumes that a kind is defined by a property called an ‘essence’, which has two functions. First, the essence determines the metaphysical identity of the kind. An object belongs to the kind if and only if the object possesses the essence. Second, the essence causally brings about the various properties characteristic of that kind, grounding the explanatory importance of natural kinds. For instance, the essence of oxygen is its atomic structure (including having eight protons and electrons). This atomic structure explains various physical and chemical properties of oxygen, in particular which compounds it can form and how it behaves in chemical reactions. Typically, essences have been taken to be intrinsic (internal) properties, as the example of chemical elements illustrates.

But it is easy to see that on this construal species cannot be natural kinds. For species are able to undergo unlimited evolutionary change. Whatever intrinsic property of an organism one chooses, be it a particular phenotypic feature or a genetic property, while many organisms existing at a point in time may share this internal property, future members of this species may not. Likewise, applying the traditional view of natural kinds to species does not do justice to biological variation. While acknowledging differences among species members, the so-called essentialism still assumes that there is a natural biological state of the species governing all its members (Sober 1980). The
deviation of individuals from this alleged natural state is seen as being due to unusual intervening factors (e.g. environmental influences), so that the natural state (species essence) is seen as real, and any variation is explained away as an accident. The empirically adequate picture, in contrast, is that variation across individuals is maintained and created by biological mechanisms and is a feature of major scientific importance.

Arguing that species are not natural kinds, Ghiselin (1974) and Hull (1978) suggested that every species is an individual. Just as an organism is composed of cells, a species is an individual made up of organisms. A species does not have members as a set, class or kind does; instead, a species is a whole that has various organisms as its mereological parts. The motivation for this position is that an individual is precisely the metaphysical thing that is able to undergo change. An individual exists in a certain period of time only, being present at specific spatial locations at any point in time. An individual can substantially change across time, and similarity among its mereological parts is not required of an individual, in contrast to the view that the members of natural kinds are identical in many properties. In analogy to this, a species comes into being at a specific point in time, inhabits concrete locations, undergoes evolutionary modification and goes extinct.

The view that species are individuals became the dominant view among philosophers of biology, and in particular biologists (Ereshefsky 2007). But recently, Boyd (1999) developed a revised notion of natural kinds that is designed to capture kinds as found in biology and the social sciences, and several philosophers and a handful of biologists have picked up on it (Wilson et al. 2007; Rieppel 2008). On this approach, a natural kind is a so-called homeostatic property cluster (HPC). First, the major assumption is that a natural kind is not defined by a simple property that all kind members share (an essence). Instead, a kind is characterized by a cluster of many properties that are more or less correlated, where a kind member need not possess all of these properties, and none of the properties is shared by all kind members. In fact, exhibiting a certain kind of variation can be characteristic for being a certain biological kind, and such a correlation and variation structure is in need of explanation (Wilson et al. 2007). The correlation of properties is not accidental, but due to some causal features (dubbed ‘homeostatic mechanism’). For this reason, the grouping of the members into a kind answers to features existing in nature, so it is a natural, rather than a nominal, kind. The statistical correlations or causal relations among the properties of the cluster make it possible for the kind to figure in scientific induction and explanation.

Second, the HPC construal of kinds explicitly maintains that properties of the cluster characterizing a natural kind need not be intrinsic, but may well be relational properties. For instance, a higher taxon (e.g. mammals) includes those species that are descended from a founding species, and having
a certain ancestry is an extrinsic property of a species. Several species can share this relational property of having the same ancestry, while differing in various (intrinsic) properties. Apart from common ancestry, the identity of a species may be defined by the ability of individuals to interbreed, which is also a relational property (of an individual) consistent with evolutionary change. As a result, historical entities such as species and higher taxa can be natural kinds, assuming that the identity of the latter is at least partially characterized by relational properties (Griffiths 1999).

These developments in philosophy of biology have several implications for theorizing about natural kinds in general. First, accounts which take species to be individuals have pointed to serious drawbacks of traditional, essentialist conceptions of natural kinds, and have offered a whole new way of thinking about biological entities. In addition, the HPC construal of kinds developed a revised and novel notion of natural kinds, motivated by a naturalism according to which philosophical notions should be tied to scientific practice, in this case kinds as they occur in biology. While the individuals vs. kinds debate focused on species and higher taxa, the HPC approach has broadened this scope by considering and attempting to capture many other putative biological kinds. In fact, Boyd developed the HPC account to likewise include kinds from the social sciences. Kinds are studied by different scientific disciplines, and there may well be different types of natural kinds, a diversity to which philosophical theories should be sensitive. It has also been suggested that some things (e.g. species) can, at the same time, be an individual and a natural kind (Dupré 1999; LaPorte 2004; Brigandt 2009), offering new metaphysical possibilities and pointing to problematic philosophical dichotomies.

Second, the HPC approach has highlighted the relevance of relational properties. Samir Okasha (2002) acknowledges that relational properties often individuate biological kinds (individuation being one function of traditional essences), but he claims that relational properties do not support biological explanations (in contrast to the second function of traditional essences). However, relational properties are, in fact, an important ingredient in explanations in biology and the social sciences (Brigandt 2009). Equations in ecology and economics explain dynamic change based on the existence of stable relations among some entities. More generally, while the causal capacities of biological entities (including molecular entities such as genes) are partially based on their internal structure, these causal capacities obtain only in certain biological contexts or given a suitable relation with other biological entities. Since many causal capacities are non-intrinsic properties, biological explanations have to appeal to relational properties as well. (Biologists arguing against some versions of reductionism have always emphasized the explanatory relevance of the structured contexts in which some entities occur; Brigandt and Love 2008).
Finally, while proponents of the view that species are individuals have defended it by metaphysical arguments (Hull 1978; Ereshefsky 2007), the HPC approach highlights epistemological considerations. For a natural kind is a grouping of objects that – despite the heterogeneity among the members of a biological kind – underwrites scientific generalization and explanation (thereby ‘accommodating’ scientific demands, as put by Boyd 1999). More generally, any kind of representation scheme used by scientists (be it grouping objects into kinds, be it considering objects as part of an enduring individual) is used to meet certain epistemic purposes, such as making particular inductive inferences or putting forward certain explanations. Philosophical accounts of the metaphysical nature of scientific entities have to pay attention to the epistemological reasons that make scientists refer to these (rather than other) entities and use them successfully in their theorizing by representing them in a certain way (Brigandt 2009; Love 2009). Different representations and classifications are used in different scientific contexts, resulting in the philosophically relevant fact that the totality of kind concepts and classification schemes used by biologists cross-classify nature (Dupré 1993).

3. Conceptual Change

Recent philosophical accounts of conceptual change have often been developed in response to the incommensurability challenge. Thomas Kuhn (1962) argued that the same term can be used with radically different (incommensurable) meanings in two paradigms (different theoretical frameworks), so that effective communication across paradigms and a rational choice of one such theory over the other is impossible. Paul Feyerabend (1970) claimed that the content of two theories containing incommensurable concepts cannot be compared. The standard response to the incommensurability challenge has been to focus on the reference of scientific terms (Putnam 1975; Fine 1975; Devitt 1979). For even if a term is used with different meanings in two theories, the term may nonetheless refer to the same entity across these theories, so that the theories can potentially make conflicting claims about this common referent and are thus comparable. The causal theory of reference provided an explanation of how stable reference is possible, despite substantially different theoretical views about the referent (Sankey 1994). According to this theory, a term’s reference is not fixed by (potentially false) descriptions of the referent, but a term is introduced by causal interaction with the referent and refers to that natural kind thusly picked out (so that the structure of the world plays a part in reference determination).

Some cases from the philosophy of biology offer a broader perspective for general discussions of conceptual change. With the advent of evolutionary
theory, biological views about the nature of species clearly changed. However, given that Darwin could communicate his arguments to his contemporaries, philosophers have questioned whether the pre-evolutionary and the Darwinian conceptions of species are incommensurable and whether they are even different concepts at all (Beatty 1986). Thus, the questions are: how these two conceptions of species are related, whether they are distinct concepts, and in either case how to make room for some conceptual continuity despite theoretical change. Moreover, nowadays different species definitions are used by biologists (Mayden 1997 lists 22 major definitions), and philosophers typically assume that many of them are, indeed, distinct concepts. One species concept considers those organisms that can interbreed with each other to be of the same species. Other species concepts focus on considerations of common ancestry and phylogenetic lineages, and still others use ecological criteria to delineate species. Not only do these concepts offer different criteria of what a species is, but they also lead to different classifications of individuals into species, so that these concepts differ in their extensions and reference. Most philosophers and biologists endorse pluralism about species concepts, that is, the idea that there is not one single true species concept (yet to be found or agreed upon), but that biology needs a multitude of different species concepts (Dupré 1999; Ereshefsky 1992; Kitcher 1984b; Wheeler and Meier 2000). The justification is that each species concept picks out some causal factors that influence the formation of partially distinct groups of organisms and that different species concepts are needed for different scientific purposes or in different biological disciplines. This situation raises several philosophical questions. How are these different definitions of species related? Arguably, the fact that the term ‘species’ is used in many different ways does not lead to substantial problems in communication among scientists (Kitcher 1984b), but a philosophical analysis of this issue is needed. Does the context of use determine which particular species concept is being employed (so that no ambiguity arises), or is there some form of ambiguity that does not matter much, as some of the species concepts overlap in their intention and extension?

A case that is currently subject to extensive philosophical debate is the gene concept. The advent of molecular genetics resulted in a semantic change, with the classical gene concept and the molecular gene concept typically being considered different concepts. A philosophical question is how the molecular concept could rationally grow out of the classical gene concept (Kitcher 1982), where it has to be taken into account that the molecular concept neither replaced nor reduced the classical concept, and both continue to be used in tandem (Weber 2005). Moreover, even reference changed in this historical period (Burian et al. 1996), as, among other things, both gene concepts individuate genes differently and thus differ in their extensions. Weber argues that geneticists have tracked not one, but several (overlapping yet distinct),
kinds by the term ‘gene’, and he uses the notion of ‘floating reference’ for the idea that the reference of the gene concept has changed constantly, though in a gradual fashion, suggesting that ‘this floating of the term’s reference seems not to have diminished its theoretical importance or practical usefulness’ (Weber 2005, p. 224).

In addition to conceptual change in the course of history, nowadays the molecular gene concept is used in quite different ways by different types of molecular biologists, resulting in the meaning and reference of the term ‘gene’ sometimes varying from context to context (Beurton et al. 2000; Brigandt 2010b; Falk 1986; Stotz and Griffiths 2004). This situation is due to findings in molecular genetics and, recently, in genomics. A gene is not a continuous stretch of DNA coding for one product, where all genes have the same structural features (which would permit a unique definition of what a gene is). Instead, it has turned out that many different kinds of (structurally characterized) DNA segments are involved in the production of RNA and polypeptides. One DNA segment can code for many products, and several separate DNA segments may be needed for a single product. In the latter case, some biologists consider these separate DNA segments as different genes, and others as one gene physically spread out within the genome. As a result, philosophers discuss how to study and characterize the current semantic variation surrounding the molecular gene concept. While Moss (2003) argues that there are two distinct gene concepts used in molecular biology, and Griffiths and Stotz (2007) opt for three basic gene concepts, Waters (2000) controversially maintains that one shared molecular gene concept underlies all varying uses. Since apart from ‘gene’, nowadays further terms are used to refer to DNA elements involved in the coding for gene products (e.g. ‘transcription unit’, ‘exon’), another possible, though very unconventional, option is to suggest that in the future the term ‘gene’ will be or should be replaced by a plurality of other terms (Kitcher 1992).

These considerations from recent philosophy of biology have implications for general philosophical theories of conceptual change. Many traditional accounts have focused on reference to rebut the incommensurability challenge. However, even if a theory of reference ensures that a scientific term has the same referent in two theories (so that they can make conflicting claims), given that the term may be used with a different meaning in each of these theories, the notion of reference alone does not solve Kuhn’s original challenge – how communication across different approaches and rational theory choice are possible despite variation in meaning. Several accounts of conceptual change have already acknowledged that the historical change in the meaning of terms is to be philosophically understood (e.g. Kitcher 1982; Nersessian 1984). Cases from biology, like the change of the gene concept, show that scientists confidently discard previous definitions they used to put
forward, and philosophical accounts have to explain why it can be rational for scientists to modify a term’s definition. And even the very reference of a biological term can be subject to change, so that a philosophical approach that assumes stability of reference cannot account for this aspect of conceptual change.

Furthermore, in addition to diachronic change, terms such as ‘species’ and ‘gene’ show that, even at one point in history, both the meaning and the reference of a scientific term can vary across uses. (Other possible such cases are the homology concept (Brigandt 2003) and the notion of a stem cell (Shostak 2006).) Semantic variation raises several questions to be addressed (Brigandt 2010b). What are the reasons for variation in a particular term’s usage? How are different uses of a term related, and what structure does the semantic variation have – do different biological sub-disciplines or approaches each favour a different use, can one and the same scientist use a term differently in different research contexts? To which extent does semantic variation hamper communication? While in his early work Kuhn assumed that differences in meaning are intrinsically problematic, some instances of semantic variation may, in fact, be beneficial to scientific practice, by promoting a division of semantic labour, for example (Wilson 2006). To the extent that semantic variation is compatible with successful communication, why is this possible, and when and how does semantic variation support scientific practice?

4. Discovery and Confirmation

Accounts of confirmation in general philosophy of science typically attempt to advance a theory that is as universal as possible and covers all scientific fields and empirical domains.1 Assuming that confirmation in science is inductive, such a logic of induction describes the form of the confirmation relation between evidence statements and theory, abstracting away from the particular empirical content involved in a particular instance of confirmation – just like deductive logic determines whether premises entail the conclusion in terms of the form, but not of the content, of the statements involved. However, Norton (2003) has pointed out that the search for formal schemas of inductive inference has proven largely futile. Reviewing different kinds of inductive schemes (such as inductive generalization, hypothetico-deductive accounts including error statistics, abduction and Bayesianism) and their known flaws, he concludes that there is a strong trade-off between generality and inductive strength, and that schemas of induction that are of universal scope are either unreliable (the conclusion is not supported by premises in too many instances) or circular (and thus useless for the purposes of inductive as ampliative
His diagnosis is that we have been misled by deductive logic into thinking that universal schemas of inductive inference exist. In contrast, Norton proposes what he calls a ‘material theory’ of induction, according to which induction is grounded not in universal schemas, but rather in empirical matters of fact. To illustrate this with a simple example (of mine), formal accounts of analogical reasoning as a type of induction construe an inference from an object $a$ having property $P$ to an object $b$ having property $P$ as justified in case the objects $a$ and $b$ are similar in that they share properties $Q_1, Q_2, \ldots$ Such a formal account has to acknowledge that the inductive inference $Pa \vdash Pb$ is justified only insofar as the degree of similarity between the objects $a$ and $b$ is significant and the properties $Q_i$ are relevant for the property $P$ to be projected (Salmon 2002). However, what is relevant or significant crucially depends on features of the particular case, and thus the plausibility of the inference is essentially contingent on empirical information (matters of fact), while the inference’s logical form is actually quite insignificant for its justification. On Norton’s account, new empirical knowledge generates new inferential power (by providing additional matters of fact), but not by yielding novel abstract schemes of inference. As matters of fact relevant for an induction hold only in certain scientific domains, scientific induction is local. Individual instances of induction may be too domain-specific to be categorized together with other inductions under a general type of induction.

While Norton’s proposal is motivated by unsuccessful attempts to put forward formal schemas of induction in general philosophy of science, recent philosophy of biology provides more direct support for the idea that confirmation in science is not based on universal schemes, but domain-restricted and contingent upon specific empirical content. For instance, Weber (2005, chapters 4–5) offers a detailed discussion of confirmation in experimental biology. He scrutinizes the oxidative phosphorylation controversy, a debate in biochemistry that started in 1961 with two rival accounts, but could not be settled until 1977. Weber rejects Bayesian confirmation theory, arguing that in this scientific case it would have made erroneous normative suggestions about theory acceptance. (On Bayesian analysis, the true biochemical theory should have been accepted too early – in 1966, at a point where the total evidence did not favour one hypothesis over the other.) Mayo’s (1996) error-statistical theory aligns with experimental biology better than Bayesianism, as her approach does not assume that scientific inference solely consists of a confirmation relation between theory and evidence, and it instead captures the piecemeal production of evidence and scientists’ attempts to control error. However, Weber argues that a statistical notion of error cannot apply to experimental biology, as the relevant reference class for an experiment is unclear, so that no error frequencies can be assigned. Based on the practice of experimental biology, he concludes that epistemic norms used by biologists are not
universal rules (as in Bayesianism and error statistics), but domain-specific, empirical considerations. In the context of evolutionary biology, Sober (2008) argues for pluralism about scientific inference, where different inference methods are appropriate in different situations. He emphasizes that scientific inference methods are not a priori and valid for every case, but presuppose specific empirical assumptions and, thus, are legitimate only in a certain range of empirical cases.

A similar picture applies to discovery in biology. Originally, philosophy of science – growing out of logical positivism – did not view scientific discovery as a philosophical issue at all. Back then, the distinction between the context of discovery and the context of justification was so construed that while scientists may come up with hypotheses by any possible means (including non-rational psychological processes), the subsequent justification of a hypothesis always has to follow standards of rationality. Consequently, only confirmation was viewed as a subject for philosophy, while the process of discovery may be studied by psychology or sociology. Due to more recent developments in philosophy of science, such as the New Experimentalism movement (Hacking 1983) and naturalism, discovery in experimental and theoretical science is nowadays viewed as a rational process – in the sense that scientists use reasoning strategies and discovery heuristics that are reliable at yielding plausible hypotheses, or at least more likely to yield fruitful hypotheses than other strategies (Nickles 1980; Wimsatt 2007). Thus, discovery is in need of philosophical analysis.

However, given the traditional philosophical ideal according to which rationality consists in formal logic and the use of universal principles, early accounts of discovery in biology have followed this philosophical model. For instance, Schaffner (1974) initially argued that discovery (just like justification) is reasoning using deductive logic, though he himself came to abandon this idea. In his reanalysis of the biological case addressed by Schaffner (the lac operon model of gene regulation), Weber (2005, chapter 3) argues that this instance of discovery involved a variety of analogical reasoning, in a form that is prohibited in the context of justification. Another prominent account of discovery in biology was put forward by Darden (1991) in the context of Mendelian genetics. Her account is influenced by artificial intelligence modeling, and she lays out a set of general strategies of discovery. She acknowledges that it is historically impossible to tell whether these strategies have actually been used, but argues that they could have been used to arrive at the hypotheses put forward in the history of Mendelian genetics. However, Weber (2005) rightly objects that this does not settle one important philosophical issue – whether general and domain-unspecific rules actually are or have been used by scientists in instances of discovery. Based on his study of experimental biology, Weber is sceptical about this, and prefers the interpretation that
reasoning in discovery involves strategies that are limited to certain domains and cannot be universal, as they are based on specific empirical considerations.

In sum, many accounts of discovery and confirmation in general philosophy of science have focused on formal rules of inference, which, by abstracting away from particular empirical content, apply to any scientific field and domain. But ‘to show that a kind of reasoning can be rational . . . is not the same as showing that it employs general rules or procedures’ (Weber 2005, p. 86). A look at biological practice suggests that scientists’ reasoning in discovery and confirmation is contingent upon empirical content specific to a particular case and thereby conforms, at best, to local and domain-specific principles. This is not to say that a philosophical study of confirmation or discovery is impossible. While overarching philosophical theories may be of limited use, more insightful studies reveal the particular scientific principles and considerations used in certain cases, highlighting the concrete empirical content (in addition to formal structure) that justifies the effectiveness and rationality of this kind of reasoning in the respective empirical domain. This aligns, to a large extent, with the practice of contemporary philosophy of biology. Philosophers of biology rarely address ‘confirmation’ and ‘discovery’ in a direct fashion – at least they often do not phrase their studies in these terms from general philosophy of science. But their many case studies and discussions pertaining to biological issues shed light on scientific method and reasoning as used in different parts of biology.

5. Explanation and Reduction

Just as the precise nature of discovery and confirmation in biology can vary with the empirical domain, biological explanations can take different shapes in different cases, and what counts as an adequate explanation depends on empirical considerations tied to a particular case. Different accounts of explanation have been put forward in general philosophy of science. The deductive-nomological model assumes that an explanation is a logical deduction from law statements, which presupposes that a certain scientific discipline contains laws in the first place (Hempel and Oppenheim 1948). Statistical relevance models argue that explanations are not logical arguments at all (neither deductive nor inductive), and take, instead, statistical relevance relations among quantitative variables to be explanatory (Salmon 1971). Causal-mechanistic models contend that explanations exhibit causal structures, as they occur in physical processes (Salmon 1984) or biological mechanisms (Machamer et al. 2000). While sometimes these accounts have been seen as rival approaches, at least from the point of view of many philosophers of
biology, there is no need to choose one of them as the correct theory of explanation. For different types of explanations occur in different parts of biology, and can even be used in one discipline. For instance, explanations in evolutionary biology often combine quantitative models (in line with deductive-nomological, unification or possibly statistical relevance accounts) with their application to concrete populations that exhibit the material features accounting for fitness differences, for example (in line with causal-mechanistic models). Furthermore, following the logical positivist ideal, early accounts attempted to capture what a good explanation is in purely formal terms, so that, for instance, the deductive-nomological approach had to offer a syntactic account of statements that capture laws of nature (attempts to do so failed). In contrast, what makes a particular explanation in biology scientifically acceptable is likely to depend on empirical considerations that cannot be characterized in a formal-syntactic fashion.

A related philosophical issue is reduction. In the philosophy of mind, accounts of reduction have centred on ontological issues (emergence, supervenience, how to construe physicalism). In contrast, while taking some notion of ontological reduction for granted (e.g. that every particular biological system is constituted by nothing but molecules and their interactions), the reductionism debate in philosophy of biology has turned on epistemological and methodological issues. In a nutshell, epistemic reduction is the idea that knowledge about one scientific domain (typically about higher level processes) can be reduced to another body of scientific knowledge (typically concerning a lower level). This broad idea can be spelled out in different (stronger and weaker) ways, so that the reductionism debate concerns not only the question of whether reduction is possible, but which notion of reduction is the appropriate one for biology – leading to a diversity of philosophical accounts of reduction.

One basic notion of reduction is theory reduction. Originally developed by Ernest Nagel (1961) as a general framework for science, it is (roughly speaking) the idea that one theory (construed as a formal system) can be logically deduced from a more fundamental theory. Inspired by the rise of molecular biology, Schaffner (1976) applied it to biology by claiming that the theory of classical genetics can deduced from the theory of biochemistry. However, the notion of theory reduction was immediately criticized (Hull 1976; Kitcher 1984a). The proponents of theory reduction acknowledged that a reduction of classical genetics to molecular biology has not been achieved yet (and that theory reduction is not an aim of scientists), while arguing that theory reduction is in principle possible. Yet the critics wondered why such a notion of reduction is relevant for understanding biological research in practice, including reductionistic methods and explanations (Wimsatt 2007). As a result, the notion of theory reduction was largely abandoned in favour of another class of models, which may be called accounts of explanatory reduction (Brigandt
There are different such models, but they differ from the earlier approach (which assumes that whole theories are reduced) by assuming that generalizations of varying scope, mechanisms or even individual facts are the features reductively explained. A reductive explanation need not be a logical deduction from a theory containing laws; and while it proceeds by explaining a complex whole in terms of its (lower-level) parts, the *explanans* need not exclusively contain terms referring to lower-level entities (in contrast to the logical premises in theory reduction consisting of statements of the reducing theory only). Due to these differences, models of explanatory reduction are much better able to capture the piecemeal nature of research progress and explanation in molecular biology. Furthermore, an individual philosopher may be inclined to focus on one notion of reduction and use it as an overarching model of ‘reduction in biology’. However, when biologists discuss the possibility or impossibility of reductions, they typically mean very specific explanatory or methodological issues about their particular empirical cases, so that ‘reduction’ as used across different research contexts may express different scientifically relevant claims.

Many arguments against epistemic reduction fall into one of two categories (Brigandt and Love 2008). One type of consideration is that the effects of a biological system, especially of lower-level molecular processes, depend on the overall biological context in which they occur. One molecular kind can correspond to *many* higher level kinds, if this molecular kind is part of different overall systems. As a result, reductions must usually be able to take into account the context of and the relations among molecular entities and processes. The second argument against epistemic reduction proceeds from the fact that *many* molecular kinds correspond to a single, higher level kind, due to the fact that a higher level biological structure or process can be instantiated by different kinds of molecular structures or processes (in line with multiple realization arguments familiar from the philosophy of mind). As a result, explanations in terms of such higher level kinds can be more general (encompassing several different lower level accounts) or point to causally more robust features (picking out entities that are stable across time while their underlying molecular constitution changes). These arguments are effective against accounts of theory reduction, and possibly also against some models of explanatory reduction. In addition to developing alternative notions of reduction, the rejection of theory reduction has led many philosophers of biology to endorse pluralism, the idea that biology needs a diversity of methods, models and modes of theoretical reasoning (Mitchell 2003). Sometimes this has been phrased in a bold anti-reductionist stance, as an argument for the disunity of biology (Dupré 1993; Rosenberg 1994).

Theory reduction, as a model that assumes the reduction of various biological fields to a single theory, does not do justice to the diversity of
different biological approaches. At the same time, the mere endorsement of pluralism or disunity of science fails to recognize the intellectual relations that exist across different sub-disciplines of biology. For this reason, arguably the most promising philosophical approaches in the last few decades concern the coordination and integration of different concepts, models and explanations. For instance, Darden and Maull (1977) developed a model (of ‘non-reductive unification’, as they put it) that argues that integration proceeds by two fields coming to be linked by so-called ‘interfield theories’. For example, the chromosome theory of inheritance is an interfield theory that in the early twentieth century came to connect Mendelian genetics (the study of phenotypic inheritance across generations) and cytology (the study of the material contents of cells). This was a non-reductive integration because neither did genetics reduce cytology, nor did cytology reduce genetics. In contrast to the possibility of unification in virtue of several theories reducing to a new, more fundamental theory, an interfield theory connects theories without reducing them.

Philosophical studies of integration in biology have often been conducted in the context of molecular, cellular and experimental biology, with related attention to disciplinary change and institutional factors (Bechtel 1986). Accounts of biological mechanisms in these biological domains have shown that mechanistic explanations often appeal to entities on several levels of biological organization at the same time (Craver 2005; Darden 2005). In a similar vein, the units of selection debate in evolutionary biology has led to some agreement that units of selection can be found at different levels, so that there is no single level that would a priori be the most fundamental one in evolutionary explanations, but in each case it is an empirical question at which level(s) selection takes place (Brandon 1988; Okasha 2007). Studies of the emerging integration of evolutionary biology and developmental biology suggest the epistemological importance of problems addressed by biologists (Love 2008). A complex scientific problem such as the explanation of the evolutionary origin of novel structures necessitates the integration of different ideas and approaches. Apart from motivating integration, a problem also structures integration by determining which disciplines, methods, concepts and models are to be integrated. Integration (or unification) is not a regulative ideal or aim in itself for biologists, but may be needed to solve a particular biological problem (Brigandt 2010a).

6. Naturalism

Naturalism is a characteristic feature of contemporary philosophy of science. Yet ‘naturalism’ as used by different philosophers can denote different ideas (Papineau 2009). It can express certain metaphysical tenets, such as the
commitment to a purely physicalist ontology, or the idea that humans (including their intellectual and moral capacities) are part of nature. A *methodological* variety of naturalism is particularly germane to the practice of philosophy of biology (Wimsatt 2007). Here the idea is, first, that given philosophy of science’s task to understand science and its success, philosophical accounts have to do justice to and reflect actual science. Second, the practice of philosophy of science is continuous with scientific practice, and empirical methods of science can be useful tools for philosophers in their attempts to form philosophical notions. This was illustrated by several themes from this chapter’s earlier sections. Given the plurality of kinds of explanation used in biology and the fact that ‘reduction’ as used by biologists expresses different commitments in different contexts, naturalistic philosophy has to capture this diversity. If biologists’ reasoning involved in discovery and confirmation is dependent on empirical considerations and the particular shape of discovery and confirmation (and what makes it justified) can vary across biological domains, philosophical accounts of discovery and confirmation must accommodate this. In the case of conceptual change, sometimes the notion of reference and the existence of stable reference across theoretical change have been considered to form a satisfactory reply to the incommensurability challenge. While some discussions in general philosophy of science have taken place under these assumptions and within these limits, I have argued that the notion of reference alone does not explain how rational theory choice is possible in the face of meaning differences, and that the reference of biological terms can be far from stable, prompting the need to philosophically address this.

In philosophy as a whole, concepts are often put forward and debated through the *use of intuitions*. Various possible scenarios are considered, and it is determined whether the concept applies to a scenario (according to one’s intuitions). However, the practice of philosophers of biology suggests that a more fruitful approach is to view philosophical concepts as technical terms that – just like scientific terms – are introduced for certain theoretical tasks. Any philosophical account is to be defended in terms of its fruitfulness and adequacy of performing this task (an a posteriori question), rather than conforming to a priori intuitions. For instance, according to widespread intuitions, the identity of a natural kind is determined by an internal essence. Yet an actual justification for the traditional notion of natural kind is that it captures features of physical and chemical kinds and how they figure in scientific explanation (which can, in principle, be empirically contested). A novel notion of natural kind, such as the notion of homeostatic property cluster, is needed to the extent that it is necessary to deal with biological kinds. Above, I indicated that different notions and models of epistemic reduction have been put forward. It is not sufficient to simply motivate one definition of ‘reduction’ and then argue that classical genetics can (or cannot) be reduced to molecular
biology on this definition, but rather the philosophical relevance of this notion of reduction has to be defended, which also involves reference to scientific practice. Indeed, the notion of theory reduction (defended by the mere in-principle possibility of theory reduction) was challenged for failing to shed light on reductionistic research in practice.

In philosophy of mind and several other fields, providing a naturalistic account is often understood as giving a *reductive definition* of a philosophical concept in terms of an alleged ‘scientific’ vocabulary. For instance, the aim is to offer a naturalistic reduction of semantic and intentional notions (e.g. ‘reference’, ‘meaning’, ‘mental representation’), in particular to reduce the normativity associated with correct vs. incorrect representation. (Some such attempts are causal theories of reference, teleosemantic accounts of representation, or Fodor’s asymmetric dependency account of representation.) However, such a naturalistic vocabulary is somewhat of a philosopher’s fiction. For as we saw in the section on ‘Explanation and Reduction’, scientists use concepts referring to entities on many levels of organization and do not attempt to reductively define their concepts in terms of the vocabulary of a particular field, such as molecular biology. Biologists provide reductive explanations only when it is empirically beneficial to do so. As a result, a methodologically naturalistic approach in philosophy of science – attempting to do philosophy in analogy to scientific practice – is not committed to settle a priori on a specific vocabulary and attempt to reduce philosophical concepts to it.4

When philosophers talk about ‘explanation’ (e.g. ‘explaining reference’), they often mean a definition of a philosophical term by means of necessary and sufficient conditions. Explanation in biology (e.g. ‘explaining life’) involves something different. Biologists do not aim at giving a definition of life. Instead, the aim is to gain partial, but ever increasing, insights into the causal workings of various life processes. Likewise, naturalistic philosophy of science is not so much concerned with defining ‘knows that p’ in terms of necessary and sufficient conditions, but to understand the various aspects of knowledge production. This includes empirical issues such as the cognitive factors involved in belief formation. Various social aspects of the organization of scientific research are a major reason for the success of modern science, so that many philosophers of biology pay attention to the sociology of science and history of science (Hull 1988; Downes 1993). This use of social and other empirical factors is not inconsistent with philosophy of science dealing with normative concerns such as the rationality of discovery and the justified endorsement of theories, as certain social factors can be precisely those that underlie the rational and effective generation of knowledge (Longino 2002). Giere (1988) assumes that naturalistic philosophy of science consists in an empirical-descriptive approach, but cannot be concerned with normative considerations. However, biologists freely use normative notions, for instance, when talking about proper method, adequate
representations, justified hypotheses, relevant problems, and the proper intellectual aims of their discipline. In a similar vein, naturalistic philosophers may arguably engage in normative considerations in their subject matter, the study of science (Laudan 1990). Just like biologists fruitfully study the causal interactions of entities at several levels of organization, the aim for philosophers of science should be to shed light on the relations among phenomena referred to by philosophical (including normative) concepts and the phenomena studied by the cognitive, natural and social sciences. Rather than being reduced or eliminated, normativity is thereby one out of many factors. Due to its methodological construal of naturalism and its interaction with biologists, historians and sociologists, philosophy of biology has become an interdisciplinary approach.

Acknowledgements

I am indebted to Marc Ereshefsky and Alan Love for comments on a draft of this essay. The work on this essay was funded with Standard Research Grant 410–2008-0400 by the Social Sciences and Humanities Research Council of Canada.

Notes

1 See Hawthorne (this volume) on Bayesian confirmation and Forster (this volume) more generally on evidence.
2 See De Regt (this volume) on philosophical accounts of explanation.
3 See Walter and Eronen (this volume) on reduction.
4 To mention another way of using biological practice as a guideline for philosophical method, Burian et al. (1996) argue that there is probably no unique philosophical account of meaning and reference. Just like biological concepts are often used in varying ways – with there being several gene concepts – philosophical concepts require local analyses and are best employed in different versions for different purposes.

References


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


Towards a Mechanistic Philosophy of Neuroscience

Carl F. Craver and David M. Kaplan

1. Introducing Neuroscience

Neuroscience is an interdisciplinary research community united by the goal of understanding, predicting and controlling the functions and malfunctions of the central nervous system (CNS). The philosophy of neuroscience is the subfield of the philosophy of science concerned with the goals and standards of neuroscience, its central explanatory concepts, and its experimental and inferential methods. Neuroscience is especially interesting to philosophers of science for at least three reasons. First, neuroscience is immature in comparison to physics, chemistry and much of biology and medicine. It has no unifying theoretical framework or common vocabulary for its myriad subfields. Many of its basic concepts, techniques and exemplars of success are under revision simultaneously. Neuroscience thus exemplifies a form of scientific progress in the absence of an overarching paradigm (Kuhn 1970). Second, neuroscience is a physiological science. Philosophers of biology have tended to neglect physiology (though see Schaffner 1993; Wimsatt 2007). Physiological sciences study the parts of organisms, how they are organized together into systems, how they work and how they break. Its generalities are not universal in scope. Its theories intermingle concepts from several levels. Neuroscience thus offers an opportunity to reflect on the structure of physiological science more generally. Finally, unlike other physiological sciences, neuroscientists face the challenges of relating mind to brain. The question arises whether the explanatory resources of physiological science can be extended into the domains involving consciousness, rationality and agency, or whether such phenomena call out for distinctive explanatory resources.

The philosophy of neuroscience should also interest neuroscientists. Because neuroscience is so young, and because it faces unique explanatory
demands, it is useful to reflect on what the science hopes to accomplish and how it might most efficiently achieve its objectives. Philosophers have much to contribute to that discussion.

Below, we sketch five debates in the philosophy of neuroscience. In Section 2, we discuss the debate between predictivists and mechanists about the norms of explanation. In section 3, we apply lessons drawn from this debate to the question of whether dynamical systems explanations are legitimate, non-mechanistic types of explanation. In Section 4, we discuss the debate between mechanistic realists and mechanistic pragmatists about how to build a ‘periodic table’ of the mind-brain. In Section 5, we discuss the strengths and limitations of deficit studies as an instance of experimental strategies for discovering mechanisms. In Section 6, we discuss the thesis that the brain is a computer, considering debates about how to define computation and about whether computational explanations are mechanistic explanations. In Section 7, we discuss the relative merits of reductive and integrative visions of neuroscience. A consistent theme throughout is that the philosophy of neuroscience can make considerable progress if it starts with the central idea that neuroscience is fundamentally a mechanistic science.

2. Mechanistic Explanation

Neuroscientists are united by the collective project of explaining how the nervous system and its parts work. Yet there is no consensus about what counts as a successful explanation. Consider two clearly contrasting views. Predictivists believe that any good predictive model is explanatory. Mechanists believe that, in addition, explanation requires knowledge of a mechanism.

Predictivists assert, for example, that to explain a phenomenon is to show that its occurrence follows from universal or statistical regularities, together with the relevant initial conditions. This view of explanation once dominated the philosophy of science (Hempel 1965). On a very liberal interpretation of predictivism, any mathematical or computational model that predicts all the relevant features of the phenomenon in a wide range of conditions counts as an explanation.

Mechanists, in contrast, insist that explanatory models describe the relevant causes and mechanisms in the system under study. In their view, a predictively adequate mathematical model might fail to explain how the system behaves. To explain a phenomenon, one has to know the mechanism that produces it, one has to know what its components are, what they do and how they are organized together (spatially, temporally and hierarchically) such that they give rise to the phenomenon to be explained (see Bechtel and Richardson 1993; Craver 2007; Machamer et al. 2000).
Predictivists and mechanists have argued over how to understand a central exemplar of modern neuroscience: the Hodgkin-Huxley model of the action potential. Action potentials are fleeting changes in the voltage difference across the neuronal membrane that serve as a basic unit of neural signalling. Hodgkin and Huxley hoped to discover how neurons generate action potentials, but their techniques could not discern the molecular mechanisms we now know to be involved. Neuroscientists have since discovered that action potentials are produced by voltage-sensitive ion channels that control the diffusion of sodium and potassium ions across the membrane. To understand how the debate between predictivists and mechanists plays out in this example, some historical context is needed.

Hodgkin and Huxley initially focused on a more tractable problem: to characterize changes of membrane conductance for sodium and potassium as a function of membrane voltage. They used a voltage clamp to hold the membrane voltage at particular values by balancing the resulting flow of current through the membrane. They could then calculate the membrane conductance at a particular voltage from the current required to balance the flow of current through the membrane at that voltage. Their primary modelling achievement was to generate equations that described the voltage-conductance relationships discovered with the voltage clamp. Using basic concepts from electrical theory, Hodgkin and Huxley represented the membrane as an electrical circuit, with currents carried by ions along parallel paths, each with its own battery (driving force) and variable resistor (the conductances). The electrical circuit schema could then be described mathematically, providing a formal model of the membrane that predicts much of the electrical behaviour of neuronal membranes. This mathematical model remains in wide use.

The core of the model is the total current equation (Eq. 1), expressed as a set of coupled partial differential equations describing the voltage-dependent changes in ionic conductances across the neuronal membrane:

\[
I = C_M \frac{dV}{dt} + G_K n^4 (V - V_K) + G_{Na} m^3 h (V - V_{Na}) + G_l (V - V_l)
\]

(1)

where \(I\) is the total current crossing the membrane, composed of a capacitive current \(C_M \frac{dV}{dt}\), a potassium current \(G_K n^4(V - V_K)\), a sodium current \(G_{Na} m^3 h (V - V_{Na})\), and the leakage current \(G_l(V - V_l)\), a sum of smaller currents for other ions. \(G_K, G_{Na}\) and \(G_l\) are the maximum conductances to the different ions. \(V\) is displacement of \(V_m\) from \(V_{rest}\). \(V_K, V_{Na}\) and \(V_l\) are the differences between equilibrium potentials for the various ions (where diffusion and driving force balance, and no net current flows) and \(V_m\). Crucially, the model also includes the three coefficients, \(n, m\) and \(h\), whose values determine
Towards a Mechanistic Philosophy of Neuroscience

how conductance varies with voltage and time (see Hodgkin and Huxley 1952).

Mechanists and predictivists debate whether aspects of the Hodgkin-Huxley model genuinely explain, or merely predict, how the membrane behaves at different voltages (Bogen 2005, 2008; Craver 2006, 2008; Machamer et al. 2000; Schaffner 2008; Weber 2005, 2008). The debate focuses on the rate coefficients n, m and h. In building the model, Hodgkin and Huxley considered a hypothetical mechanism involving the movement of activation and inactivation particles in the membrane. The coefficient n is raised to the fourth power in the rate equation for potassium conductance, one might think, because the maximum conductance depends on the probability that each of four activation particles has moved as it must. A similar interpretation is available for the term m3h in the equation for sodium conductance. It is important to remember, however, that Hodgkin and Huxley had only vague and speculative ideas about how membranes change conductance. The idea of ion-selective channels did not appear until the 1960s, and neuroscientists debated their existence well into the 1980s (see Hille 1992). This is why Hodgkin and Huxley insist that, ‘the success of the equations is no evidence in favour of the mechanism of permeability change that we tentatively had in mind when formulating them.’ They explain: ‘An equally satisfactory description of the voltage clamp data could no doubt have been achieved with equations of very different form, which would probably have been equally successful in predicting the electrical behaviour of the membrane.’ (1952, p. 541) That is, they did not know the mechanism.

Hodgkin and Huxley were mechanists. Mechanists can agree with predictivists that explanatory models should make accurate predictions. They insist, however, on the distinction between explanatory models and phenomenal models. Phenomenal models characterize a phenomenon without explaining it, or even pretending to explain it. For example, Snell’s law describes how light refracts as it passes from one medium to another, but the law does not explain why the path of light changes as it does. Theoretical neuroscientists Dayan and Abbott take this mechanistic stance when they distinguish purely descriptive models that ‘summarize data compactly’ from mechanistic models that ‘address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry’ (2001, p. xiii).

How might predictively adequate models fail as explanations? One might build a model to predict a cause on the basis of its effects. A car’s heat gauge affords accurate predictions of the engine’s temperature, but it does not explain it, presumably because the heat gauge indicates, but does not cause, the engine’s changes in temperature. To continue with the example, the heat gauge reading can predict radiator blowouts, but does not explain them when they
occur. In this case, the correlation between them is explained by a common cause (engine heat). For these reasons (and others like them) mechanists hold that explanatory models reveal causal mechanisms. Hodgkin and Huxley insist that their model merely predicts how the membrane changes conductance with voltage, because they know it does not describe the mechanism for those changes. Finally, this example illustrates the mechanist’s distinction between how-possibly and how-actually explanations (Dray 1957; Hempel 1965). Hodgkin and Huxley’s activation and inactivation particles are how-possibly fictions, used as heuristic tools to interpret their mathematical model. It does not count against the Hodgkin-Huxley model that the how-actually model has turned out to be quite a bit more complicated (see, for example, Doyle et al. 1998; Yu and Catterall 2003). That part of the model is intended for prediction, not explanation.

Summarizing these considerations, mechanists hold explanatory models to a strict model-to-mechanism-mapping (3M) requirement:

(3M) A model of a target phenomenon explains that phenomenon when (a) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism.

Why do mechanists think these mapping relations mark a crucial difference between explanations and non-explanations? Two reasons are primary. First, mechanists think that explanation and know-how are deeply connected to one another. The mechanist reserves the term ‘explanation’ for descriptions that reveal how things work precisely because knowing how things work is a reliable route to identifying ways of making systems work for us. Second, as illustrated in the above examples, predictivism violates scientific common sense about what does and does not count as an adequate explanation (see, for example, Salmon 1984). Predictions can be made on the basis of correlations, and only some correlations are explanatory. Perhaps predictivism can be restricted to block these devastating counter-examples, but the efforts to do so will only bring it closer into line with the mechanist point of view. One might require that explanations appeal to laws of nature, or that they unify diverse phenomena within a single explanatory framework. Mechanists think that such demands are ill-fitted to the explanatory domains of physiological sciences (Bechtel and Abrahamsen 2005), and they doubt that such corrections can be made to predictivism without collapsing into a version of mechanism.
3. Need for Non-mechanistic Explanation?

Some dynamicists argue that complex systems require a non-mechanistic form of explanation. Mechanistic explanations suffice for simple systems, in which parts work sequentially and can be represented in a simple flow chart. But few target systems have such a unidirectional, sequential architecture. Neural systems often have tens to millions of parts interacting promiscuously with one another, often through reciprocal feedback connections. They exhibit non-linear changes in behaviour; minor changes in a single part or in the operating conditions have dramatic effects on the behaviour of the whole. As the number of parts and the diversity and interdependence of their interactions increases, it becomes harder to see the system as made up of discrete parts. Contrary to the demands of 3M, dynamicists argue, the phenomenon must be understood on its own level, in terms of system variables and order parameters (see Beckermann et al. 1992; Stephan 2006). For such systems, ‘the primary explanatory tools’ are the mathematical methods of non-linear dynamical systems theory (Chemero and Silberstein 2007).

Some mechanists respond that complex systems are, in fact, decomposable. Bechtel (1998) shows that fermentation and glycolysis, which have numerous feedback loops and non-linearities, can be broken down into subsystems and can be localized to specific parts of the cell. Action potentials are produced by complex feedback loops: one governing the behaviour of sodium channels and one governing the behaviour of potassium channels. They might be described as coupled oscillators. Nonetheless, it is possible to understand how those oscillators are implemented in the behaviour of decomposable subsystems in the cell (the sodium and potassium channels).

Mechanists might claim further that dynamical models fail as explanations so long as they violate the 3M requirement discussed above, for precisely the reasons mentioned in favour of that requirement. Consider how the arguments of the previous section might be applied to a dynamical model in cognitive neuroscience, such as the Haken-Kelso-Bunz (Haken et al. 1985) model of human bimanual coordination (henceforth, HKB). Mechanists argue that such models are neither mechanistic nor explanatory of the phenomena they describe.

The HKB model (Eq. 2) describes a behavioural phenomenon arising in certain bimanual coordination tasks. Subjects are instructed to repeatedly move their index fingers up and down in time with a pacing metronome either in-phase (both fingers move up and down together) or antiphase (one finger moves up and the other moves down, and vice versa). Researchers systematically increase the tempo. Beyond a certain critical frequency, subjects can no longer maintain the antiphase movement and switch involuntarily into
in-phase movement. Subjects who begin in phase do not shift. Only in-phase movement is possible beyond the critical frequency.

Dynamicists characterize this system in terms of state-spaces and attractors representing the behaviour of two coupled oscillators. At slow tempos, the oscillators can be stably coupled to one another in both antiphase and in-phase movement modes. The state-space of the system is thus said to have two basins of attraction (or attractors) – one for antiphase and one for in-phase movement. At high tempos, only the in-phase mode is possible – the state-space has a single attractor. At the switch point, the attractor landscape of the system is said to change.

The HKB model provides an elegant mathematical description of this phenomenon, including the rate of change in the phase relationship between the left and right index fingers and the critical switch point from the antiphase to the in-phase pattern. The core of the model is a differential equation describing the coordination dynamics of these coupled components:

$$\phi = -A \sin \phi - 2B \sin 2\phi$$

(2)

where \(\phi\) is the phase difference (relative phase) between the two moving index fingers (when \(\phi = 0\) the moving fingers are perfectly synchronous), and \(A\) and \(B\) reflect the experimentally observed oscillation frequencies of the fingers (for further discussion of the HKB model see Haken et al. 1985; Kelso 1995; Kessler and Kelso 2001).

Crucially, the modellers do not intend HKB to be a description of mechanisms or of the motor systems responsible for the experimentally observed behavioural dynamics. Instead, it is a mathematically compact description of the temporal evolution of a purely behavioural dependent variable (relative phase) as a function of another purely behavioural independent variable (finger oscillation frequency). None of the variables or parameters of the HKB model map on to components and operations in the mechanism, and none of the mathematical relations or dependencies between variables map onto causal interactions between those system components (as required by 3M).

Predictivists might argue that, for complex phenomena, explanation is just prediction. Mechanists will insist, however, that this view commits the predictivist either to accepting as explanations many things that are not accepted in any science as explanatory (such as mere correlations, relationships between an effect and its causes, and inferences from effects of common causes), or to proposing some means of distinguishing phenomenal redescriptions from explanations, for distinguishing how-possibly explanations from how-actually explanations, and for tracking progress in the search for deeper explanations. Dynamicists such as Kelso appear to recognize the importance of mechanistic explanation. After developing HKB, Kelso and colleagues began a research
Towards a Mechanistic Philosophy of Neuroscience

programme to understand how this behavioural regularity results from features of the underlying organization of component neural systems and their dynamics (see, for example, Jantzen et al. 2009; Jirsa et al. 1998; Schöner 2002; Schöner and Kelso 1988). Dynamical models do not provide a separate kind of explanation subject to distinct standards. They are tools in the search for mechanistic explanations. Or so says the mechanist.

4. Functional and Structural Taxonomies

To discover a mechanism, one must first identify a phenomenon to explain. The taxonomy of cognitive capacities guides the search for their neural mechanisms. Neuroscience currently has nothing like the periodic table of chemical elements to guide its research projects: no systematic scheme of basic functional capacities of the mind or their neural and molecular mechanisms. To the extent that anything like such a taxonomy currently exists, it remains actively under revision, even in highly developed areas of neuroscience. As Bechtel and Richardson (1993) emphasize, one is often forced to revise one’s taxonomy of phenomena as one learns about underlying mechanisms. Conversely, one’s understanding of the mechanistic structure of the brain depends, to a large extent, on one’s characterization of the phenomenon one seeks to understand. Thus, neuroscience is apparently engaged in iterative cycles of recalibration between models at different levels of organization, each of which must fit its findings to what is known about other levels. Mechanistic realists and mechanistic pragmatists differ from one another in their understanding of this process.

Mechanistic realists expect to ground the taxonomy of cognitive and neural phenomena in objective facts about the mechanistic structure of the brain. By learning which mechanisms are distinct from which others, one can carve the mind-brain at its joints. This process terminates in the periodic table of the mind-brain (or something asymptotically approaching it). This pattern of reasoning is evident among proponents of ‘massive modularity’, who believe the mind is subdivided into functionally distinct, domain-specific subsystems (Boyer 2002; Carruthers 2006; Fodor 1983; Pinker 1999). It is also evident in a host of investigative techniques used by neuroscience to localize functions in the brain (such as clinical dissociations, lesion studies and neuroimaging). Dissociation experiments show that the mechanisms underlying one cognitive capacity can be disrupted independently of the mechanisms underlying another. They drive taxonomic reform by splitting capacities into mechanistically distinct units. If cognitive capacities can sustain damage independently of one another, the realist claims, the putative capacity is heterogeneous, and a revised taxonomy, splitting the capacity in question, will be more useful for
prediction, explanation and control than is the taxonomy that includes the heterogeneous capacity.\(^4\)

Pragmatists question the appropriateness of the periodic table as a metaphor for thinking about the taxonomic structure of neuroscience. They question whether there is a uniquely correct way of subdividing the functional capacities and mechanisms of the brain, and they emphasize that functional and structural decompositions depend in subtle ways on researchers’ explanatory goals, pragmatic interests and theoretical orientations (see, for example, Uttal 2003). They charge that the realist faces a dilemma: either the realist view regresses, or it is viciously circular. The regress threatens because the view presupposes some method of carving the brain into distinct types of mechanisms, and that method appears no less difficult than the method of carving the mind into functionally distinct capacities. On the other horn, circularity threatens because our ability to carve the brain into distinct mechanisms requires some idea of what those mechanisms do, and this requires some commitment about which capacities require explanations. Mechanistic characterizations necessarily filter out the parts, activities and causal interactions among parts that are irrelevant to the capacity to be explained. Filtering judgements such as these are guided by a prior understanding of the capacities one hopes to understand. The pragmatist argues that there is no antecedently intelligible mechanistic structure in the brain that carves the mind into objectively correct functional units – functional and structural taxonomies are interdeterminate.

Pragmatists also charge that antecedent commitments about the correct cognitive taxonomy creep into the analysis of dissociation experiments. (Similar considerations apply, \textit{mutatis mutandis}, to other localization techniques in cognitive neuroscience, including neuroimaging.) Neuropsychologists typically use a variety of tasks to provide a broad profile of the subject’s motor, sensory and cognitive competencies and deficits. From the observed pattern of competencies and deficits, inferences can be drawn about the existence of damage to one or more neurocognitive mechanisms. Assumptions about the correct taxonomy of the mind-brain enter into task design, task analysis and diagnosis.

Furthermore, such assumptions reciprocally influence views about brain organization. Some cognitive neuropsychologists (e.g. Dunn and Kirsner 1988) argue that dissociation experiments establish that the brain has a modular architecture only on the assumption that the brain has such a modular architecture. By relaxing this assumption and allowing, for example, that cognitive dissociations might be produced through damage to single neural processing systems, one becomes more cautious about inferring distinct functions from deficit dissociations. Olton (1989) shows that apparent dissociations can also arise by varying task demands on memory tasks. Relatedly, Plaut (1995) shows that non-modular connectionist architectures can exhibit
apparent double dissociations, thus providing further reason for caution in drawing inferences about underlying mechanisms from task dissociation evidence. Lambdon Ralph and Rogers (2007), for example, show that one can produce category-specific naming deficits in connectionist networks without separate modules for the storage of particular semantic categories. For these reasons, pragmatists argue that much of the evidence for the currently accepted taxonomy of brain functions makes nontrivial and contentious assumptions about the taxonomic structure of the mind-brain.

Regardless of how the debate between realists and pragmatists resolves, however, everyone can agree that the taxonomy of cognitive capacities is simultaneously anchored at multiple levels of organization. Churchland (1986) describes this process as the coevolution of theories at different levels. This coevolution develops as researchers with different techniques, vocabularies and even distinct standards of assessment attempt to negotiate their way to a consensus model of a mechanism. Neuroscience thus exemplifies a kind of conceptual development substantially different from that observed in more mature sciences, such as physics and chemistry.

5. Experiment and Evidence: Deficit Studies of Brain Mechanisms

Let us leave the debate between pragmatists and realists behind and assume we know which cognitive capacities require explanation and that we have tasks to distinguish them unambiguously. Can one, then, justifiably infer the organization of brain mechanisms on the basis of deficit studies (such as single and double dissociations)? Must we make additional assumptions beyond these? Are there kinds of mechanism that deficit studies alone are incapable of revealing under any reasonable set of assumptions?

Clark Glymour (1994, 2001; also see Bub 1994 for response) uses computational methods to investigate the conditions under which one can draw reliable inferences about underlying mechanisms on the basis of deficit studies. In order to apply his formal analysis, he makes a number of instructive assumptions in addition to those mentioned above (2001):

1. That behavioural deficits are all or nothing, not graded (p. 136);
2. That all normal individuals have the same cognitive architecture (p. 136);
3. That the cognitive architecture in patients is a sub-graph of the normal architecture (that is, the normal architecture minus one and only one component) (p. 136);
4. That the normal architecture has no cycles or feedback loops (p. 137);
5. That the normal architecture is non-redundant (p. 137).
Glymour shows that if all of these are true and, in addition, we eventually see all of the patients that the cognitive architecture allows, then it is possible in many cases for cognitive scientists to converge on the correct cognitive architecture on the basis of deficit studies alone. This is a powerful result. However, some cognitive architectures are indistinguishable on the basis of deficit studies alone. For example, it is impossible for deficit studies to distinguish a cognitive architecture with one variable (V) between input and output and a cognitive architecture in which two variables (V1 and V2) lie in sequence between the input and output (Glymour 2001, pp. 144–5). Deficit studies also have difficulty distinguishing cases in which two cognitive mechanisms have been shown to be distinct from those in which a single cognitive mechanism has merely lost the resources (e.g. computational resources) required for it to perform both of its typical functions (Glymour 2001, pp. 146–8). Indeed, adding resource considerations into the models of brain function makes even simple causal architectures exceedingly difficult to discover.

Glymour’s starting assumptions, as he acknowledges, are frequently false. Patients often have multiple deficits resulting from remarkably different kinds of brain damage and disease, and the deficits lie anywhere on a spectrum from mild to global impairment (see Van Orden et al. 2001). Brains vary considerably from individual to individual (see Ojemann 1991). Brains reorganize in response to damage and disease, and new systems can sometimes take over the functions of the damaged system (see Hardcastle and Stewart 2002). Among the most dramatic examples of neural plasticity is the discovery that congenitally blind subjects use the primary visual cortex when performing demanding tactile discrimination tasks, such as braille reading, in comparison to normal subjects, who show deactivation in the same areas in equivalent tasks (Sadato et al. 1996; Cohen et al. 1997). In deficit studies, we do not study what the missing part did, but rather what the rest of the brain can do without it. Finally, brain systems are paradigmatically complex, involving multiple feedback loops at multiple levels and multiple redundant pathways that make development and recovery possible.

How does neuroscience make progress in the face of such challenges? Neuroscientists have an arsenal of techniques and strategies beyond deficit studies to study how brain mechanisms work. These techniques and strategies have non-overlapping strengths and weaknesses; they are more valuable when taken together than when taken singularly (see Bechtel 2002, 2007; Craver 2002, 2007). Deficit studies are a species of bottom-up inhibitory experiments in which one intervenes into a mechanism to inhibit one of its components, while monitoring the behaviour of the mechanism as a whole. Other examples include gene knockouts, transcranial magnetic stimulation and the use of pharmacological agents to inhibit systems in the brain. Such
interventions can often be delayed and reversed, allowing for subtly controlled manipulations.

Activation experiments, in contrast, involve activating a mechanism as a whole and monitoring the behaviours of its components. For example, subjects are engaged in tasks while brain activity is measured (using, for example, single-unit and multi-unit microelectrode recording and functional neuroimaging methods, including functional magnetic resonance imaging (fMRI), diffusion-tensor imaging (DTI), positron emission tomography (PET), electroencephalography (EEG), and magnetoencephalography (MEG)). Finally, stimulation experiments involve intervening to drive the components of a mechanism, while monitoring the behaviour of the mechanism as a whole. One might, for example, stimulate neurons in a brain region in order to produce specific motor patterns (Graziano et al. 2002) or to alter visual decision-making about motion direction in non-human primates (Salzman et al. 1990). Normative assessment of any single technique or strategy in isolation will yield more pessimistic results than would an analysis of the full range of mutually reinforcing techniques. Philosophers have focused on the integration of theories (Oppenheim and Putnam 1958) and the integration of fields (Darden 1991), but contemporary neuroscience affords the opportunity to consider how experimental strategies are integrated in the search for mechanisms.5

6. Computation and the Brain

Contemporary neuroscientists use computation to analyse, interpret and cross-validate experimental data; to generate testable predictions and refine theoretical hypotheses; and to optimize experimental designs (see, for example, Koch 1999; Koch and Segev 2000; Sejnowski 2009). Some neuroscientists and philosophers, computationalists, also believe that the brain is a computer. They hold that brain systems perform computations and employ neural codes (see, for example, Rieke et al. 1999). Fictionalists think computation is in the eye of the modeller. It might be useful to describe the brain as a computer, but there is no well-defined sense in which neural systems, in fact, compute. The brain is simply a collection of causal mechanisms – full stop (Searle 1992).

To resolve the debate between computationalists and fictionalists, one must first decide what it means to compute. Grush (2001) cautions that an account of computation must navigate between the twin perils of triviality and irrelevance. On one hand, the notion cannot be left so permissive that everything computes (see Piccinini 2007a). For example, if one holds that a system computes if and only if a computer can simulate it, then brains compute. But so do thunderstorms, turbulent flows and orbiting planets. Or, if one holds that a system computes when its input-output transformations can be described
as Turing computable functions, then most worldly causal processes count as computational. Churchland and Sejnowski (1992), and Sejnowski, Koch and Churchland (1988) hold something like this view: the brain computes in the sense that (a) its input-output behaviour can be interpreted in terms of some computable mathematical function, and (b) this interpretation is something we find useful or otherwise revealing. Clause (b) adds only the pragmatic thesis that, for whatever reason, computational explanations are especially revealing in a particular investigative context. This view postpones the question of why computational explanations are revealing in some cases and not others. It also looks very close to the fictionalist thesis that computing is in the eye of the beholder.

On the other hand, the notion of computation cannot be so restrictive as to make it utterly implausible from the start that brains compute. If one defines computation in terms of the behaviour of serial, von Neumann architectures or as operations on well-defined symbol strings according to a set of explicitly stored rules or lines of programme code (e.g. Cummins 1985), then the brain is not really a candidate for computing. The available models of computation from computer science (including abstract connectionist or neural network models (Rumelhart et al. 1986)) are unlike brains in many respects. It would appear that computational neuroscience needs a proprietary model of neural computation.

Grush (2001), Eliasmith and Anderson (2004), and Shagrir (2006, 2010) have each attempted to develop such a proprietary notion. Grush argues that brain systems compute in the sense that they process information or transform internal representations. This view fits standard examples of computational neuroscience. Single neurons in the primary visual cortex (V1) are maximally active for stimuli presented at preferred orientations. They are commonly said to represent those orientations (Hubel and Wiesel 1968). Neurons in visual area MT fire preferentially for objects moving at specific speeds and in specific directions. Neurons in the motor cortex (M1) exhibit cosine tuning, that is, each cell is maximally responsive (fires most action potentials or spikes) for a particular direction of movement with activity declining as a function of the cosine of the angle between the preferred and actual movement direction (Georgepoulos 1982, 1986). Perhaps, as Grush claims, the proprietary notion of computation involves the manipulation of representations or the processing of information.

Why is this particular idea of computation useful to neuroscientists? Churchland (Churchland and Grush 1999) and Shagrir (2006) argue that mechanistic descriptions of neural systems in terms of firing rates of neurons across populations, for example, without any reference to what those activity patterns represent or how those representations are being transformed, would leave the functional role of those activity patterns utterly obscure. Shagrir argues that computational terms are essential when the project is to explain
how a semantic task, that is, a task defined in terms of representational content, is performed (2006, p. 393). This semantic view of computation is a received view in the philosophy of mind and in mainstream cognitive science. As Fodor (1981) demands: ‘no computation without representation.’

Fictionalists hold out against these suggestions. They worry that the appeal to representation only delays the twin threats of triviality and irrelevance. To the extent that any causal system can be described as carrying information (as a river’s height carries information about rainfall) or processing information (as a submerged river rock processes inputs related to current speed and outputs some transformation of it), the idea that the brain computes appears trivial. It amounts only to the claim that brain processes are causal processes, or that its various states are correlated with one another or with aspects of the external environment. Alternatively, one might attempt to apply a more restrictive notion of information that does not apply equally to weather patterns and turbulent flows.

One might require, for example, that states of the system have been selected for their being correlated with something else, or that the system has been trained up through development and learning so that certain states are so correlated (Dretske 1988, 1995). One might require that the states in question should be available for use by the organism in guiding behaviour. One might require further that information in the parochial sense refers only to states that are re-presenting (in the sense that they are duplicating some other input to the system). Critically, the fictionalist can argue that these more restrictive notions create bigger problems than they solve. If granted, such notions potentially render neuroscience obsolete for the task of identifying neural systems that implement computations. If computations are defined over representations that count as such if and only if they have the appropriate causal or evolutionary history, for example, then neuroscience will, in principle, be unable to answer questions about whether, and what, a given neuron or neuronal populations represents, or what computation a neuron performs. And this, one might reasonably think, is a far worse outcome for neuroscience than any consequences of adopting a fictionalist stance (e.g. utilizing a notion of computation with built-in subjectivity or observer-dependence). Thus, fictionalists will suspect that none of the proposed specifications will avoid both of Grush’s twin perils: they will remain either overly broad, applying to systems that do not seem to be computing, or overly restrictive, ruling out systems that we think are genuinely computational.

The syntactic view of computation, in contrast, holds that computation can be defined without appealing to semantic notions (e.g. Egan 1995; Piccinnini 2007b). Piccinnini (2007b) argues that computing mechanisms are defined structurally in terms of system transformations of input strings of digits into output strings according to a rule defined over them, without reliance on
semantic interpretation of the inputs and outputs (or internal states). Anything that has these structural features is a computer. The view avoids the evil of triviality by placing strict constraints on systems that count as computers. Thunderstorms and rivers will not count as computers on this view. Concerning the claim of irrelevance, this view can respond that the account of computation comes with the credentials of the classical theory of computation (for overview, see Sipser 2005). Who better to settle the dispute about what does and does not count as a computer than the researchers responsible for developing our most rigorous, formal treatment of computation? And given that something like this abstract notion is implemented in the myriad digital computers that now surround us, there is at least a large class of things to which the account applies.

Fictionalists, however, will no doubt deny that the syntactic view satisfies Grush’s challenge. For although the class of syntactic computing mechanisms is much smaller than the class of systems supporting computer simulation, there are still more computing mechanisms than a gripping computationalist hypothesis would likely countenance. DNA replication mechanisms will count as computing, as will gumball machines. Relevance is also a worry. The syntactic account is well suited to describe computation in finite state automata, Turing machines and digital computers. But there is good reason to wonder whether such a rigidly specified syntactic account can find application in the wetware of the brain. As Shagrir argues, some of the central successes of computational neuroscience, such as Robinson’s (1989) network model of the oculomotor integrator, would not count as computing mechanisms, for the simple reason that the system’s inputs and outputs have continuous values and so are not well-defined strings in any sense. The fictionalist will, therefore, likely object that this notion is at once too permissive (in countenancing too many systems to be interesting) and too restrictive (in limiting the application of the term ‘computation’ to an overly restrictive class of system).

A second debate about computation in neuroscience concerns whether computational explanations have a distinctive kind of explanatory value over and above non-computational explanations. Rather than survey the vast range of positions on this matter, we hope to motivate a view of computational explanations in neuroscience that fits neatly into the mechanistic perspective discussed in Sections 2 and 3: computational models that shoulder explanatory weight do so in virtue of being mechanistic explanations.

There is a widespread intuition that computational explanation is fundamentally distinct from mechanistic explanation. Marr (1982) expressed this view in his famous tri-level framework for cognitive science, in which information-processing and computational levels are independent of lower levels of analysis. Sejnowski, Koch, and Churchland (1988) likewise insist that ‘Mechanical and causal explanations of chemical and electrical signals in the
Towards a Mechanistic Philosophy of Neuroscience

brain are different from computational explanations’ (p. 1300). They explain: ‘The chief difference is that a computational explanation refers to the information content of the physical signals and how they are used to accomplish a task’ (p. 1300). The mechanist will argue that one can consistently agree about the distinctiveness of computational models, and yet disagree that the form of explanation they provide is essentially distinct from other more familiar mechanistic explanations.

From a mechanistic perspective, computational explanations must satisfy the same criteria of adequacy as other mechanistic explanations (see Section 2). In particular, they should respect the realistic assumptions built into 3M: components and transformations in the computational model should map onto components and activities in the mechanism. However, mechanists need not deny that computational descriptions are distinctively valuable as interpretive tools: computational explanations of neural systems help us to keep track of the interpretive mapping between internal states of the system and mathematically representable quantities in the world (as Shagrir 2006 argues). However, mechanists insist that providing such informational mappings is insufficient for computational models to have explanatory force. After all, one might develop such mappings, not on the basis of components in the system that produce the target behaviour or cognitive function in question, but on the basis of components whose states are merely correlated with aspects of the target behaviour or cognitive function. For the mechanist, computational explanations are explanatory to the extent that they describe the causal structure of a system. Computational interpretation without causation is not explanation. Furthermore, one can explain information processing tasks mechanistically, without reference to representational states. Computational explanations are, in many cases, more readily intelligible than uninterpreted mechanistic explanations would be, but hard-to-understand explanations are still explanations.

The mechanist will thus disagree with Shagrir’s understanding of the role of computation in providing explanations: ‘[W]e adopt the computational approach because we seek to explain how a semantic task can be carried out, and computational explanations are able to do this. But what is the source of this explanatory force? Why are computing mechanisms able to explain semantic tasks? I suggest that the explanatory force of a computing mechanism derives from its correspondence to mathematical relations between the represented objects and states’ (Shagrir 2006, p. 394). Mechanists grant that Shagrir identifies crucial features warranting the computational interpretation of a system (that its states map onto mathematical relations in the world). However, they argue that the explanatory force of such models lies not in the interpretation the model supports, but in the causal structure the model correctly describes.
7. Mechanism and the Unity of Neuroscience

We close on the topic with which we began. Neuroscience is a multi-field discipline with (as yet) no grand unifying theory. If the goal of neuroscience is to elucidate the workings of the mind-brain by bringing together diverse investigators studying different aspects of brain structure and function, what form should we expect this union to take? And what benchmarks measure our progress?

Some take a reductionist viewpoint, arguing that the unity of neuroscience is achieved by assimilating higher-level phenomena to lowest-level phenomena. Bickle (1998, 2003), for example, argues that the success of contemporary neuroscience consists in its ability to reveal cellular and molecular mechanisms for cognitive and mental phenomena. This position is evidenced by the fact that one can often change organism-level behaviours by intervening on genes, molecules and ions. The rest of neuroscience – what lies between behaviour and the cellular and molecular mechanisms – is of tremendous heuristic value in defining which mechanisms are explanatorily relevant, but ultimately, it is the molecular mechanisms that do all the explanatory work. The most predictively adequate and instrumentally useful neuroscience will forge mind-to-molecule linkages directly, leaving out the purely heuristic middle realm of cells and physiological systems. Bickle calls this view ‘ruthless reductionism’.

Ruthless reductionism has a cousin in Weber’s heteronomy thesis. Weber (2005) argues that biological generalizations have no distinctive explanatory value and that the laws of physiology are evolutionarily contingent. As such, they lack the requisite necessity to underlie explanations. This requisite necessity can be found only at the physical and chemical level. Thus, biology is heteronomous, subject to the external authority of physical law. To explain phenomena in neuroscience, it is necessary to relate physiological phenomena (such as action potentials) to physico-chemical laws (such as Ohm’s law and the Nernst equation). Biology merely fixes the conditions on which physical laws operate.

The integrationist alternative to these views accepts that there are adequate explanations at multiple levels of organization and insists that many adequate explanations in neuroscience must span multiple levels. Mechanists tend to be integrationists (Bechtel 2007; Craver 2005; Darden 2006; Machamer et al. 2000). They see the nervous system as composed of nested hierarchies of mechanisms within mechanisms linking behaviours in social environments to the behaviours of molecules and ions via intermediate levels of organization. For example, neuroscientists studying spatial learning in rodents have now found crucial linkages between multiple levels, from the lower-level behaviour of ion channels, through forms of synaptic plasticity among neurons, to the
behaviour of brain regions, to the coordination of social behaviour among conspecifics (Squire and Kandel 2000; see also Bickle 2008).

Integrationists deny the reductionist thesis that biological generalizations lack explanatory force. For them, any generalization that holds up under the conditions of a well-controlled experiment arguably counts as explanatory (Craver 2007; Woodward 2003), regardless of whether the regularity in question is evolutionarily contingent and regardless of whether the generalization has any direct connection with physical or chemical laws. Here, we offer three considerations favouring an integrationist perspective.

First, a reminder: no area of neuroscience is metaphysically fundamental. Neuroscience is a science of the middle range, bottoming out well before the physical bedrock that reductionists hope to someday reach. To the physicist, ion channels are complex and messy biological objects. Yet such channels are at, or near, the rock bottom of today’s neuroscientific hierarchy.

Second, neuroscientists have discovered causal generalizations that satisfy the requirements of controlled experiments at multiple levels. Levels of causal regularity are to be expected in a world in which interacting parts are organized into mechanisms, within mechanisms, within mechanisms. Basic features of evolution by natural selection make it likely that nature exhibits a hierarchy of nearly decomposable levels of organization that are explanatorily salient and practically useful (Simon 1969; Steel 2008; Wimsatt 2007). Levels are not merely heuristic guides to the search of molecular mechanisms. They supply veritable handles in the causal structure of the mind-brain that can potentially be used for the purposes of curing diseases, improving function and manipulating the brain for good or for ill. Research at, and across, levels is a crucial guide to making interesting predictions (diagnoses, prognoses) and designing interventions (therapies, drug treatments). Were neuroscience to aim for an endgame of ruthless reduction or heteronomy, it would discourage the search for these higher-level forms of intervention (see, for example, Ramachandran et al.’s (1995, 1998) work on phantom limbs). If one thinks of intermediate levels of organization as mere shadows of lower-level activity, inefficacious abstractions about the behaviour of molecules, one is less likely to think about designing interventions that take advantage of higher levels of organization.

Finally, integrationist strategies are useful tools for concept development. By integrating descriptions at multiple levels, scientists place their concepts under constraints from other levels of organization. The action potential is taken to be a robust phenomenon, in part, because of how it figures in neurotransmitter release and in the behaviour of networks of neurons. The membrane mechanism that Hodgkin and Huxley hoped to discover is visible as a mechanism only on the assumption that the action potential is a meaningful unit to be explained in the first place. To echo an earlier theme, at the molecular
level, the world is a busy, buzzing confusion. The molecular level has to be segmented into mechanisms by subtracting out irrelevant parts and irrelevant interactions. To do that, one needs to know which higher-level phenomena to take seriously. The effort to bridge levels of organization is, in this sense, a procedure for testing whether the concept is empirically adequate. It is an epistemic benefit of integrationist thinking.

8. Conclusion

Despite exponential growth in the neurosciences over the last several decades, the goals, methods and practices of neuroscience have received scant attention from philosophers of science. This chapter surveys a number of current trends and open debates in the philosophy of neuroscience, and it shows that the mechanistic framework affords a fruitful perspective from which to develop answers to many of the core issues in philosophy of neuroscience, including explanation, methodology, computation and reduction. The 1990s, hailed as the ‘Decade of the Brain’, has long since come to an end. Given the exciting developments unfolding in philosophy of neuroscience in recent years, the prospects for a decade of the philosophy of the brain are now extremely bright.

Notes

1 See Bechtel et al. (2001), Machamer et al. (2001) and Bickle (2009) for recent collections dealing with these and related issues.

2 Making explicit his original aims, Kelso (1995) states: ‘one of the main motivations behind these experiments was to counter the then dominant notion of motor programs, which tries to explain switching (an abrupt shift in spatiotemporal order) by a device or mechanism that contains “switches.”’ (1995, p. 57). Appearances notwithstanding, Kelso’s motivation was not to replace one flawed mechanistic hypothesis with another. In an earlier passage, Kelso rails against all models of motor behaviour that invoke underlying neural mechanisms, including those that invoke motor programmes (the stored sequence of instructions to drive the muscles during a movement), asserting that ‘[a]ny time we posit an entity such as reference level or program and endow it with content, we mortgage scientific understanding.’ (1995, pp. 33–4). For Kelso, the HKB model was a full-fledged explanation, even though it does not describe mechanisms.

3 Models such as HKB at most involve behavioural ‘components’, such as the phase relationship between the index fingers, but these are features of the phenomenon and should not be confused with parts of a mechanism (see Bechtel 1998). Van Gelder makes a similar point about dynamical models of cognition: ‘the variables [dynamical models] posit are not low level (e.g., neural firing rates) but, rather, macroscopic quantities at roughly the level of the cognitive performance itself’ (1998, p. 619). Hence, the HKB equation resembles other well-known empirical
Towards a Mechanistic Philosophy of Neuroscience

‘laws’ of human motor behaviour such as Fitts’ law (Fitts 1954), which quantifies the robust empirical relationship between target size and distance, movement speed and accuracy.


5 Sullivan (2009) claims that the unity of neuroscience is impossible, owing to the heterogeneity of experimental practices and protocols across and within research laboratories.

6 In computer science, a computable function is defined as a mathematical function in which the mapping relation between input and output can be specified by a rule or algorithm, or step-by-step procedure. See Sipser (2005), or for a seminal treatment, see Turing (1936).

7 Weber (2005) allows that there might be laws at higher levels of organization than physiology, such as at the level of population genetics.

References

—(2002), ‘Aligning multiple research techniques in cognitive neuroscience: why is it important?’, Philosophy of Science, 69, 48–58.


—(2003), Philosophy and Neuroscience: A Ruthlessly Reductive Account. Dordrecht: Springer.


Towards a Mechanistic Philosophy of Neuroscience

— (2003), ‘What can we infer from double dissociations?’, Cortex, 39, 1–7.


Towards a Mechanistic Philosophy of Neuroscience


1. Introduction

Chemistry studies chemical substances and chemical change, using a range of compositional concepts: substance, element, compound, mixture. Chemists use detailed theories of the microscopic structure of substances to explain the differences between various chemical substances and the different reactions they undergo. This makes elements, compounds, molecular structure and the chemical bond chemistry’s fundamental concepts, and their physical basis one of the central questions for the philosophy of chemistry. Structure and the bond present surprisingly problematic issues, as we shall see. Not surprisingly, chemistry is of central interest to philosophers of chemistry, who are motivated by an intrinsic interest in the science, but it should also be of interest to philosophy at large. It offers some clear and widely cited examples of central philosophical notions: natural kinds (gold, water); theoretical identity (‘water = H₂O’); mechanisms (of chemical reaction); reduction (chemistry to physics of course, but also of genes to molecules) and scientific explanation by appeal to unobservables (atomic and molecular structure). Surely it is the business of the philosophy of chemistry to provide informed critical examination of chemical examples in philosophy: whether to debunk them, to defend them, or to point out the complexities and subtleties they involve. Yet until recently, philosophers have almost entirely ignored chemistry, for reasons which remain unclear (see van Brakel 2000, chapter 1; 2011). In the last ten years this has changed, and in what follows I will survey the more recent developments, concentrating on the conceptual and metaphysical issues raised by substances, bonds, structure and reduction. Hopefully this will provide a persuasive case that philosophy of chemistry can offer an arena for fruitful interaction between general philosophy of science, metaphysics, philosophy of language and the philosophy of mind.¹

2. Chemical Substances

Since chemistry is the science of the transformation of substances, one central issue in the philosophy of chemistry is how substances are individuated. What
makes gold the metal that it is, and water the liquid that it is, and why? This is chemistry’s version of the more general issue of natural kinds, mirroring the species question in the philosophy of biology. As such, it raises many familiar philosophical questions. How ‘natural’ are the divisions between chemical kinds? Is there more than one such possible system of divisions? In what follows, I will critically examine two broad philosophical accounts of the classification of substances, which emphasize microscopic and macroscopic properties, respectively. A word on scope: I will concentrate on substances as studied by chemistry. Some kinds of substance are clearly picked out by classificatory interests that are different from chemistry’s: artificial silk, for instance, might be chemically identical to silk, but there is also a sense in which, because it is made in a factory rather than in a silkworm, artificial silk is not silk. In this case, the causal history of the substance, and in particular its relationship to a biological species, seems more important than its chemical composition. Conversely, what counts as silk in virtue of its origin in a silkworm might well be heterogeneous from a chemical point of view. Foodstuffs, too, may be identified by where they have come from, rather than what they are made of, and the same goes for wool and wood. Jade provides a different kind of example: it is well known to philosophers that ‘jade’ applies to two quite different chemical substances, jadeite and nephrite, application of the term being constrained by appearance and economics as well as constitution (for details, see LaPorte 2004, pp. 94–100). The salience of non-compositional factors like causal history or appearance, however, comes from classificatory interests that are foreign to chemistry. In the following discussion I will concentrate on substances from the chemical point of view.

2.1 Microscopic Conceptions of Substance
Judging by appearances, modern chemistry regards microstructural properties as central to the identity of chemical substances. Since 1923, the International Union of Pure and Applied Chemistry (IUPAC) has identified the elements in terms of the nuclear charge of their atoms, overturning the tradition handed down from Dalton and Mendeleev that atomic weight is the key property (see Hendry 2010a for discussion). Furthermore, IUPAC’s rules of systematic chemical nomenclature for compound substances involve properties at the molecular level: macroscopic properties are not mentioned (see, for instance, Thurlow 1998). Microstructuralism about chemical substances is the thesis that microstructural properties dominate the classificatory interests of chemistry, and is a natural extrapolation from the above facts about chemical nomenclature. In the 1970s, Saul Kripke and Hilary Putnam famously advanced microstructuralist claims about two chemical substances, arguing that scientists had discovered that the essences of gold and water, respectively,
are their atomic number and molecular structure (Kripke 1980, Putnam 1975). Their philosophical aims were quite different. Putnam was trying to develop a semantics for theoretical terms in which it is possible for meaning to be stable across radical theoretical change, accommodating the realist thought that science investigates a world in which divisions between kinds are given by nature. Kripke’s more ambitious project was to resurrect essentialism and ground it in necessity (see Hendry 2010a). They shared some tools, however: in particular, the assumptions that nature is divided into kinds, and that kind terms function somewhat like proper names. Just as famously, Kripke and Putnam have been attacked on many fronts. First, their examples were presented in a scientifically simplistic manner, and further detail appears to undermine their identifications (Needham 2000, 2002; van Brakel 2000). Secondly, the model of reference Kripke and Putnam proposed does not really entail that microstructural properties constitute the essences of chemical substances that have been discovered empirically (LaPorte 2004). Lastly, Needham (2011) contends that, contrary to Kripke’s arguments, it is perfectly possible to use macroscopic properties to individuate substances. There seems no need for any appeal to hidden microstructural essences after all.

Let us consider those chemical details: the objection is that elements and compounds can be heterogeneous at the microstructural level. It is worthwhile questioning the principle behind this objection. Historically, atomism has often been assumed to see pure substances as collections of qualitatively identical atoms and molecules. Pierre Duhem (2002, 86) and Jean Timmermans (1941, chapter 8) criticized atomism on just those grounds. But elements as defined by IUPAC consist of atoms of like nuclear charge, which allows that they may differ in other respects, like mass. Putnam seemed to imply a requirement of homogeneity in saying that the extension of ‘water’ is ‘the set of all wholes consisting of H2O molecules’ (1975, p. 224). But liquid water is far from homogeneous at the molecular level (see Needham 2000, 2002; van Brakel 2000, chapter 3): a small but significant proportion of H2O molecules dissociate, forming H3O+ and OH− ions; H2O molecules are polar and form hydrogen-bonded chains similar in structure to ice. One response to this is to regard the ionic dissociation products as impurities, but this seems too revisionary of chemical usage, because the presence of H3O+ and OH− ions is central to understanding water’s electrical conductivity. Paul Needham and Jaap van Brakel’s response is to reject the microstructuralist contention that to be water is to be composed of H2O molecules, which raises the question of just what makes something water, if not its molecular composition (I will address some macroscopic alternatives in the next section). But molecular heterogeneity is consistent with microstructuralism if one sees water not as a mere assemblage of H2O molecules, but as a heterogeneous molecular population that is generated by bringing H2O molecules together (for details, see Hendry 2006).
So microstructuralism need not require that substances are homogeneous populations of molecular or atomic species, or even that every member shares some property, although elements are, in fact, homogeneous in that way. The requirement is only that the populations as a whole have some significant microstructural property. It is ironic that Putnam chose water as an example. It is a difficult case for the microstructuralist, and other molecular compounds are less problematic, simply because they are more homogeneous at the molecular level. Yet the heterogeneity issue may still have some bite. Once we realize that populations of molecules have many properties – some which are shared across the population, and some which vary – we might ask which property is ‘the’ microstructural essence. A failure to find an answer to that question will surely undermine the whole idea of microstructural essences. Thus, LaPorte (2004, chapter 4) argues that in past usage, the names of substances may have been indeterminate in respect of which isotopes they apply to. If there is no fact of the matter whether, for instance, heavy water (deuterium oxide, D₂O) fell within the extension of ‘water’ as used by scientists before the twentieth century, then chemists’ current usage, according to which heavy water is water, must reflect a decision rather than a discovery. But LaPorte does not consider the actual history of chemistry, which may well reveal that, long before the twentieth century, there were classificatory interests in chemistry that rule out the referential indeterminacy on which his argument depends (see Hendry 2006, 2010a and Bird 2010; see LaPorte 2010 for responses).

One important challenge for microstructuralism is whether it can provide some account of the distinction between compounds and mixtures. This distinction would seem to be fundamental to chemistry and is closely related to that of the individuation of substances, because something is a mixture just in case it contains two different substances. Yet it presents a difficulty because there are penumbral cases, such as solutions (see van Brakel 2000, pp. 82–7; Needham 2011). Should these be counted as mixtures, or as compounds? At the macroscopic level, solutions are homogeneous, unlike (for instance) a mixture of iron filings and powdered sulphur, in which one can see the distinct components with the naked eye. Interstitial compounds are another example: these are formed by forcing the smaller atoms of one kind into the ‘spaces’ in the lattice formed by much larger atoms of another kind (typically, transition metals). Interstitial compounds have been compared to solid solutions of gases in metal lattices, so one might say that the components of a compound are chemically combined, rather than merely juxtaposed as in a mixture, but what is ‘mere juxtaposition’? Does it preclude physical interaction? Surely not: molecules always interact always interact, even when in mixtures. The weakest chemical interactions between atoms and molecules are van der Waals forces: dipole interactions between instantaneously asymmetrical charge distributions
of non-polar species, inducing dipoles in their neighbours. Though very weak, these forces are no different in kind to the stronger hydrogen bonds that constrain the shapes of large molecules like proteins. From a microscopic point of view, it would seem that the forces that act within molecules to hold them together are continuous with the forces that act between molecules of different kinds in a mixture, and the distinction between compounds and mixtures appears correspondingly vague. But distinguishing substances from mixtures is the same as distinguishing between substances, so if nature does not make the first distinction, it does not tell us how many different substances there are. It seems that the microstructuralist must accept this consequence. Yet, as we shall see in the next section, there are attempts to distinguish macroscopically between compounds and mixtures. If these are successful, then there is a determinate macroscopic distinction between substances and mixtures for which there is no microscopic basis. In that case, microstructure alone would seem to be insufficient for a philosophical understanding of chemistry.

2.2 Macroscopic Conceptions of Substance

What, if not nuclear charge and molecular structure, determines the identity of substances? One might consider the aims and interests of chemistry as a discipline: according to Jaap van Brakel (2000, chapter 3), chemistry is a ‘science of stuffs’ that manipulates and transforms macroscopic quantities of substances, investigating their location in a network that ‘contains all possible substances’ (2000, p. 72). Substances are the nodes, while chemical interactions form the connecting relations. This is an interesting suggestion, but further discussion would require that it be developed well beyond this metaphor. While it is true that synthesis is central to chemistry (see Schummer 1997), the microstructuralist will object that no such structure as van Brakel’s network of substances appears explicitly in works of chemical classification and nomenclature. Operational definitions of substances are also conspicuous by their absence. As we have already seen, the IUPAC rules of nomenclature concern molecular structures. Moreover, as scientific understanding has advanced during the nineteenth and twentieth centuries, chemists have increasingly described substances, and understood the transformations between them, in terms of the way their constituent atoms are bonded together. This suggests that the aims and interests of chemistry point straight to microstructural properties.

An alternative macroscopic approach is to employ the notion of separability: something is a mixture just in case that it is composed of physically separable components. Wilhelm Ostwald (1904) and Jean Timmermans (1941) offered such conceptions of substance, as has Paul Needham more recently (see Needham 2007, 2010, 2011). Since classical thermodynamics provides the general theoretical framework for understanding the processes by which
chemical substances are separated, it is the most likely source for a macroscopic conception of substance.

The general idea is that the thermodynamic behaviour of a system is dependent on the number of distinct chemical substances present in it, which provides a criterion for when a substance is pure, rather than mixed. This, in turn, implies a criterion for the sameness and difference of substances. This will be a macroscopic criterion in so far as classical thermodynamics describes the behaviour of matter independently of its microscopic structure (Zemansky 1957, pp. 1–3; Denbigh 1966, p. 3). In fact, the thermodynamic behaviour of a system depends in a number of ways on the number of substances or components present in it, three of which I consider below.

(i) When a mixture of substances is separated by distillation, it is boiled and the vapour is collected and cooled, providing a liquid that is enriched in one component. This distillate can successively be redistilled, purifying the relevant component to whatever degree is required. This practice relies on the fact that a mixture and its vapour typically differ in composition, while a pure substance has the same constitution as its vapour. Constancy of composition over phase change (solid to liquid, or liquid to gas) might seem to provide a thermodynamic criterion for distinguishing pure substances from mixtures (Ostwald 1904; Timmermans 1941, chapter 2), but an immediate problem concerns azeotropes. Azeotropes are mixtures of different compounds (e.g. water and ethanol, or water and hydrogen chloride) which, under certain conditions, can be distilled without affecting their composition. Ostwald distinguished between azeotropes and compounds proper on the grounds that azeotrope composition varies with pressure (Ostwald 1904, pp. 514–16; Timmermans 1941, chapter 2). In fact, altering the pressure at which distillation occurs is one way to separate the components of an azeotrope. In this respect, azeotropes are more like solutions and other mixtures. The composition of a saturated solution of salt in water, for instance, depends on temperature. In contrast, the composition of most compounds is constant within the range of physical conditions in which they exist. But why does variable composition disqualify an azeotrope from being a compound? There was an extended debate in the first half of the nineteenth century over whether compounds must have a composition that is fixed and independent of physical conditions, because a range of substances (Berthollides) display variable composition and not all can be regarded as solutions or mixtures (see Needham 2008, pp. 67–8; Needham 2009). Constancy of composition over phase transitions therefore seems to fail as a criterion for being a pure substance, since it requires appeal to an independent, and indeed contentious, criterion for purity of substance.
(ii) A phase is a homogeneous body of matter with a definite boundary. Gibbs’ elegant and abstract Phase Rule relates the number of components (C) in a multiphase system to the number of phases (P) present in it, and its variance (F), which is the number of independent physical variables required to characterize the thermodynamic state of the system:

\[ F = C - P + 2 \]

Does the phase rule provide a criterion of sameness and difference for substances? The difficulty is that ‘components’ cannot simply be equated with chemical substances, and in some systems it is a matter of some delicacy how to map one onto the other so as to preserve the truth of the phase rule. Needham (2010, Sections 4 and 5) discusses the three-component system of calcium carbonate (CaCO₃), carbon dioxide (CO₂) and calcium oxide (CaO): according to the phase rule, this system contains only two components. It seems that the phase rule doesn’t itself say how many substances are present in a system, but only once a case-by-case matching of phase-rule components to chemical substances has been achieved. If so, then the phase rule doesn’t really offer a criterion for distinguishing substances and mixtures, but only a series of relationships upon which we can impose a carefully constructed interpretation which does so. The question is whether or not such an interpretation amounts to a bookkeeping exercise, rather than an autonomous conception of substance.

(iii) Whenever samples of two distinct substances in thermal equilibrium are allowed to mix, there is an increase in entropy. In contrast, samples of the very same substance generate no entropy change on isothermal mixing. This gives rise to ‘Gibbs’ paradox’, that entropy of mixing does not depend on the degree of similarity between the two substances, but is a step function of the number of distinct substances being mixed (see Denbigh and Denbigh 1985, chapter 4; Denbigh and Redhead 1989 for discussion of the paradox and solutions to it). Entropy of mixing generates a criterion for sameness and difference of substance in an obvious way, one which Needham (2008) applies to water, arguing that it is a mixture because it contains traces of isotopic variants like \(^{1}\text{H}_{2}^{16}\text{O} \) and \(^{2}\text{H}_{2}^{18}\text{O} \) as well as the more common \(^{1}\text{H}_{2}^{16}\text{O} \): such isotopic variants display non-zero entropy of isothermal mixing. However, entropy change on mixing is displayed by many pairs of species that chemists do not normally regard as different substances, but rather as the same substance in different physical states: examples include spin isomers like ortho- and para-hydrogen and populations of otherwise identical atoms in orthogonal quantum states, like the two beams of silver atoms emerging from a Stern-Gerlach apparatus.
If one accepts that all these are different substances, it seems that, on the entropic criterion, any physical difference (except in respect of temperature) must count as a difference of substance. This may be logically consistent, but hardly seems consonant with the aims and classificatory interests of chemistry, according to which chemical substances are continuants that may vary in their physical properties. This is an important consideration when the question is how to understand chemical sameness and difference (for further details of the entropic criterion, see Hendry 2010c).

2.3 Conclusion
Criticizing the microstructuralism of Kripke and Putnam for its scientific naiveté has become something of an industry, but it is important not to throw the baby out with the bathwater. Is it really so difficult to construct more nuanced versions of microstructuralism that are adequate to the detailed science? I surveyed some of the options for doing so in Section 2.1. Is the macroscopic account of substances so clearly true that it should be adopted by default when microstructuralism encounters criticism? A more balanced discussion of how chemical substances are individuated requires that we weigh up the merits and demerits of all the available theories. That, in turn, requires that the macroscopic view be more fully articulated, as discussed in Section 2.2. But the exploration of these positions has only just begun, and many intriguing issues remain.

3. Molecular Reality
Molecular structures figure prominently in explanations of how and why chemical reactions happen, what their products are, and how much energy is released or absorbed in the process. Molecular spectra are explained as arising from bonded groups of atoms vibrating and rotating in interaction with radiation. Chemical and physical methods provide a wealth of information about particular types of chemical bond, and applications of quantum mechanics offer deep theoretical insights into the structure and bonding of molecules. Molecular structure is central to chemical explanation, and the chemical bond is central to molecular structure, because a structure is just an arrangement of atoms of specified kinds connected in a particular way by bonds.

3.1 The Chemical Bond
What, however, is a chemical bond? One obvious way to answer that question is to consider the explanatory role of the bond in chemistry. The classical
chemical bond is a creature of the structural theory that developed within organic chemistry in the nineteenth century (see Hendry 2010b). Once the elemental composition of compounds began to be analysed quantitatively and represented in formulae in the 1830s, chemists became aware that there could be distinct chemical compounds – isomers – which are composed of the same elements in the same proportions. Isomers were assumed to differ in the ways in which the elements are combined within them, that is, in their internal structure. Although in the first half of the nineteenth century there was no general agreement about what ‘structure’ might be (see Brock 1992, chapter 6), the 1850s saw general acceptance of the idea that atoms in molecules are linked to fixed numbers of other atoms. August Kekulé applied this idea, which came to be known as ‘valency’, to carbon, assigning a fixed tetravalency to carbon. By allowing it to link to other carbon atoms, he reduced the aliphatic hydrocarbons (methane, ethane, propane, etc.) to a homologous series (Rocke 1984, chapter 9). Kekulé later introduced double bonds and extended his treatment to aromatic compounds, producing the famous hexagonal structure for benzene. The characteristic visual representations of structure known as valence formulae emerged in the 1860s. They were interpreted cautiously at first, as representing only the ‘chemical structure’ of substances – the topology of the bonds between atoms – rather than the ‘physical’ positions of the atoms in space (see Hendry 2010b, section 1). Structural formulae became truly spatial in the 1870s, when they were employed in stereochemical theories that were intrinsically spatial, because their explanatory power depended precisely on their describing the arrangement of atoms in space. Thus, for instance, Jacobus van’t Hoff explained why there are two optical isomers of (for instance) lactic acid by assuming that four different functional groups A, B, D and E are arranged tetrahedrally around a single carbon atom (see Ramberg 2004, chapters 3 and 4).

If the explanatory role of the chemical bond was mapped out in the nineteenth century, twentieth-century chemists and physicists began trying to work out what realizes that role. G. N. Lewis (1923) provided the first widely accepted account: the electron-pair bond, which unified chemists’ understanding of bonding in polar substances (like sodium chloride) and non-polar substances (like methane). The electron-pair bond helped to create a whole new level of understanding of reaction mechanisms in organic chemistry during the 1920s and 1930s (see Goodwin 2007) and remains central to such explanations today (see Goodwin 2011). Lewis’s ideas also played an important part in early quantum chemistry: applying quantum mechanics to molecules turned out to be far beyond 1930s computing power, so semi-empirical bonding models were developed, such as the valence-bond approach championed by Linus Pauling, who regarded it as a synthesis of quantum mechanics with Lewis’ ideas (Hendry 2008). Although Pauling’s efforts
provided molecular quantum mechanics with some important early successes, such as an explanation of the tetrahedral structure of bonds around the carbon atom, the valence-bond approach was just one approximate scheme among others (the chief rival being molecular-orbital theory). Moreover, the chemical structure seemed to have been put into the quantum mechanics by hand, rather than derived from it. From the point of view of physics, Lewis’ localized pairs of electrons seem quaint. It was clear even in Lewis’ time that molecules must be dynamic entities, constantly in motion, and that the electrons within them cannot be static. Quantum mechanics only makes things worse: electrons are delocalized within molecules, and spend relatively little of their time between bonded atoms; moreover, they are fermions, which means that the electronic states of molecules must be anti-symmetric with respect to electron permutation. So individual bonds cannot be understood as composed of individual pairs of electrons. For all these reasons, some physicists and theoretical chemists have been sceptical about how far bonds are a genuine part of quantum-mechanical reality, even though they seem indispensible to chemical explanation (see Hendry 2004, 2008). Bonding is, of course, a real phenomenon: the question is whether understanding this phenomenon requires something like the structural chemical bond of the nineteenth century, and how far quantum mechanics is compatible with it (see Hendry 2008; Weisberg 2008). Should the structural conception of the bond be discarded in favour of something that is unproblematic at the level of the underlying physics? A more physically respectable conception of the bond might just identify it with the energetic stabilization of a molecular system, yet it is debatable whether this would be rich enough to capture the explanatory role of the bond in organic chemistry.

The chemical bond continues to be a matter for foundational debate in theoretical chemistry, its physical nature remaining elusive. In the ‘Atoms in Molecules’ programme, Richard Bader and his co-workers have sought to recover the traditional bond structure of molecules as a topological feature of the electron-density distribution (see Bader 1990; Gillespie and Popellier 2001). From the electron-density distributions for many different molecules, one can define ‘bond paths’ between atoms that generate ‘molecular graphs’ which are strikingly close to the classical molecular structures of those molecules. As Bader puts it, ‘The recovery of a chemical structure in terms of a property of the system’s charge density is a most remarkable and important result’ (1990, p. 33). The interest of Bader’s elegant results is twofold. First, if the scheme is extensionally adequate, finding bonds just where classical valence formulae would put them, this would suggest that it provides some real insight into the underlying physical nature of the chemical bond. Secondly, the fact that this account is based on electron density, which is undeniably a quantity of fundamental significance in quantum mechanics, suggests that the structural
conception of the bond can be regarded as quantum-mechanically rigorous after all.

But the correspondence between bond path and chemical bond is not perfect. The main problems concern repulsive (rather than attractive) interactions between neighbouring atoms in a molecule, and troublesome cases are provided by phenanthrene and the inclusion complex of helium in adamantane (see Haaland et al. 2004 and Poater et al. 2006). ‘Atoms in Molecules’ finds bond paths corresponding to these repulsive interactions, even though chemists would not normally regard the mutually repelling pairs of atoms as bonded to each other. One response to these cases is to defend these ‘bonds’ as bonds (see Bader 2006). This looks like bullet-biting, but too much of that will undermine the revisionary analysis one is seeking to defend. Does ‘Atoms in Molecules’ really identify the underlying physical nature of the classical chemical bond, or does it identify too many non-bonding interactions between atoms as bonds? Jerome Berson (2008) has argued that extending quantum-mechanical analysis to exotic chemical species, such as the ‘fleeting molecules and molecular fragments’ (2008, p. 956) that appear in the course of chemical reactions, challenges some long-held assumptions about chemical bonds. One such assumption is that the formation of a bond must always stabilize a molecule (2008, p. 952). Perhaps the subtle and complex physical interactions within these molecules show that ‘The bond concept allows us to understand much of chemistry, but not all of it’ (2008, p. 954). Interestingly, the assumption that is challenged by Berson’s cases appears to be just what underlies some of the criticisms of Bader’s programme.

3.2 Molecular Structure and the Symmetry Problem
Quantum chemistry is the interdisciplinary field that applies quantum mechanics to atoms and molecules in order to explain their structure and bonding. Textbooks often present the application as follows: we enumerate the electrons and atomic nuclei present in a molecule and identify the forces acting between them. This determines a Schrödinger equation for the molecule, whose solutions correspond to the quantum states it may have. Schrödinger equations for chemical systems typically consider electrostatic interactions only (Woolley 1976), which is a reasonable approximation, because the other interactions (gravitational, weak and strong nuclear forces) are orders of magnitude weaker at the molecular scale and can effectively be ignored when chemical behaviour is considered. There is an exact analytical solution to the non-relativistic Schrödinger equation for the hydrogen atom and other one-electron systems, but these cases are special, owing to their simplicity and symmetry properties. The Schrödinger equation for the next simplest atom, helium, cannot be solved analytically, and to solve the Schrödinger equations for more complex atoms,
or for any molecule, quantum chemists apply a battery of approximate methods and models which have become very sophisticated with the development of powerful digital computing.

Can such calculations be regarded as explaining molecular structure? One problem concerns isomerism. Ethanol (CH₃CH₂OH) and dimethyl ether (CH₃OCH₃) are different compounds with distinct molecular structures, but contain precisely the same electrons and nuclei. If the Schrödinger equation is determined only by the electrons and nuclei present, then the alcohol and the ether share a Schrödinger equation, and it is difficult to see how their structures could be recovered from it (see Woolley 1998). Symmetry properties pose a deeper problem: arbitrary solutions to exact Coulombic Schrödinger equations should be spherically symmetrical, but real molecules cannot be (see Woolley and Sutcliffe 1977, 2005; Hendry forthcoming). The problem is usually avoided by disregarding the exact Schrödinger equations in favour of lower-symmetry structures. This is motivated by the Born-Oppenheimer (or ‘clamped nucleus’) approximation, which separates the nuclear and electronic motions, then considers the massive slow-moving nuclei to be approximately at rest. Only the electrons are assumed to move quantum-mechanically, and while this makes little difference to the energy of the solution (see, for instance, Atkins 1986, p. 375), it does mean that the internuclear structure is assumed, rather than explained. Woolley points out that the clamping of nuclei cannot really be regarded as an approximation, because although it makes only a small difference to the calculated energy of a molecule, it makes a big difference to its symmetry properties. To give an example, chirality is a form of molecular asymmetry in which a molecule is not superimposable on its mirror image, because (for instance) a carbon atom is bonded to four different groups of atoms (which are arranged at the corners of a tetrahedron). Hence, chirality gives rise to a form of isomerism (the different forms are called enantiomers), and it has been known since the nineteenth century that in some cases the two enantiomers will rotate plane-polarized light in opposite directions, but by the same angle. If a determinate molecular structure is assumed (as, for instance, within the Born-Oppenheimer approximation), then it is possible to calculate the observed optical rotation angles. Exact solutions to the isolated molecule Hamiltonian, in contrast, will yield an optical rotation angle of zero. The symmetry problem is not specific to optical activity: asymmetries in molecular structures are essential to all kinds of explanation at the molecular level. It is worth emphasizing that Woolley’s symmetry problem has nothing to do with the insolubility of Schrödinger equations for molecules, or the computational complexity of numerical methods for solving them. According to Woolley, the problem is not that molecular structure is difficult to recover from the exact quantum mechanics, but that it is not there at all.
One obvious response to Woolley's argument is to point out that 'exact quantum mechanics' cannot just be the Schrödinger equation for the isolated molecule. In one sense, this is right; it should not, perhaps, be surprising if a Hamiltonian that was set up to describe an isolated molecule fails to apply in the bulk-matter environments that are of interest to chemistry: so much the worse for the isolated molecule Schrödinger equation as an account of molecular reality. On the other hand, there is evidence from molecular beam experiments that the isolated-molecule Hamiltonians are accurate in some rarefied contexts, which suggests that quantum-mechanical systems of nuclei and electrons can, in fact, be prepared in states in which they do not have determinate molecular structures. A determinate molecular structure is therefore something a quantum-mechanical system of nuclei and electrons may or may not have, depending on its interactions with its environment. For this reason, one might conclude that molecular structure is not really part of physical reality, because the physical states on which it depends are themselves dependent on 'our' interactions with it. However, Jeffry Ramsey (1997, 2000) points out that environment-dependence is not the same as observer- or mind-dependence, even if a molecule's physical interactions with its environment do sometimes constitute observations of it. Having clarified that issue, Ramsey does argue that the fact that molecular structure is a relational, and hence a non-intrinsic property of molecules 'provides good reasons for giving up certain kinds of strict realism and reductionism' (2000, p. 120), and also essentialism about molecular structure. It is hard to see why either essentialism or realism is challenged. If microstructural essentialism is the thesis according to which (for instance) being H₂O is what makes something water, then it could be true even if being H₂O involves being a system of one oxygen nucleus, two protons and ten electrons bonded in a certain way, and quantum-mechanical systems consisting of one oxygen nucleus, two protons and ten electrons can sometimes fail to have that particular structure. This would just entail that the relevant kind of system might not be water all of the time. Furthermore, it is not clear that the weaker ‘contextual’ realism that Ramsey advocates in place of ‘strict’ realism is anything more than the recognition that a molecule's state is dependent on its interaction with its environment. This seems compatible with any sane version of realism. The challenge to reductionism is clear, however, as we shall see in the next section.

Those who are comfortable with the idea of physics revising or even eliminating the central explanatory concepts of the special sciences might choose another response to Woolley's symmetry problem and say 'so much the worse for molecular structure.' Although physicists of the stature of Max Born have said similar things (see Hendry 2004), there is a vast amount of chemical and spectroscopic evidence that chemists currently explain by appeal to determinate molecular structures. It is easy to advocate radical revision of
existing explanations, but a lot harder to come up with the required revisions. Until there are new explanations of (for instance) carbon dioxide’s chemical and spectroscopic behaviour, or the complex mechanisms of organic reactions, explanations that do not appeal to determinate molecular structure, the call for explanatory revision is idle. A subtler version of this view is to regard molecular structures as artefactual, even while recognizing their utility (or even indispensability) to chemical explanation. For instance, according to Hans Primas (1975, 1983), quantum-mechanical holism means that it is an idealization to suppose that a molecule has a quantum state to call its own, a state which then interacts with its environment. All there is, from a quantum-mechanical point of view, is the state of the system-plus-environment (in effect, the universe as a whole). The quantum-mechanical properties normally associated with structure (determinate internuclear distances, moments of inertia) are not derived from an exact quantum-mechanical description. Rather, explanation in quantum chemistry involves the construction of model states that replicate the phenomenal patterns we read into a structureless quantum world (see Hendry 1998, pp. 127–30, for a more detailed account of Primas’ views).

4. The Reduction of Chemistry?

As long as the idea of reduction has existed, philosophers have tended to assume that chemistry is no more than physics. As we have seen, chemistry deals with substances, and also with chemical microspecies like molecules and their physical bases. Hence there are two layers to the question of reduction in chemistry. The first layer concerns whether chemical substances are nothing over and above their constituent microspecies. The next layer concerns the relationship between molecular structure and physics: are molecules just quantum-mechanical systems of charged particles? I will address the first question only to note that microstructuralism about chemical substances does not commit one to reductionism: even if substances are individuated by their microstructural properties, this does not entail that they are nothing over and above the bearers of these properties. The second reduction issue merits a longer discussion, because the symmetry problem about molecular structure, introduced in the last section, raises some important questions.

Classical intertheoretic reduction should, in principle, require the derivation of the properties of atoms and molecules from their Schrödinger equations. But quantum chemistry does not meet the strict demands of classical intertheoretic reduction, because its explanatory models bear only a loose relationship to exact atomic and molecular Schrödinger equations and are calibrated by information drawn from chemistry itself (see, for instance,
Bogaard 1981; Hofmann 1990; van Brakel 2000, chapter 5; Scerri 2007, chapters 8 and 9; Hendry forthcoming, chapter 7. In the case of atomic calculations, quantum-mechanical calculations assign electrons to one-electron orbitals that, to a first approximation, ignore interactions between electrons. Scerri (2007, chapters 8 and 9) argues that although the orbitals are artefacts of an approximation scheme, they seem to play an important role in explaining the structure of atomic electron shells, and the order in which they are filled is determined by chemical information, rather than fundamental theory (see Needham 1999 for a response). In practice, molecular calculations nearly all work within the Born-Oppenheimer approximation: the nuclei are constrained within empirically calibrated semi-classical structures, the electrons moving in the resultant field. As we have seen, this makes it difficult to see how molecular structure, which is so central to chemistry, can be said to have been explained.

Reductionists can make two responses here. The first is that the models are just ad hoc, but since these models provide much of the evidence for the explanatory importance of quantum mechanics in chemistry, this response would seem to undermine the motivation for reductionism. The second response is that inexact models are common in computationally complex parts of physics, and may not signal any deep explanatory failure. There is something in this response, but it requires that the atomic and molecular models that are used in explanations are justifiable as approximations to solutions of exact Schrödinger equations, and stand in for them in explanations of molecular properties (hence, call this the ‘proxy defence’ of inexact models). This is a more stringent condition than it may sound, requiring that the inexact models attribute no explanatorily relevant features to atoms or molecules that cannot be justified in the exact treatments. The Born-Oppenheimer approximation seems to offer a justification for the assumed semi-classical molecular structures, because the masses of atomic nuclei are thousands of times greater than those of electrons, and so move much more slowly. Fixing the positions of the nuclei makes little difference to the calculated energy, so in calculating the electronic motions the nuclei may be considered to be approximately at rest. But if Woolley’s symmetry argument is correct, the proxy defence must fail, because the symmetry properties of the Born-Oppenheimer models are explanatorily relevant, yet they cannot be justified in the exact treatments.

But the reduction issue does not begin and end with relationships between theories. Intertheoretic reduction of X to Y concerns whether or not a theory about X entails (or approximately entails) a theory about Y, or perhaps explains everything that the theory about Y explains. Ontological reduction concerns whether or not X itself is something ‘over and above’ Y in some robustly metaphysical sense that may be independent of any logical or explanatory relationships between current or future theories about X and Y. Robin Le Poidevin (2005) argues rightly that the unfeasibility of intertheoretic (or as he calls it,
‘epistemological’) reduction does not settle the issue of ontological reducibility. He attempts to identify just what could count as an argument for ontological reducibility of the chemical to the physical: chemical properties (like being a particular element) are not merely correlated with microphysical properties (like having atoms with a particular nuclear charge), they are exhausted by them. All possible instances of chemical properties are constituted by combinations of discretely varying physical properties. It is just not possible that between (say) helium and lithium there is another element. Now the physical explanation of the discreteness of the elements does seem interesting and important from an ontological point of view, but there are two lines of objection to a reductionist argument of the kind Le Poidevin envisages (see Hendry and Needham 2007). First, it addresses only the elements, but the elements do not exhaust the whole of chemistry. As we have seen, isomers are distinct substances that are identical in respect of their elemental composition, yet differ in respect of their molecular structures. Furthermore, molecular structure appears to be defined in terms of continuously varying quantities, like bond lengths and bond angles. Hence, the argument for reduction doesn’t seem to apply to the whole of chemistry, but only to elemental composition. Secondly, it is not clear just why the exhaustion of chemical properties by combinations of physical properties should be thought to establish the ontological reducibility of the chemical. Here’s why not. In recent philosophy of mind, ontological reducibility has been understood in terms of causal powers: A is ontologically reducible to B only in the case that the causal powers conferred by possession of A-properties are exhausted by those conferred by possession of B-properties (see Kim 1998, chapter 4). On this formulation, neither Le Poidevin’s combinatorial determination, nor microstructural supervenience is sufficient for ontological reduction, for the A-properties may confer ‘additional’ causal powers. If, for each cluster of B-properties corresponding to an A-property, there is a sui generis law of nature conferring distinct causal powers that are not conferred by more fundamental laws governing the B-properties, then the A-properties are irreducible to the B-properties in a robustly ontological sense.

Is this more than a mere logical possibility? The symmetry problem discussed earlier would seem to indicate that it is. For over a century, chemical explanations of the causal powers of molecules, and of the substances they compose, have appealed to molecular structures attributed on the basis of chemical and physical evidence. Yet the existence of such structures does not appear to have an explanation in the exact quantum mechanics of isolated systems of electrons and nuclei. If the explanation of molecular structure is that it arises from a quantum-mechanical system’s interactions with its environment, then the truth or falsity of the reductionist view of chemistry will depend on the precise nature of this explanation. To be an ontological
reductionist is to think that molecular structures arise from more fundamental laws, so the reductionist must expect that the ability of the environment to collapse the molecular symmetry is ultimately due to its falling under some more fundamental physical law. The opposing emergentist view is that for each molecular structure there is a *sui generis* law of nature that can be expressed in the language of quantum mechanics, but is an instance of no deeper physical law. To address the issue of the ontological reduction of chemistry is to assess the relative plausibility of these two interpretations, and how well they account for the explanatory relationship between physics and chemistry. The issue is not settled by the existence of quantum-mechanical explanations of molecular structure and bonding. Both reductionism and emergentism are compatible with there being such explanations, although they differ over their structure, and the degree to which the laws that appear in them are unified. It seems that the reduction of chemistry is more open than philosophers have assumed (see McLaughlin 1992 and Hendry forthcoming, chapters 9 and 10, for differing views).

5. Conclusion: Prospects for the Philosophy of Chemistry

For any science X, the autonomy of X’s subject matter is an obvious and engrossing issue for philosophers to address. Moreover, a careful study of chemistry and its theories promises to cast interesting new light on the issue of reductionism more generally, because if reduction to physics can proceed anywhere, it should do so in chemistry. So it is not surprising that, up to now, reduction has been a central topic in the philosophy of chemistry. Yet it is surely time to move on. If the sub-discipline is to find a wider audience, it must identify other disciplinary conversations – in philosophy, science and the history of science – from which it can learn and to which it can contribute. What are the candidates?

Chemical kinds are of clear interest to the philosophical discussions of natural kinds and classification that range across metaphysics, philosophy of language and the philosophy of science. Deeper chemical understanding would also introduce a more realistic atmosphere into philosophers’ discussions of theoretical identities like ‘water is H₂O’, an example which has been tortured in so many poor arguments for scientifically implausible conclusions in the philosophy of language and philosophy of mind. Other topics of relevance to philosophy of chemistry include structure and composition, causal mechanisms and the measurement problem in quantum mechanics (via the symmetry problem).

What about dialogue between chemistry and the philosophy of chemistry? Kinds and classification would not be a hopeful topic for discussion, because
classification and nomenclature have been dead science since the early twentieth century. The IUPAC systematic nomenclature proceeds exclusively in microstructural terms, and the normal science of classification is based on the assumptions of one side of a debate that philosophers would wish to regard as open. Reduction tends to be regarded by chemists as a philosopher’s issue, although in conversation with practitioners from different parts of chemistry, one finds a whole range of opinions on the autonomy of their science with respect to physics. A more promising subject is the quantum-mechanical basis of molecular structure, but most promising of all is the chemical bond. This is a notion with which every chemist is familiar, and one which is central to chemical explanation. Yet every theoretical chemist knows it to be a foundation puzzle, and as we have seen, its physical basis continues to be a topic for debate. Here philosophers might hope to clarify the arguments and make more explicit the opposed positions and their disagreements.

Note

1 I will have little to say about methodology, because I do not think that chemistry has its ‘own’ methods, although it does face its own distinctive versions of wider methodological issues. One such concerns visual reasoning, and its relationship to the kinds of explanation offered by physical theories which are expressed via mathematics (see, for instance, Woody 2000). Another concerns its distinctive interest in synthesis (van Brakel 2000; Schummer 1997). However, metaphysical and conceptual issues will dominate this survey.

References


—(2006), ‘Elements, compounds and other chemical kinds’, *Philosophy of Science*, 73, 864–75.


—. (2005), ‘Molecular structure calculations without clamping the nuclei’, Physical Chemistry, Chemical Physics, 7, 3664–76.

1. Introduction

For many philosophers of science, mathematics lies closer to logic than it does to the empirical sciences like physics, biology and economics. While this view may account for the relative neglect of the philosophy of mathematics by philosophers of science, it ignores at least two pressing questions about mathematics. First, do the similarities between mathematics and science support the view that mathematics is, after all, another science? Second, does the central role of mathematics in science shed any light on traditional philosophical debates about science, like scientific realism, the nature of explanation or reduction? When faced with these kinds of questions, many philosophers of science have little to say. Unfortunately, most philosophers of mathematics also fail to engage with questions about the relationship between mathematics and science, and so a peculiar isolation has emerged between philosophy of science and philosophy of mathematics. In this introductory survey I aim to equip the interested philosopher of science with a roadmap that can guide her through the often intimidating terrain of contemporary philosophy of mathematics. I hope that such a survey will make clear how fruitful a more sustained interaction between philosophy of science and philosophy of mathematics could be.

2. Historical Background: From Kant to Hilbert

The late nineteenth century and early twentieth century saw a sharp rise in philosophical discussions of mathematics. This can be plausibly traced to the continuing interest in Kant’s philosophy, along with the dramatic changes taking place in mathematics itself in the nineteenth century. Kant had argued that our knowledge of mathematics was grounded in our special access to the features of our mind responsible for our perception of objects in space and time. This appeal to pure intuition, as Kant called this special kind of representation, was meant to explain both the a priori and synthetic character of mathematical knowledge. Such knowledge was a priori because it could be justified independently of any appeal to perceptual experience. At the same time, it
extended our knowledge in a way that a mere analysis of our concepts did not, so such knowledge was synthetic. Kant presented mathematics as the clearest case of synthetic a priori knowledge, but also argued that synthetic a priori knowledge was at the basis of any scientific knowledge or, indeed, of any knowledge. Mathematics was crucial to defending Kant’s system of transcendental idealism according to which our knowledge is restricted to appearances whose features are due, in large part, to the nature of our mind. Later interpreters of Kant as well as his critics took pains, then, to account for mathematical knowledge and its relationship to the rest of our knowledge.

For our purposes, the most influential critic of Kant is John Stuart Mill. In his *System of Logic* (1843), Mill conceded that mathematics is synthetic or, as Mill would prefer to put it, involves real propositions. But Mill argued that mathematics is justified via ordinary perception and induction, and so it is empirical and not a priori. This empiricism about mathematics combined a claim about justification with a position on the subject-matter of mathematics. For Mill, a claim like ‘7+5=12’ is about physical regularities, such as the result of combining 7 ducks with 5 geese to get 12 birds. Similarly, a geometrical theorem such as ‘Every triangle has interior angles summing to 180 degrees’ is about ordinary physical objects whose shape is approximately triangular.

While these philosophical positions were being articulated and debated, mathematics itself was undergoing significant changes. One view of these developments is that they took mathematics in a more abstract direction, away from the traditional arithmetic and geometry at the heart of Kant’s and Mill’s discussions of mathematics. This abstraction countenanced new mathematical entities and techniques that extended mathematical knowledge in new directions. The most well-known examples of this trend are the development of calculus or analysis as well as non-Euclidean theories of geometry. The clarification of the central concepts of these new areas of mathematics such as function, derivative, limit and dimension occupied the attention of leading mathematicians like Weierstrass, Cauchy and Dedekind. Collectively referred to as the rigorization of analysis, this trend was incredibly successful in providing precise definitions of previously vague concepts and in pushing mathematical research in even more abstract directions.

Standing at the confluence of these two streams of philosophy and mathematics are Frege and Russell and their respective versions of logicism. In *Foundations of Arithmetic* (1884) Frege argued that arithmetic was analytic in the sense that all arithmetical concepts could be defined in purely logical terms and all arithmetical theorems could be proved by making only logical assumptions. For Frege, this project went hand in hand with the development of a new approach to logic. His earlier *Concept-script* (Begriffsschrift) (1879) provided the first system of polyadic predicate logic so that it became possible to logically articulate the difference between, for example, the true claim that
for every number there is a number greater than it and the false claim that there is a number greater than every number ($\forall x \exists y x < y$ vs. $\exists y \forall x x < y$). In *Foundations*, Frege focused most of his discussion on the natural numbers and presented several devastating objections to Kant’s and Mill’s philosophies of arithmetic. This cleared the ground for his own logicist proposals so effectively that neither approach to mathematics has had many contemporary defenders.

Frege’s proposed definition for the natural numbers can be best approached by first considering a definition that he deemed flawed. This is to define the numbers by saying that they obey the following principle, commonly known as Hume’s Principle:

$$(HP) \text{NxFx = NxGx if and only if } F \approx G.$$ 

In HP ‘Nx’ is a variable-binding operator that turns a concept expression like ‘Fx’ into a name for the number of Fs. The relation symbol ‘$\approx$’ stands for a relation between concepts that obtains just in case the objects that fall under the concepts can be correlated one-to-one. Let us call this the relation of being equinumerous. For example, suppose our concepts are $x$ is a knife on the table and $x$ is a fork on the table. Then HP says that the number of knives on the table is identical to the number of forks on the table just in case the knives can be exactly paired with the forks. Nobody would dispute the truth of HP for our pre-philosophical concept of number, but Frege offered clear reasons for thinking that it is not an acceptable definition of the concept of number. HP leaves open which objects the numbers are, so in particular, it does not settle the question of whether or not Caesar is identical with some number. Frege tried to solve this Caesar problem, as it is sometimes called, by picking specific objects that could be identified with the individual numbers in line with HP. This explicit definition deployed extensions of concepts, but we will put our discussion in terms of the set of things that fall under a concept. With this adjustment Frege’s proposed explicit definition is

$$(ED) \text{NxFx = } \bigwedge \neg G(G \approx F).$$

In words, this is the claim that the number of Fs is identical to the set of concepts that are equinumerous to the concept F. Using our earlier example, the number of knives on the table would then be a set that has as its members all the concepts that are equinumerous to the concept $x$ is a knife on the table. Supposing that there were 8 knives, then this set would include the concept $x$ is a year strictly between 2000 and 2009 and the concept $x$ is a planet in the Solar System.

To vindicate his logicism for arithmetic, Frege had to defend the claims that ED involves only logical concepts and that it was sufficient to prove all
Philosophy of Mathematics

arithmetic theorems using only logical principles. This is what he attempted in his Basic Laws of Arithmetic (vol. 1, 1893; vol. 2, 1903). The project failed because of Frege’s reliance on certain principles of set existence that, in the context of his strong logical system, rendered his theory inconsistent. The general problem is usually referred to as Russell’s Paradox, for it was Russell who communicated the problem to Frege in 1902 and who also sought to resolve it in a way consistent with a logicist philosophy of mathematics. To avoid the contradictions that ruined Frege’s approach, Russell developed his theory of types, which places strong restrictions on which sets exist. These restrictions, in turn, forced Russell to deploy debatable axioms for his logical system like the axiom of infinity. It states that there are infinitely many individuals, where an individual is an object that is not a set. Russell needed this assumption to prove many mathematical theorems, but it is hard to justify the view that the axiom of infinity is a logical truth.3

Russell’s paradox and similar foundational puzzles, along with the perceived failure of logicism, prompted an intense period of reflection on the foundations of mathematics, especially in the 1920s and 1930s. Following Frege and Russell, the participants in this debate seemed to assume that it was possible to give a comprehensive description of the nature of mathematics and that it was crucial to this project to explain how mathematics related to logic. The two main alternatives to logicism developed in this period are the intuitionism championed by Brouwer and the formalist programme advanced by Hilbert. Brouwer’s intuitionism involved a return to many aspects of Kant’s philosophy of mathematics, especially the focus on non-perceptual intuition and the constructive aspects of mathematics. Brouwer traced the foundational crisis to the failure to observe appropriate restrictions on principles like the law of excluded middle. In a proof by contradiction, for example, a classical mathematician may prove p by showing that ¬p entails a contradiction with antecedently known theorems. The intuitionist insists, however, that this inference assumes that p or ¬p is true, while there may be claims which have no determinate truth-value. Similarly, the intuitionist does not accept a proof from the premise that ¬∀xFx to the conclusion that ∃xFx. This is because the proof of an existence claim requires the production of the entity that witnesses the truth of the existence claim. Brouwer grounded these restrictions in his positive conception of the nature of mathematics based on a basic intuition of what he called ‘two-ity’. This is tied to our temporal experience. These aspects of Brouwer’s intuitionism proved difficult to motivate, and later philosophers of mathematics influenced by Brouwer have sought to ground restrictions on classical reasoning by other means. The most influential instance of this legacy is Michael Dummett’s criticisms of the use of the law of excluded middle in mathematics based on his views on meaning.4

The second main alternative to logicism in the foundational period is Hilbert’s formalism. Hilbert granted the intuitionist the worry that reasoning
about certain objects like sets of infinitely many things could be the source of the contradictions. But he sought to address this worry in a way that would preserve classical mathematics and its logical principles. Hilbert's strategy began by separating off a core body of real mathematics that was taken as the valid starting point for any further foundational investigations. A classical mathematical theory, for example, the theory of functions on the real numbers, could then be approached as the result of adding various symbols for 'ideal' mathematics to the original symbolism expressing the real mathematics. Hilbert argued that the meaning of these new symbols could be ignored, leaving only formal rules for their manipulation, if it was possible to give a proof that the new theory was consistent using only the sub-theory of real mathematics. This philosophical programme yielded a rich body of mathematical results in the form of proof theory. Here the mathematical symbolism was treated as a subject for mathematical study in its own right. The ultimate goal of these investigations was to show that the theories of classical mathematics lacked a configuration of signs among its theorems which would yield a contradiction, for example, '0=1'.

The viability of Hilbert-style formalism was called into question by Gödel's incompleteness theorems. The first incompleteness theorem applies to axiomatizable theories which include a certain minimal arithmetic, that is, a set of sentences T that are derivable from a set of axioms that include some minimal arithmetical claims Q. The theorem claims that if such a theory T is consistent, then it will be incomplete in the sense that there will be a sentence G in the language where neither G nor ¬G are derivable from the axioms of the theory. Gödel's second incompleteness theorem extended this result to show that for such theories, subject to some further fairly weak conditions, the sentences missing from T include an arithmetical claim Con(T) that is materially equivalent to the claim that the theory T is consistent. This convinced many that Hilbert's quest for a proof of the consistency of classical mathematics that would persuade the intuitionist to abandon her restrictions was impossible to fulfil. This is because a theory that was stronger than T was necessary to prove the consistency of T, and so any part of T restricted to real mathematics would not have the resources to prove the consistency of T. The case is not completely closed, however, and some philosophers of mathematics continue to champion more modest versions of Hilbert's formalism as an adequate philosophy of mathematics.

3. Current Context: From Platonism to Nominalism

The central role of sets in the foundations of mathematics prompted some mathematicians to develop an axiomatic theory for sets that seemed strong
Philosophy of Mathematics

enough to unify all of mathematics in a single framework. The most commonly discussed theory of this kind is ZFC (Zermelo-Fraenkel set theory with the axiom of choice). On this approach, certain truths about sets are taken for granted, and other mathematical entities, like the natural numbers, are identified with particular sets. If these axioms are adopted as the basis for mathematics, then it is quite tempting to view mathematics as the study of non-physical, abstract objects that exist outside of space and time. This is because there does not seem to be any physical interpretation of the axioms which renders them true. This proposal is often referred to as ‘set-theoretic platonism’, although the associations with Plato’s actual views are far from clear. In particular, few philosophers have been tempted to combine the view that the subject matter of mathematics is a domain of abstract objects with the epistemic claim that our knowledge of the axioms of mathematics is based on a special faculty of reason or intuition. A notable exception here is Gödel’s philosophy of mathematics as presented in ‘What is Cantor’s Continuum Problem?’ The Continuum Hypothesis (CH) is a claim involving sets that is easy to formulate, but which resisted proof or disproof using the axioms of ZFC. Eventually, it was shown that CH is logically independent of ZFC. Contemplating this possibility, Gödel insisted that there was, nevertheless, a fact of the matter concerning the truth of CH and that mathematicians should deploy various methods to try to determine whether or not CH was true. These methods include further reflection on the concept of set itself: ‘the very concept of set on which [the axioms of set theory] are based suggests their extension by new axioms which assert the existence of still further iterations of the operation “set of”’ (Benacerraf and Putnam 1983, p. 476).

Quine is the most influential proponent of a version of platonism for mathematics that tries to justify our mathematical knowledge empirically. This combination of views can be seen as a consequence of Quine’s naturalism. Quinean naturalism is the view that all knowledge is scientific knowledge and so the only form of justification of any purportedly non-scientific belief is its contribution to the overall success of science. For mathematics, this led Quine to offer what has come to be called an indispensability argument for platonism. This argument proceeds by noting the crucial role of mathematical theories in our best scientific theories. As these mathematical theories, when suitably regimented in the language of first-order logic, entail sentences like ‘(∃x)(x is a number)’, then our best science gives us empirical evidence for the existence of numbers. Quine rejects Mill’s view that the subject matter of mathematics is physical regularities. He insists that the best interpretation of mathematical language is in terms of sets and that sets are abstract objects (Quine 1981).

Quine’s views on mathematics set the agenda for much of the philosophy of mathematics in the 1960s and 1970s. The central issue for the discipline became
Quine’s claim that a platonistic interpretation of mathematical language in terms of sets was required or was in some way superior to a non-platonistic interpretation. Alternatives to Gödel’s and Quine’s respective versions of set-theoretic platonism received new life from two influential papers by Benacerraf. In ‘What numbers could not be’ (1965, reprinted in Benacerraf and Putnam 1983), Benacerraf questions the motivation for choosing to identify things like numbers with sets in this or that particular way. For example, the natural numbers 0, 1, 2, 3, . . . could either be identified with the series of sets $\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \ldots$ or with the series $\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \ldots$ (Here $\emptyset$ refers to the empty set). Benacerraf argues that there was no mathematical reason to prefer one identification over another. This complicates the set-theoretic reduction of numbers to sets, at least if it is conceived of as revealing what the natural numbers were all along. Benacerraf concludes from these problems that not only is there no fact of the matter whether or not numbers are identical with this or that series of sets, but also that numbers are not objects at all: ‘in giving the properties (that is, necessary and sufficient) of numbers you merely characterize an abstract structure – and the distinction lies in the fact that the “elements” of the structure have no properties other than those relating them to other “elements” of the same structure’ (Benacerraf and Putnam 1983, p. 291). This enigmatic formulation inspired a variety of structuralist interpretations of mathematics. The most ambitious version of structuralism is eliminative structuralism. It aims to eliminate all reference to abstract objects from the interpretation of mathematical language by conceiving a mathematical claim C as a tacit conditional of the form

$$(\forall x)(\forall y)(\ldots)(A^{(x, y, \ldots)} \rightarrow C^{(x, y, \ldots)}).$$

Here the axioms for the domain in question ‘A’ and the original claim ‘C’ have been amended so that, as the superscripts indicate, all the constants and predicate symbols are replaced by variables $x, y, \ldots$. The resulting generally quantified conditional claim says that for any domain satisfying the axioms, the claim will hold for that domain. The eliminative structuralist tries to recast all ordinary mathematical claims in this way so that their truth does not require the existence of any abstract objects. This undercuts set-theoretical platonism, but only at the cost of ascribing a non-standard logical form to the statements of mathematics.

A further objection to eliminative structuralism motivates the platonistic or ante rem structuralism of Resnik and Shapiro (Resnik 1997, Shapiro 1997). Consider a purported domain that has no concrete instances such as Zermelo-Frankel set theory. Here the only non-logical symbol is the sign for the membership relation ‘$\in$’. So, the eliminative structuralist suggests interpreting the result R of ZFC as $(\forall x)(ZFC^{(x)} \rightarrow R^{(x)})$. But given that there are no concrete
domains that satisfy ZFC\(^{(n)}\), it follows that this conditional is vacuously true and so \((\forall x)(ZFC^{(n)} \rightarrow \neg R^{(n)})\) will come out to be true as well. This deprives the eliminative structuralist of an adequate interpretation of mathematics as it fails to preserve the truth-value of the original mathematical claims. The ante rem structuralist responds by positing abstract structures that will satisfy the axioms of the traditional theories of mathematics, like number theory and set theory. This does not mark a return to set-theoretic platonism, though, because these structures are conceived of on the model of structured universals, which are prior to the positions which play the role of traditional mathematical objects.

Ante rem structuralism is similar to traditional platonism to the extent that it faces a serious epistemic objection. In another influential paper ‘Mathematical truth’ (1973, reprinted in Benacerraf and Putnam 1983), Benacerraf argued that a literal or standard interpretation of mathematics is strongly supported by the grammatical role of terms for mathematical objects like ‘five’ and ‘seven’. For example, the claim ‘There are at least three perfect numbers greater than 17’ seems to be of the same logical form as ‘There are at least three large cities older than New York.’ But then ‘17’ is a name for an object just like ‘New York’. However, Benacerraf continued, if we adopt this standard interpretation, then we have difficulty explaining how we can refer to such objects and how we can know any truths about them or even that they exist. Benacerraf originally pressed this point using the then fashionable causal theory of reference and causal theory of knowledge. But the worry remains in force for many pictures of reference and justification. As Field summarized the worry, the epistemic challenge ‘is to provide an account of the mechanisms that explain how our beliefs about these remote entities can so well reflect the facts about them . . . if it appears in principle impossible to explain this, then that tends to undermine the belief in mathematical entities, despite whatever reason we might have for believing in them’ (Field 1991, p. 26).

The theme set by ‘Mathematical truth’ was played out in many ways in subsequent philosophy of mathematics with philosophers finding ingenious strategies either to account for our knowledge of abstract objects or else to recast the logical form of mathematical statements so that mathematical knowledge could be assimilated to knowledge of logic and ordinary scientific knowledge. The most vigorous strand of the former strategy is neo-Fregeanism as developed by Hale and Wright, among others. Hale and Wright argue that Frege’s HP definition was adequate after all and that Frege was mistaken in thinking that the further ED definition was needed. Neo-Fregeanism received significant support from a result known as Frege’s Theorem. This is the fact that the theory of arithmetic given in second-order logic with HP as its only non-logical axiom, called Frege Arithmetic (FA), is sufficient to derive Peano Arithmetic (PA) (Boolos 1999). PA is typically thought to represent our knowledge of the natural numbers. Furthermore, FA can be shown to be consistent just in case
PA is consistent. The main epistemic advantage of FA over PA is that HP is treated as a definition, and so is supposed to be easier to justify than alternative axiomatizations of arithmetic. The success of this approach for the natural numbers has led to a sustained investigation of other principles like HP for other areas of mathematics, such as the real numbers and set theory (Fine 2002, Burgess 2005).

The alternative strategy that retains its popularity is nominalism. Strictly speaking, nominalism for mathematics is just the claim that mathematics does not motivate us to accept the existence of any abstract objects. One version of nominalism achieves this result by adopting a non-standard interpretation of the logical form of mathematical statements. For example, Chihara treats mathematical statements in terms of the constructability of certain linguistic items (Chihara 1990), while Lewis offered an interpretation of these statements using the mereological apparatus of part and whole (Lewis 1993). Building on the eliminative structuralist proposal, Hellman considered taking mathematical claims to involve only the possibility of a certain kind of structure (Hellman 1989). All three stop short of countenancing abstract objects, but aim to assign objective truth-values to the right mathematical statements. They thus reject both Benacerraf’s claim that mathematical statements should be given a logical form that matches ordinary statements and Quine’s argument that the role of mathematics in our scientific theories requires a metaphysics of abstract objects. A second variety of nominalism argues that mathematics is not, after all, essential to the success of our best scientific theories. This kind of nominalist presents non-mathematical versions of these best theories and uses these versions to determine what we should believe exists. Here Field’s proposal is the most developed, but we will defer a discussion of it until Section 5. Finally, a more recent version of nominalism is typically dubbed ‘fictionalism’. This is the view that even though mathematical statements have a literal content that accords with their standard logical form, they also have a non-literal or fictional content that we can use to fix our ontological commitments. Different fictionalists obtain their fictional contents in different ways. For example, Balaguer focuses on the causal isolation of abstract objects from the physical world (Balaguer 1998), while Yablo speaks more metaphorically of ‘the real-world condition that makes it sayable that $S$’ (Yablo 2002, p. 229).

Despite these innovations, it is fair to say that many philosophers of mathematics would agree with Burgess’ verdict in ‘Why I am not a nominalist’. Combining Quine’s test for ontological commitment with Benacerraf’s standard interpretation of the logical form of mathematical statements, Burgess ridicules philosophers who would impose philosophical standards to adjust the successful practice of scientists and mathematicians (Burgess 1983, p. 98). While the platonist-nominalist debate has reached a sort of standoff, new innovations continue to inject new life into these traditional positions.
The appearance of Parsons' *Mathematical Thought and Its Objects* is perhaps the best example of this in some time (Parsons 2008). Parsons combines aspects of structuralism with a novel account of our intuitive access to what he calls quasi-concrete objects to try to overcome the persistent problems with a platonist interpretation of mathematical language.

4. Naturalism and Practice

The debates between platonism and nominalism continued through the 1980s and 1990s, but by this time the philosophy of mathematics had come to seem to many to be too disconnected from the mathematics studied in most mathematics courses and pursued by most mathematicians in their research. Unlike the time of the original foundational crisis, these critics argued, philosophy of mathematics had drifted away from mathematics and risked losing track of ongoing developments within mathematics itself. A major strand of this ‘maverick’ tradition appealed to the range of topics that were being pursued in the philosophy of science as an inspiration or even a model that could be imitated by the philosophy of mathematics. Lakatos's *Proofs and Refutations* remains one of the earliest and best examples of this trend (Lakatos 1976). Written in a lively dialogue style, Lakatos offered a kind of rational reconstruction of developments from the history of mathematics centred around the theorem that for any polyhedron, the number of vertices minus the number of edges plus the number of faces is identical to two. A central lesson of the dialogue, though, is that the different proofs offered for this theorem serve as much to clarify the central notions like polyhedron as they do to place some determinate theorem beyond doubt. Also, counter-examples to previous versions of the theorem make a positive contribution to the development of new theorems. In the philosophy of science, Lakatos aimed to reconstruct scientific knowledge using his theory of progressive research programmes and the similarities between the mathematical and scientific cases are significant. In both cases, our knowledge is the product of a historical process of scrutiny and innovation.

Lakatos's call to focus on mathematical practice and the history of mathematics was taken up with enthusiasm by Philip Kitcher in his *The Nature of Mathematical Knowledge* (1984) and the collection co-edited with Aspray, *History and Philosophy of Modern Mathematics* (1988). On the one hand, Kitcher engaged with the foundational tradition by presenting a thoroughly empiricist interpretation of mathematics inspired by Mill. On this view, mathematics is the study of the operations of an ideal agent stripped of many of the limitations of ordinary human agents. Different mathematical theories correspond, then, to different conceptions of the abilities of this ideal agent. On the other
hand, Kitcher unleashed a polemic against the methods typically employed by most philosophers working in the foundational tradition. Writing with Aspray, Kitcher identifies a ‘minority tradition’ of philosophers, including Lakatos, who ‘share the view that philosophy of mathematics ought to concern itself with the kinds of issues that occupy those who study other branches of human knowledge (most obviously the natural sciences)’ (Aspray and Kitcher 1988, p. 17). This historical turn proved of more lasting significance than Kitcher’s empiricist interpretation of mathematics. For while this interpretation seemed to be just one more attempt to give a non-standard nominalist spin on mathematics, the methodological turn that Kitcher argued for linked the philosophy of mathematics to the history of mathematics and so reopened many philosophical questions about mathematics that had been overlooked for some time. A similar call for change can be found in the collection edited by Tymoczko (1986/1998). Here, a range of different approaches to philosophy of mathematics are presented, each of which places an emphasis on the similarity between scientific methods of investigation and more recent developments in mathematics, as with the use of computers to prove the four-color theorem.

More recently, Maddy has pushed this turn to practice in new and interesting directions. She has developed her views under the aegis of a revised form of Quinean naturalism, beginning in her earlier work (1997) and continuing with her more recent study (2007). Maddy retains the traditional focus on set theory, but she criticizes Quine’s view that scientific knowledge is to be the standard against which other areas of knowledge should be judged. Instead, she presents a more thoroughgoing naturalism which accords mathematical practice a default valid status. Maddy links this shift to a change in the sorts of questions that the philosopher of mathematics should be asking. Rather than trying to incorporate mathematics into our overall metaphysics and epistemology, philosophers should instead aim to articulate and clarify the standards internal to mathematical practice. For example, on the issue of which axioms should be added to ZFC to resolve CH (see Section 3), Maddy argues that philosophers should focus on the principles that set theorists themselves seem to be employing when they debate this issue. As a result, Maddy aims to reform philosophy of mathematics so that it is more descriptive in orientation (see also Corfield 2003).

The most thorough presentation of this new approach to the philosophy of mathematics is Mancosu’s edited collection (2008). As he makes clear in his introduction, Mancosu aims to enrich the philosophy of mathematics by combining ‘local studies’ on specific issues with a more global attempt to integrate these case studies into an epistemology and metaphysics for mathematics as a whole (Mancosu 2008, p. 19). The focus on practice can be used to serve the same aims as the traditional foundational approaches, but the tools used to achieve these aims are completely different. Unlike the top-down foundational
approach, which focuses almost exclusively on set theory and its interpretation, and which seems to assume that all of mathematics works in the same way, the new bottom-up practice approach takes the variety of mathematics across sub-disciplines as a starting point and uses this to construct a more adequate metaphysics and epistemology for mathematics. The parallels with earlier developments in the philosophy of science are striking. For while mid-twentieth-century philosophy of science had typically assumed that philosophy should study how science works in general, contemporary philosophy of science has developed into a series of different sub-disciplines focused more on the specific sciences like physics, biology, economics and so forth than on any overarching general lessons for how science as a whole functions. But few argue that this should displace traditional topics like realism, reduction or explanation. Instead, one would hope that these topics would now be pursued in a more enlightened and informed way.

The turn to practice indicated by Mancosu’s volume has similar ambitions. This comes through clearly in Manders’ discussion of the role of diagrams in Euclidean geometry. His aim is ‘to account for the justificatory success of diagram-based geometry’ (Mancosu 2008, p. 67) on its own terms. Manders’ reconstruction of the value of diagrams tries to explain how the Euclidean geometrical tradition could be so stable for so long. The broader implications of this reconstruction are not drawn in his contribution to the volume, but it would provide a valuable input into any attempt to understand why mathematicians might have abandoned Euclid’s methods or what epistemic shifts might arise across these different mathematical traditions. Similarly, in his own contribution with Hafner, Mancosu explores the status of explanation within mathematics. Mathematicians commonly distinguish proofs that are explanatory from those that are not, but it remains difficult to characterize the mathematical value of the pursuit of explanatory proofs. Hafner and Mancosu draw on a case study of the mathematician Brumfiel’s work on real closed fields. Brumfiel explicitly indicates that he adopts a certain method of proof, even though another method is available that would provide a more unified method of proof. This case is then used to undermine Kitcher’s proposal that unification is the key to explanation not only in science, but also in mathematics. While this shows that unification is not always the source of explanatory power in mathematics, Hafner and Mancosu remain open to the possibility that more than one account of explanation may be necessary to explain how mathematicians do mathematics.

This rapprochement between the topics and priorities of philosophy of mathematics and philosophy of science naturally raises the more general question of how mathematics is both like and unlike a science. As we have already seen, there are general proposals, like Mill’s, Quine’s and Kitcher’s, that align mathematics very closely with empirical science, either in its subject
matter, methods of justification or both. What all these views have in common, though, is a conviction that mathematics and science as a whole can be characterized in uniform terms so that it makes sense to ask how the two disciplines are alike. By contrast, the turn to practice in the philosophy of mathematics and the philosophy of science suggests that we look for more local similarities and differences between sub-disciplines, or even more narrowly to particular historical episodes or contemporary methodological debates. This approach promises to deliver a more realistic picture of the links between this or that part of mathematics and science and should help us to improve our understanding of the remarkable success of both fields.

5. The Role of Mathematics in Science

There is an additional point of contact between the philosophy of mathematics and the philosophy of science having to do with the role of mathematics in science. The contribution that mathematics makes to science is difficult to summarize, and one suspects that mathematics might contribute in different ways to different parts of the scientific enterprise. Setting this complexity aside, both philosophers of mathematics and philosophers of science have sought to draw substantial conclusions from the central place of mathematics in nearly all contemporary successful science. After reviewing the respective debates on the mathematics and science sides, I will conclude by suggesting an additional, hopefully fruitful approach that would involve the expertise of both philosophical communities.

On the philosophy of mathematics side, most discussion has been concerned with Quine’s indispensability argument for platonism. This argument received an influential formulation by Putnam (1979), but the most careful presentation is by Colyvan (2001). As Colyvan presents the argument, it has only two premises:

1. We ought to have ontological commitment to all and only those entities that are indispensable to our best scientific theories;
2. Mathematical entities are indispensable to our best scientific theories.

Therefore:

3. We ought to have ontological commitment to mathematical entities (Colyvan 2001, p. 11).

The first premise is supported by Quinean naturalism and another claim that Quine argued for, namely confirmational holism. This is the view that the
success of a scientific theory leads to the confirmation of all of its sentences. Holism of this sort rules out any kind of selective confirmation that would assign differential confirmation to the mathematical and non-mathematical parts of the theory. This point has been questioned by Azzouni who argues that ‘posits’ in science are treated differently, depending on our epistemic access to them (Azzouni 2004). While other aspects of Quine’s naturalism have been disputed (Maddy 1997), by far the most effort has been expended on the second premise. Notice that many nominalists can concede that mathematical language is indispensable to the presentation of our best scientific theories, while still rejecting premise 2’s claim that mathematical entities are themselves required. This is the response suggested by Chihara’s, Lewis’ and Hellman’s different nominalist interpretations of mathematical language. Quine’s response to these sorts of proposals is that they make use of non-logical resources like an appeal to modality that is illegitimate in the context of a proper regimentation of the language of science. This strategy is not very attractive for contemporary philosophers, who reject many of Quine’s views, and Colyvan has at times conceded the point that the conclusion of the indispensability argument does not require a platonist interpretation of mathematics (Colyvan 2001, section 7.1; Pincock 2004).

A more aggressive reply to premise 2 is offered by Hartry Field in his (1980). Field argues that mathematical language is only pragmatically indispensable to reasoning with our best scientific theories, but that there are non-mathematical formulations of these theories which we can use to determine our ontological commitments. Field’s attempt to develop a non-mathematical version of Newtonian gravitational physics provoked an extended debate on the adequacy of such a formulation and the border between the mathematical and non-mathematical. He proceeded by presenting an axiomatic theory whose intended domain included only space-time points and regions. Scalar magnitudes like temperature or gravitational potential were then assigned to these points using physical relations like being greater in magnitude. Given certain axioms for these relations and the background space-time, Field then proved what is known as a representation theorem to the effect that these physical properties and relations can be accurately represented by assignments of real numbers to space-time points and regions. Field further claimed that he could prove that the original mathematical theory M+T was a conservative extension of his non-mathematical theory T. That is, for any sentence s that is entailed by the original theory M+T and that is formulated entirely in non-mathematical terms, s will also be entailed by T. This led Field to present T as a non-mathematical version of M+T.

Field’s programme was criticized in a number of directions (MacBride 1999). Some of the most important objections concerned the extension of Field’s strategy to other sorts of scientific theories (Malament 1982), the
acceptability of the logical resources deployed by Field (Shapiro 1983) and
the nominalistic credentials of Field’s background space-time (Resnik 1985).
More recently, Pincock has argued that the physical assumptions of these non-
mathematical theories are unwarranted by the evidence that is adequate to
support the original mathematical theories (Pincock 2007). Still, Field’s
programme prompted extended investigations into the prospects for non-
mathematical scientific theories and offered indirect illumination of the posi-
tive role that mathematics has in science. The most thorough investigation
of the possibilities for nominalizing science can be found in Burgess and Rosen
(1997).

In his own response to Field, Colyvan emphasizes the scientific benefits
that traditional mathematical scientific theories tend to have over their pur-
tportedly non-mathematical rivals. These include the power to unify scientific
theories (section 4.4) as well as an increase in explanatory power (section 3.3).
The examples in the book prompted an exchange with Melia, which has since
expanded into an extended discussion (see Melia 2000, Colyvan 2002, Baker
2005, Bangu 2008 and Baker 2009). The platonists in this debate seek to
provide a positive explanatory argument for premise 2 in the original indis-
ponsibility argument, which parallels in many ways the explanatory
‘no miracles’ argument championed by scientific realists like Psillos (Psillos
1999). For example, Baker has highlighted the scientific explanation of the
primeness of the length in years of the life cycle of some species of cicada to
support such an explanatory indispensability argument. But as with anti-real-
ists about science, anti-platonists have several responses. Bangu, for one, has
complained that such an argument begs the question in favour of platonism
because the explanandum is already given in a mathematical form. The issue
has recently been taken up by Psillos himself, who exploits the role of mixed
abstract–concrete objects like the Equator and models like the linear harmonic
oscillator to conclude that scientific realists should also be platonists to some
degree (Psillos unpublished). Given that the realist needs some abstract
objects, it is hard to see why mathematical objects are so objectionable.

Most philosophers of science have shown little interest in the indispens-
ability argument for platonism, but ongoing debates about modelling and ide-
alization have placed the role of mathematics in science on the philosophy of
science agenda. A central issue in discussions of models is the relationship
between models and theories and the metaphysical nature of models them-
selel. The semantic view of theories offered clear answers to these questions
by identifying theories with collections of models that satisfied a certain range
of descriptions and by identifying models with sets of the sort used in logic to
interpret formal languages. On this approach, a model represents a situation
when there is a certain kind of structural correspondence between the model
and the situation. The traditional advocates of the semantic view seemed to

328
downplay the distinctively mathematical nature of the descriptions that were used to pick out their models. Here they followed much of the foundational tradition in the philosophy of mathematics in assuming that first-order logic and set theory were sufficient to characterize scientific models and their representational relationship to the world.

Many of the assumptions of the semantic view of theories were challenged by an alternative tradition which advocates the need for ‘mediating models’ which would stand in between the abstract theories picked out by mathematical descriptions and the concrete physical situations which these theories purport to represent. The most influential contributor to this approach is Nancy Cartwright. Many of her assumptions about mathematics are reflected in the early paper ‘Fitting facts to equations’. As the title suggests, Cartwright conceives concrete situations in a thoroughly non-mathematical way. It is only through a considerable distortion and selective focus that we can achieve any kind of structural correspondence between mathematical descriptions and actual situations. As Cartwright puts it, her ‘basic view is that fundamental equations do not govern objects in reality; they govern only objects in models’ (Cartwright 1983, p. 129). The objects in models are artificially devised through idealization and abstraction so that they can mediate between the abstract mathematics and the concrete physical world. This has far-reaching implications for Cartwright for the proper understanding of what scientific knowledge tells us about the world, especially concerning the limited scope of the fundamental laws of our best science (see also Morgan and Morrison 1999).

An important response to Cartwright was offered by McMullin in his ‘Galilean idealization’ (1985). McMullin argued that since Galileo, the scientific realist has had a clear strategy for reconciling the sort of gap between abstract, mathematical scientific theories and the concrete, messy physical world. This is to posit a series of corrections between the highly idealized mathematical representation and a fully realistic description (McMullin 1985, p. 261). While no doubt a productive response to Cartwright’s pessimism about bridging the gap between the mathematical domain and the concrete systems of the physical world, McMullin’s realist correction strategy left several open questions that have been pursued in more recent work on modelling and idealization. To start, McMullin gave no general assurances that this sort of correction was always possible and left little for the realist to say in those cases where it seems de-idealization is impossible (Butterman 2009). A deeper issue concerns the need to distinguish between the semantic question of whether or not a mathematical representation is about a target system from the epistemic question of the respects in which the mathematical representation is an accurate representation of the target system. McMullin gave the realist a means to address the epistemic worry, but had little to say against Cartwright’s and others semantic worries (Suárez 2003).
A more recent proposal by Bueno and Colyvan aims to address these semantic worries (Bueno and Colyvan forthcoming; see also da Costa and French 2003). They argue that the attempt to understand representation using a single stage of structural correspondence is too rigid to account for the inferences that scientists carry out using their scientific models. Instead, they offer a three-stage process of immersion, derivation and interpretation. In the immersion step, an aspect of the target system is associated with some mathematical structure. Then in the derivation step certain consequences of this association are drawn. Finally, in the interpretation step the resulting mathematical claims are related back to predicted features of the target system. The new flexibility offered by Bueno and Colyvan’s framework is sure to provoke continuing refinements from advocates of a broadly semantic view of theories and a new round of objections by their critics from the mediating models tradition.

We have seen how philosophers of mathematics are most concerned with questions about the metaphysical interpretation of mathematics, while philosophers of science tend to focus on the status of scientific models, representation and the realism debate. A new way forward is suggested by the success of the local studies of the success of science that are the shared data for the positions of the semantic theory tradition, the mediating models tradition as well as both realists and anti-realists. What is needed is a survey of the different ways in which mathematics has contributed to the success of science. These contributions may turn on issues of representation and structural correspondence, or they may vindicate Cartwright’s conception of the gap between the mathematical world and the physical world. Whatever the results, it seems clear that this sort of investigation would have broad implications for all of the debates covered in this survey. Carrying it out successfully would require the skills and background of both philosophers of mathematics and philosophers of science and might provide the motivation for further collaboration and philosophical innovation.

Acknowledgements

For comments on an earlier draft I would like to thank Jody Azzouni, Sorin Bangu, James Hawthorne, Paolo Mancosu, Stathis Psillos, Bryan W. Roberts and Esther Rosario.

Notes

1 See Shapiro (2000, chapter 4) for discussion of Kant and Mill.
2 An accessible summary of this period in the history of mathematics is Kline (1972, especially chapter 43).
4 See Posy’s and McCarty’s contributions to Shapiro (2005) for more discussion of intuitionism.
6 See Hale and Wright’s contribution to Shapiro (2005).
7 See also Laymon (1995). A recent classification of approaches to idealization is offered by Weisberg (2007).

References


—(unpublished), ‘Scientific realism: between Platonism and nominalism’.


This page intentionally left blank
Part III

Past and Future
1. Introduction

This chapter discusses some emerging trends, new directions and outstanding issues in philosophy of science. The next section places contemporary philosophy of science in context by considering its relationship to analytic philosophy at large, to the history of science and to science itself. Section 3 then takes a look at a selection of interesting trends emerging from current research and some important issues calling for further work. The presentation is inevitably coloured by our personal perspectives, and we cannot hope to do justice to even the majority of new developments. But we hope to be able to convey some sense of the range of problems, issues and areas in which new and interesting research is being conducted.

2. Philosophy of Science in Context

What does the future hold for the philosophy of science? Certainly the field overall appears to be in rude health, as we hope the essays in this volume testify. At the ‘global’ level, its changing and maturing relationships with both the history of science and philosophy – with epistemology and metaphysics in particular – offer the possibility of new perspectives from which to not only view but perhaps contribute to science itself. In its analyses of particular features of science – the nature of evidence, the role of explanation, the possibility of reduction and so on – new resources are being brought into play and fresh approaches delineated.

2.1 The Place of Philosophy of Science in Analytic Philosophy

Philosophy of science is central to many areas of contemporary analytic philosophy, due to the broadly naturalistic inclination of the latter. In as far as naturalistic philosophers see philosophy as ‘continuous with science’ and as
being informed – to some extent, at least – by our best science, it is the philosophy of science that naturally assumes a mediating role between science and other sub-disciplines of philosophy. Such a mediating role can be fulfilled in many ways – as illustrated through examples below – and it need not be a one-way process: philosophical analyses of a more general sort can feed into philosophy of science and, ultimately, even into science itself.

Many current philosophers of science are taking the mediating role seriously, and rightly so. Take, for example, the concept of ‘natural kind’, central to numerous philosophical debates ranging from the philosophy of language to metaphysics. Several people have recently argued that the simple essentialist concept of ‘natural kind’, originating from Kripke and Putnam and still forming a staple of many a traditional debate, is problematic in the face of actual science. Regarding paradigmatic chemical kinds, for example, some have argued that there’s an element of stipulation in specifying what a natural kind name like ‘water’ refers to, and others have argued that water, say, cannot be identified on the basis of some microstructural essence, as being simply H₂O. (For recent discussion on these issues, see Beebee and Sabbarton-Leary 2010.) Brigandt, in his contribution to the present volume, discusses related issues in the context of biology, and Hendry covers them in the context of chemistry. Both authors argue that the relevant scientific detail should be brought to bear on general philosophical debates, in as far as the concept of ‘natural kind’ is to be grounded at all in the way that science actually classifies the world.

‘Natural kind’ is but one example of a central philosophical concept that can potentially be sharpened by scientifically informed philosophy of science. Another example is that of ‘cause’. To give an illustration of how much in analytic philosophy depends on our understanding of the concept of causation, consider the following. Many philosophical positions (in the philosophy of mind, for example) accept ‘Alexander’s dictum’, according to which (qualitative) properties are real only if they are ‘causally powerful’. Premised on this assumption, and armed with various kinds of conceptual analyses of causation – grounded mainly on armchair intuitions – philosophers of mind have debated for years about the reality of mental states and mental causation. The challenge has been to avoid the so-called causal exclusion problem, according to which the higher-level mental properties are causally otiose, and therefore epiphenomenal. Recently, some very profitable moves have been made in the debate by approaching the concept of causation from a more science-driven perspective: the counterfactual ‘interventionist’ theory of causation, due to Woodward (and others), is inspired by the actual structure of causal explanation in physics and in randomized experiments, for example. According to Woodward (and others who follow similar approaches to causation) much of the intuitions-based literature on the topic fails to analyse the nature of causation and causal explanation correctly ‘from the armchair’, and
arguably their science-driven account – motivated by very different problems at the core of philosophy of science – provides a natural solution. (For recent discussion, see Hohwy and Kallestrup 2008.)

The debate about the reality of the mental is largely driven by the adoption of some kind of physicalism, according to which fundamental physics describes causal facts of the world at ‘the bottom level’ on which all the other facts supervene. In connection with this idea, it is very important to note that according to one rather influential line of thought in philosophy of science – drawing on modern physics, in particular – there is no objective, wholly mind-independent causation to be found in fundamental physics at all! (See, for example, Corry and Price 2007.) This causal anti-fundamentalism is a subject of ongoing debate, of course, but given that it is not at all clear that fundamental physics and its properties are naturally understood in causal terms, one ought to wonder where this possibility leaves the mental causation debate thus presented, and ‘Alexander’s Dictum’ to boot! It is quite possible that the reality criterion for (qualitative) properties at the fundamental level is not best construed in causal terms at all. (We do not want to suggest that only metaphysicians and philosophers of mind outside contemporary philosophy of science have ignored such potential ‘complications’ arising from actual physics. Within philosophy of science itself, there are many programmes of research that similarly turn a blind eye to these matters. The recent revival of neo-Aristotelianism in ‘metaphysics of science’ is a case in point: there has been very little discussion so far of how a metaphysics based on fundamental causal powers ties in with the actual physics and the role of symmetries and conservation laws, for example, in identifying the fundamental properties in the world. From the point of view of philosophy of physics, dispositional essentialism is very much ‘armchair’ philosophy of science in its spirit.)

Here’s another example of a notion central to analytic philosophy: the concept of ‘concept’ itself has recently attracted a similar kind of science-influenced attention. Wilson (2006) provides a book-length argument, drawing heavily on engineering and applied mathematics in particular, against (what he calls) the ‘classical picture of concepts’. Even if Wilson’s unorthodox ‘patchwork’ picture doesn’t get purchase among philosophers of mind and language, the latter are likely to benefit from this extremely rich and detailed discussion explicitly inspired by subtle workings of concepts in actual science. A very different perspective on concepts is offered by Machery’s approach (2009), which derives its eliminativism about concepts from psychology research and its historical development. Again, even if Machery’s own position fails to convert philosophers who approach concepts from a more a priori perspective, his work is significant in uncovering important complexity and detail in the rapidly evolving scientific disciplines that ultimately must inform any philosophical analysis of ‘concept’.
Machery’s work draws on psychology, including neuropsychology. The rapid development of neuroscience has given rise to philosophy of neuroscience, a relatively new field of philosophy of science. Many classic issues in the philosophy of mind get discussed in a different light in philosophy of neuroscience. Take the debates around the notion of ‘multiple realizability’ of psychological states, for example. While the original arguments for multiple realizability were based on thought experiments and conceivability claims, the current debate relies much more heavily on the actual neuroscientific findings and the methodological nature of the discipline itself. (See, for example, Wilson and Craver 2006.) And it is not only philosophy of neuroscience that has contributed to the debate on multiple realizability and related issues, such as the possibility of ‘higher-level’ causation and explanation: it is becoming increasingly clear that closely connected conceptual questions (regarding reduction, say) crop up already in physics, and that a physics-based perspective may throw significant light on the debates in the context of philosophy of mind. (See, for example, Batterman 2002, Wilson 2010.) There is considerable scope here for fruitful interaction between philosophers operating within different sub-disciplines, and there are signs that inter-sub-disciplinary research along these lines is an emerging new direction.¹

Philosophy of mind stands to the philosophy of neuroscience much like (general) metaphysics stands to the philosophy of physics: in as far as (general) metaphysics concerns the fundamental nature of reality, fundamental physics can (surely) direct and throw light on it. Callender, towards the end of his essay, lays down hints for a ‘scientific metaphysics’ that represents another new direction for the philosophy of science through developing fruitful interaction with a thriving area of philosophy. Already there are signposts indicating the way to go. However, there are practical issues: how should metaphysicians be encouraged to join in partnership with philosophers of science? There have been any number of collaborative workshops and conferences, of course, and even large-scale funded projects in the metaphysics of science, but should this relationship be approached in global terms, looking at the methodologies adopted in the philosophy of science and metaphysics respectively, or via the frameworks they adopt, or should it be approached in piecemeal fashion, on the basis of particular issues to which both sides can contribute without fear of losing disciplinary identity?

Certainly when it comes to methodology, some metaphysicians themselves appear to be modelling their support of particular metaphysical views on the stances they assume scientists to be adopting with regard to their theories (see, for example, Sider 2009). Competing metaphysical claims are treated as if they were alternative hypotheses about the world, and are assessed by what are taken to be the criteria for theory choice we find in science: simplicity, unificatory power and depth of integration across domains. Close attention to
the philosophy of science will reveal just how complex and multivaried those stances and these criteria actually are (and, of course, unlike the case of scientific hypotheses, empirical success plays little, if any, role when it comes to metaphysics). But perhaps Callender’s aim of a scientific metaphysics is best achieved through piecemeal work that is focused on particular problems and issues. Hawley, for example, distinguishes between optimistic and pessimistic views of the relationship and argues that the choice as to which to adopt can only be made on the basis of a case by case examination of potential contributions of science to metaphysics (Hawley 2006).

Not surprisingly, perhaps, a number of metaphysicians wish to remain aloof from such collaborative endeavours, insisting, as Callender indicates, that the scope of metaphysics extends far beyond the limits of science and, hence, the philosophy of science. Indeed, they may insist that tying metaphysics too closely to science will constrain the field too tightly, shackling its fertility. The philosopher of science can still treat this kind of ‘science-free’ metaphysics as a rich source of plunder, drawing on both the concepts developed and the moves made in that development. One might draw an analogy here with the relationship between mathematics and science: the former generates considerable structure that at a particular state of scientific development may look surplus to requirements. However, as science progresses, this surplus mathematical structure may turn out to be heuristically fruitful, leading to new theoretical developments. The classic example of this is Dirac’s famous equation for spin $\frac{1}{2}$ particles like the electron: the solutions of this equation include some that appear to refer to particles with negative energy and thus could be regarded as non-physical, surplus mathematical structure. However, they came to be reinterpreted in terms of a new particle, subsequently identified as the positron (a form of antimatter), illustrating just how useful a resource this surplus structure can be. Now, metaphysics is not mathematics, of course, but still, even the far reaches of the field may contain elements that the philosopher of science can put to use: think of the metaphysical notion of ‘grounding’, for example, which may help explicate the relationship between different levels of ontology, or different fields, such as physics and chemistry (as touched on in Hendry’s essay); or take the idea of property ‘fusion’ (Paul, forthcoming) and think how it might be applied as a metaphysics of entangled states in quantum physics.

So, even if metaphysicians resist the call to attune their subject more acutely to scientific concerns, there are still opportunities for the philosopher of science to explore.

Turning from metaphysics to epistemology, it may be that epistemological issues in philosophy of science often appear to be just a subset of the various topics that fall under general epistemology. Philosophers of science are typically more concerned with beliefs that are ‘science mediated’ – gained through
scientific theories, instruments or a combination thereof – and less concerned with scepticism regarding more mundane forms of belief. But as Bird indicates in his essay, there is much in general epistemology that can guide these more specific epistemological debates in the philosophy of science: debates regarding general epistemological frameworks can have significant repercussions at the level of more specific arguments advanced in philosophy of science. The debate between internalism versus externalism with respect to justification and knowledge, for example, can rule out (or otherwise) certain argumentative strategies in the scientific realism debate. Thus Psillos, in his classic work on this debate, explicitly adopts an externalist perspective and uses it to analyse the relationship between the ‘no miracles argument’ and ‘inference to the best explanation’ (Psillos 1999), for instance.

More generally and, perhaps, problematically, conceptions of knowledge as ‘situated’ in a particular social context bring social factors into our epistemological framework and from there, into the philosophy of science. How to handle such factors alongside such familiar ‘epistemic values’ has been a problem at least since Kuhn’s famous book was taken to sound the death knell of logical positivism (Kuhn 1962). Giere (1988) represents one attempt to bring epistemic and social factors together within a single framework, but many philosophers of science continue to insist that the social has no bearing on matters of justification, truth, empirical success and the like. One group who maintain the significance of at least one such factor are feminist philosophers of science who argue that gender, in particular, exerts a significant influence on not only the choice of problems and issues to be tackled by scientists, but also on the selection of relevant evidence, for example. Gender bias has been taken to have a major impact on scientific objectivity in particular, but although examples from primatology and early hominid evolution appear to carry some force, it is difficult to see how similar biases might be prevalent in physics, say. Early work in this area tended to focus on the deficiencies of positivist views of science, but recent trends have moved beyond this to embrace stances such as ‘standpoint theory’ and pluralistic perspectives. (For a discussion of gender and epistemology with useful sections on feminist philosophy of science, see Anderson 2010.) It is with regard to the inclusion of social factors in general that work in the history of science has tended to diverge from that in the philosophy of science, although recently, concerted efforts have been made to bring the two back together in a more productive engagement, as we shall see below.

### 2.2 Philosophy of Science versus History of Science

As Howard indicates, the relationship between the history of science and philosophy of science is entering a new productive phase, in which the old prejudices and distrust have been overthrown, or at least set to one side.
Friedman’s neo-Kantian approach, which Howard considers in detail, offers one framework in which this relationship can be reconfigured. Chang’s ‘complementary’ approach offers another (Chang 2004). Here the history of science and philosophy of science are brought together to not only present a complementary understanding of science, but also to contribute to scientific knowledge itself. The history and philosophy of science, conceived of as a unitary enterprise, does this ‘by serving as the cold-case squad of science’, (Sumner 2005, p. 411), unearthing those features of science that have been lost or papered over by the respective field’s need to focus on certain problems, and generally subjecting both the foundations and framework to critical evaluation. Chang offers a reconciliation of Kuhn and Popper: to progress, science must become ‘normal’ in Kuhn’s sense, but then it becomes less open and critical. Complementary history and philosophy of science then takes on the critical role insisted upon by Popper, and by illuminating anomalies and reanimating overlooked ideas, it can contribute to science itself. By doing so, history and philosophy of science engages with science in a way that is neither merely descriptive, nor crudely prescriptive. In this respect, Chang’s proposal chimes with Callender’s in seeing philosophy, and philosophy of science in particular, as broadly continuous with science, and thus capable of contributing to it as well as drawing upon it. (We shall return to this issue in the next section.)

In the reoriented relationship that Chang envisages, the history of science becomes a source of forgotten questions and overlooked facts that, once dusted off and appropriately scrutinized, become fresh and scientifically significant. It is the philosophy of science that provides that critical scrutiny, reaching back into history to not only reassess the treasures that are unearthed, but also to initiate and guide historical investigations. As Chang emphasizes, this leads to a complex intertwining of the history and philosophy of science (2004, p. 240) that requires the philosopher of science to become much more historically engaged and the historian of science to become much more philosophically aware.

Both sides may balk at this bringing together, however. The historian of science may insist that there is more to her field than a kind of interdisciplinary conceptual archaeology. In particular, she may complain that such a conception leaves out the social aspects of science, with which the history of science has been preoccupied for 40 years or more (see Sumner op. cit.). Indeed, it is precisely this preoccupation that many see as having created the barrier to a more productive relationship with the philosophy of science. One option is to situate this approach firmly within social history, thereby introducing problematic demarcations within the history of science itself. Another is to seek ways in which these social aspects can be introduced into the relationship with the philosophy of science, as touched on above. Kitcher, for example, accepts that theories develop over time, with contributions from numerous researchers, thus allowing some space for social factors to be
considered (1993). More recently, he has argued that science should not be characterized as seeking objective truth; rather, this is an ideal and, in practice, what scientists pursue are significant truths, where significance reflects social interests and perspectives (2003). Giere, likewise, has attempted to accommodate the social context of science, as we have noted, by introducing a perspectival view, but one that remains objective (see also Giere 2006). Such approaches have obvious consequences for scientific realism, particularly of the form advocated by Psillos: if the kinds and classifications put forward by science are perspectival and interest relative, then how can a realist stance be maintained towards them?

Such approaches seek to reconcile the philosophy of science and socially oriented history of science by reducing or reconfiguring the realist commitments of the former and picking out those features of the latter that could plausibly be said to contribute to scientific objectivity. Whether this is a viable way to address the issue, or whether we should accept that such reconciliation is simply not possible, represents a further set of issues for the philosophy of science to address.

In addition to unearthing forgotten facts and ignored anomalies, the complementary approach encourages the historian and philosopher of science to consider the roads not taken, to explore alternative theoretical possibilities that remained undeveloped or were not taken up by scientists at the time. Cushing represents an early example of such an explorer, in suggesting that the famous Copenhagen interpretation of quantum mechanics, associated with Bohr, would not have achieved the hegemony it did if Pauli’s apparently devastating criticism of the alternative pilot wave interpretation advocated by de Broglie (and subsequently developed by Bohm) had been seen to be misconceived at the time (Cushing 1994). More recently, and in the biological domain, Radick has suggested that Weldon’s unpublished work on non-Mendelian genetics of the early twentieth century offers a plausible alternative to current conceptions in evolutionary biology (Radick 2005). Underpinning such proposals is a view of science as significantly contingent and open, and both historians and philosophers of science may reduce the scope for such alternative histories by cataloguing the heuristic moves made in scientific discovery. Furthermore, in judging such counterfactual claims, epistemic considerations must intrude. Typically, mere conceivability is appealed to as a guide to possibility, but this is too indiscriminate to help judge whether a proposed alternative road in the history of science counts as a ‘genuine’ possibility (see the collection of essays in Radick 2008). Here, general considerations of the epistemology of modality may help (see Vaidya 2007), as there may be useful tools and devices that can be imported, and certainly the philosophical evaluation of counterfactual history represents another new set of issues for the philosopher of science to consider.
Some of these issues are closely related to the debate about the force of historical evidence against scientific realism. Stanford (2006) argues that the history of science indicates that scientists are recurrently unable to conceive of radically different theoretical alternatives that would be at least equally well confirmed by the available data. This is the latest advancement of the famous anti-realist line of thought – the so-called pessimistic induction – going back to Putnam (1975) and Laudan (1981). Stanford’s ‘new’ pessimistic induction has faced criticism from the realist quarters for its construal of contemporary realist positions, for example, but the underlying historical scholarship is to be applauded. Stanford is exemplary in extending the examination of history in this context beyond its usual confines of physical sciences to the life sciences. (For another recent example in this direction, see Turner 2007.) But despite this profitable new trend, it remains the case that this dimension of the scientific realism debate is still being conducted on the basis of a very limited number of historical case studies, and more historical data should be brought to bear on these epistemological issues. We shall return to the scientific realism debate below.

2.3 Philosophy of Science versus Science
The relationship between science and metaphysics is not all one-way, as Callender notes in his contribution, with the latter acting as a descriptive hand-maiden to the former. Arguably, metaphysically informed philosophy of science can make a genuine contribution to science itself. Of course, the interesting question is: how receptive is science to such engagement? Feynmann famously declared that ‘Philosophy of science is about as useful to scientists as ornithology is to birds.’ Another Nobel Prize winning physicist, Steven Weinberg, is well-known for his dismissal of the philosophy of science, detailing the way particular forms, such as positivism, have hindered scientific progress and suggesting that the only value of philosophers of science in general is to protect science from the views of other philosophers (Weinberg 1994). Not all scientists are so negative, however, and as a counterweight to Weinberg, one can set Mayr, as an eminent scientist who is both receptive to and reflective upon the philosophy of science (Mayr 1997). That Mayr is an evolutionary biologist is significant, and here, as Brigandt notes in his essay, there is fruitful interaction between philosophers and biologists. (See the volume co-authored by McShea (a biologist) and Rosenberg (a philosopher) for one such example in practice: McShea and Rosenberg 2008).

As another example of potentially fruitful interaction, from physics, Callender mentions the theory of quantum gravity, which seeks to unify quantum physics and our best current theory of space-time, general relativity. Already there is evidence of scientists and philosophers of science coming
together to share ideas and explore new ways forward in this area (see Rickles et al. 2006; Callender and Huggett 2001). Another example is that of string theory, where philosophers can help bring clarity to the foundations of this emerging programme. Of course, one issue that has to be faced concerns the lack of empirical evidence for such theories, given the extremely high energy regimes that are involved. In such cases, other criteria may be called upon in evaluating such theories, including unificatory power, for example (see Cartwright and Frigg 2007). Such virtues may still render a research programme ‘progressive’ in the absence of direct empirical support, as suggested by Lakatos in his ‘Methodology of Scientific Research Programmes’ (Lakatos 1970). Even in these, perhaps extreme, cases, there is useful work for the philosopher of science to do.

Yet another example of this ilk – suggested by Huggett in his contribution – concerns the foundations of quantum field theory (and indeed, Huggett himself has made significant contributions here; see Huggett 2000 for a review). As long ago as 1983, Redhead called on philosophers to address the conceptual issues in these foundations (Redhead 1983). Early hopes of productive interaction between physicists and philosophers in this area have been only partly realized (see the proceedings of the 1996 Boston University Conference on the Conceptual Foundations of Quantum Field Theory, which explicitly attempted to bring physicists and philosophers of science together; Cao 1999). However, recent work has given new impetus to these discussions. Fraser, for example, has advocated a formulation known as ‘Algebraic Quantum Field Theory’ as the proper focus of philosophical investigation of these foundations (Fraser 2009). This is a highly formal approach that emphasizes the algebraic relations holding between the observables of the theory (see Halvorson and Müger 2006). Wallace, on the other hand, argues that the ‘naïve’ form of quantum field theory that physicists actually use (and that forms the framework of the so-called Standard Model of elementary particle physics) is sufficiently coherent and conceptually robust as to withstand philosophical investigation (Wallace 2006). This obviously raises a fundamental issue in the philosophy of science in general: do we explore the science that is actually used in practice, or should we restrict our attention to logically ‘clean’ formulations? The latter avoids the conceptual infelicities and, in some cases, outright inconsistencies of the former, but at a cost of losing the interpretational richness that historically oriented philosophers of science, in particular, will be interested in (Vickers forthcoming).

Yet another exciting new direction in philosophy of physics, where there is considerable overlap with current physics, involves quantum information theory. Information theory in general abstracts away from the physical representations of information, yielding a potentially powerful tool from which to view a variety of issues. Quantum information theory takes as its fundamental unit the ‘qubit’, which, unlike its classical counterpart, can represent states that are in superpositions. This suggests potentially very useful applications
to computation and cryptography: the information in such a superposition cannot be accessed without destroying the superposition, for example. The non-local nature of such superpositions also raises the possibility of a form of teleportation, in which information is transferred between systems without the need for a traditional kind of signal (see Jin 2010). As well as these exciting applications, quantum information theory offers the possibility of shedding new light on a range of issues in the foundations of quantum physics, even if one remains unsympathetic to the more radical claim that the world is made up, in some sense, of information. (For an exploration of these issues and a sceptical stance on this last claim, see Timpson 2008).

A final, but thought provoking, example of the potential for fruitful interdisciplinary interaction between philosophers, physicists and even historians of science concerns the Large Hadron Collider experiment at CERN, which has also excited attention among members of the general public. The aim of the experiment is to recreate conditions a fraction of a second after the Big Bang and perhaps discover the famous ‘Higgs particle’, responsible for conferring mass to some elementary particles, according to the Standard Model of elementary particle physics. A recently launched large-scale research collaboration (‘Epistemology of the LHC’; www.lhc-epistemologie.uni-wuppertal.de/) aims at an epistemological analysis of this piece of ‘Headline Science’, focusing on matters such as the nature of the Higgs mechanism and the Higgs particle that the LHC has set out to detect; the interaction between a large experiment and a community of theoreticians; and the epistemological implications of an experiment of enormous complexity and the need for highly selective data gathering. Studying the epistemological dynamics of the experiment in real time allows for the possibility of creating a feedback loop to suggest modifications in the course of the LHC’s operation. Such an active role is a novel and promising one for philosophy of science to play.

In order for philosophy of science to interact with science in this way, the philosopher obviously needs to master the relevant scientific discipline well enough to collaborate and engage with the scientists. Although this can be a tall order, it seems that, over the past decade or two, philosophy of science has become sufficiently ‘local’ and focused to allow the requisite depth to be reached. Many outstanding philosophers of science are today less concerned with general features of scientific practice, and more tightly focused on conceptual issues that are specific to particular disciplines. Examples of technically sophisticated, unadulterated philosophy of particular sciences abound: one can have a look at the Handbook of Philosophy of Physics (Butterfield and Earman 2007), and its entries on ‘Algebraic Quantum Field Theory’ and ‘Symplectic Reduction in Classical Mechanics’, for example. Cutting-edge philosophy of biology can also turn on sophisticated technical detail, as in the case of Okasha’s book on evolution (2006), for example.
Increased focus and interdisciplinarity pays off in the philosophy of particular sciences, as it allows the philosopher to fully take into account the best scientific understanding in its full intricacy and complexity, as opposed to merely paying it lip service by ‘popularizing’ it to other philosophers. The philosophy of science inevitably becomes more specialized and fragmented in the process, and it is harder and harder for anyone to be a ‘jack of all trades’. But if the historical development of science itself is anything to go by, this is surely a sign of progress. Due to the nature of philosophy, there is, nevertheless, an inevitable balancing act between deeply focused, local philosophy of science and general philosophy of science that offers broad descriptions and perhaps also normative lessons about science in general. Given the breadth and complexity of its subject matter, general philosophy of science often walks a tightrope between accounts that approach triviality in their abstractness and lack of substantial detail, on the one hand, and accounts that can be falsified by a case study drawn from somewhere in science, on the other. By ‘going local’, it is easier to provide substantial accounts in agreement with scientific practice, but without any overarching accounts the whole discipline risks becoming fragmented into philosophies of various particular sciences. How to resolve this tension remains as a further issue for the field to tackle.

3. General Philosophy of Science: Some Emerging Issues

3.1 Realism, Indispensability, Models
The issue of scientific realism in one form or another has been central to philosophy of science throughout its history. Nevertheless, there are many unanswered questions in the realism debate (in addition to the issue mentioned above regarding historical evidence).

One issue concerns the nature of positive arguments for realism. In the recent literature, one can find broadly two kinds of arguments. On the one hand, one can argue for scientific realism ‘globally’, as in the case of the famous ‘miracles argument’: the success of science, suitably construed, is taken to be an indicator of science latching onto reality with its theoretical assumptions, since this is arguably the best explanation of the success (see Psillos’ essay in this volume). On the other hand, one can argue for realism more ‘locally’ with respect to some particular theoretical assumptions: given the nature of those assumptions, and what we know of the world through experiments, say, we are arguably justified in making those assumptions (on pain of thoroughgoing scepticism). In this way, one can argue locally for realism about atoms, say, or microbiological mechanisms, without simultaneously advancing an argument for across-the-board realism including fundamental physics, for example (see, for example, Achinstein 2002, Kitcher 2001). One emerging set of issues
concerns the relationship between these two very different realist strategies. Are they complementary, or in competition? Which one provides the best way for the realist to defend their ground? What can one be a realist about on the basis of the local strategy?

Another outstanding issue concerns the precise sense in which the realist takes theories to be ‘latching onto reality’. Everybody agrees that even our very best theories are not perfectly accurate representations of reality, and various kinds of approximations, idealizations and inconsistencies abound in successful theorizing. The literature contains numerous responses to this prominent issue, of course, largely under the heading of ‘approximate truth’ or ‘verisimilitude’. Nothing like a consensus has formed, however, and different ways of construing the notion of ‘scientific theory’ call for different approaches to ‘approximate’ (or ‘partial’) truth (or some cognate notion). As Contessa explains in his contribution to this volume, one currently influential way of construing theories and the way they relate to reality is in terms of non-linguistic models that connect to the world via a non-linguistic representation relation. Different views regarding models and representation have evolved recently, and there’s work to be done in spelling out what ‘approximation’ or ‘partiality’ amounts to with respect to these views. Towards the end of his essay, Contessa explores some questions in this neighbourhood.

One reason models and modelling practices have received a lot of attention in the recent realism debate is this: it has become evident that much of the success of science that so impresses the realist turns on making some specific modelling assumptions that are not directly derivable from some overarching general theory. It appears that impressive predictive or explanatory success is gained often from models or simulations that incorporate various kinds of misrepresentations and idealizations that are (on the face of it) *indispensable* for this success. Hence, the following question arises. What is ‘responsible’ for success in these cases: is success down to models latching onto reality in some relevant respects – as the realist would intuitively have it – or is it in some sense down to clever idealization schemes in a way that supports anti-realism? In order to fully answer this question, one needs have a clear view of what models are, how they represent the world, and how idealizations function in modelling. A number of different positions on these issues have emerged of late, and there’s work to be done in evaluating and contrasting the alternatives. (The essays from both Contessa and Pincock discuss these and related issues.)

One issue here concerns the role of mathematics in modelling and idealizations, and this issue has recently forcefully emerged in connection with the ‘indispensability argument’ for mathematical Platonism (covered by Pincock towards the end of his essay). The debate around this argument – in its most recent guise – in a sense concerns the limits of scientific realism. The advocates
of the indispensability argument argue that (‘standard’) scientific realism implies a form of Platonism regarding abstract models or mathematical entities, due to the indispensability of the latter in scientific theorizing and explanations. Those unsympathetic to the argument, on the other hand, wish to draw the line of realist commitment at concrete, clearly spatio-temporal things, and explain the indispensability of abstract notions in fictionalist terms, for example. (See Leng 2010 for a recent monograph.) One very closely related set of issues that has recently begun to receive significant attention has to do with the role of mathematics in scientific explanations. Here the advocates of the indispensability argument view mathematical abstracta, due to their explanatory virtues, on a par with hypothetical concrete entities (such as quarks, say) as ‘explanatory posits’. Arguably, a ‘global’ scientific realist who adopts an inference-to-the-best-explanation strategy in defending realism shouldn’t discriminate against mathematical posits in her realist commitments. This kind of confirmational holism is a source of an ongoing debate, but there’s a clear lacuna here, as no theory of scientific explanation exists that would regiment these debates by providing an overarching framework for mathematical (and more generally, non-causal) explanations in science.

3.2 Causation, Laws, Symmetries, Structures
The importance of metaphysics also features prominently in Psillos’ essay as he charts the swing to metaphysics-friendly forms of realism. Although he expresses concerns about the neo-Aristotelian nature of some of these developments, there is clearly considerable scope here for the philosopher of science to draw on recent work in metaphysics. Again, consider causation, for example, a concept that, as we have already noted, crops up again and again in these essays and is clearly highly significant for our understanding of science. Here there is no consensus as to its metaphysical status, with some commentators insisting there is nothing to causation over and above the instantiation of some regularity, while others adopt an analysis in terms of certain counterfactuals, and still others argue that it is to be understood in terms of the necessary connections that arise from the powers and capacities exhibited by the properties possessed by the fundamental objects of our ontology (see Beebee et al. 2009). Psillos himself favours a pluralist approach, according to which there is an array of causal concepts specific to particular situations and, although these may bear a family resemblance to one another, there is no single umbrella account capable of capturing them all (Psillos 2009: for a survey, see Godfrey-Smith 2009). Admitting such an array may help resolve some of the disputes that arise with regard to both the existence of causation within various domains or at various levels and how best we should understand it (see Hall, this volume).
The analysis of causation is often taken to be related to that of scientific laws. Here, too, one finds a much discussed division of views, between those who insist that law statements are nothing but condensed descriptions of regularities, to those who argue that they encapsulate necessary connections in the world (see Carroll 2006). Recently, new life has been injected into the debate by a focus on the governing role laws are supposed to play. How one understands this role is problematic, and it has been argued that laws should be removed from our ontological picture entirely, with the behaviour of physical objects accounted for in terms of the dispositional ‘powers’ they possess (Mumford 2004). In support of this radical view, it has been pointed out that the philosophers’ simplistic concept of ‘law’ covers a varied array of principles, equations, rules and so forth in science. Clearly, in responding to such a position and clarifying the nature and role of laws in general, the philosopher of science needs to both pay close attention to scientific practice in order to develop an appropriate classification (see Chakravartty 2007) and then bring aspects of metaphysics to bear on the form of governance involved (see also Roberts 2009).

Laws are just one feature of theories, of course, and in modern physics in particular, symmetry principles have also played a prominent role (Brading and Castellani 2003). Indeed, van Fraassen has argued for the epistemological significance of such principles on this basis (1989). Symmetries not only play important heuristic and classificatory roles, they have also been used to generate novel predictions. Perhaps the most famous example is from elementary particle physics, where the $\Omega$-particle was predicted on the basis of such considerations of symmetry, and was subsequently discovered (Bangu 2008). Yet most philosophical analyses of laws are either dismissive of symmetry principles or fail to mention them altogether. Such a glaring omission calls for the philosopher of science to step in and develop an account that embraces laws and symmetries in a way that is consonant with the practices of physicists themselves.

Symmetries lie at the heart of certain forms of structural realism, a position that was developed as a response to the claim that a realist stance is unsustainable, since the history of science reveals a series of changing theoretical ontologies. By focusing on the laws, equations and, more generally, structures that are retained through these changes, it is claimed, a viable form of realism can be constructed (Worrall 1989). ‘Ontic’ structural realism adds to this a close analysis of certain aspects of the foundations of physics that expands the notion of structure to embrace symmetries and conservation laws and further shifts the underlying metaphysics away from objects (Ladyman 1998; French and Ladyman 2011). Indeed, it is through such symmetries that physical particles and their kinds are classified, and the structural realist takes this to reflect how the world is, with the particles themselves conceptualized not as objects per se, but as ‘nodes’ in a structure in some sense. Here metaphysics may be turned to again to defuse the explosive mix that Psillos sees in this
view and explicate a notion of structure that can accommodate the kind of causality Psillos regards as crucial for our understanding of science. Alternatively, one could revive Russell’s position and argue that causal notions have no place in physics, a stance that is attracting increasing attention again (Ladyman and Ross 2007; Norton 2007; Corry and Price 2007). And, again, a broadly pluralist approach may be appropriate here.

Laws and symmetries play a significantly smaller role in chemistry and biology. Again, metaphysics can be pressed into service in helping us get a grip on the kinds and substances of the former. And as in the case of laws, we need to pay attention to the science itself if we are to understand what is meant by a chemical bond, for instance, a concept that – as Hendry explains – plays a key explanatory role in this field. Here, structural realism may have something to contribute in urging us to stop thinking of electrons as tiny objects that ‘compose’ a bond, and in bringing to the fore the underlying symmetry that allows us to understand their behaviour in quantum mechanical terms.

Whether structural realism can be exported into the foundations of biology is a contentious issue. Nevertheless, concern about the nature of biological objects has arisen here too, with Dupré, for example, arguing that there exists a ‘General Problem of Biological Individuality’ when it comes to the issue of how one divides ‘massively integrated and interconnected’ systems into discrete components (Dupré and O’Malley 2007). He and his co-workers have urged a shift in philosophical focus away from individual genomes and distinct organismal lineages to the collaborative interactions between communities of entities from many different reproductive lineages, the very nature of which breaks down the object-oriented boundaries between such entities. In his essay for this volume, Brigandt maps out the diverse accounts of biological ‘kinds’ and advocates a version of the ‘homeostatic cluster theory’ that emphasizes the role played by relations in individuating kinds. This is consonant with a structuralist approach and the kind of contextualism that Brigandt discusses relates nicely to forms of ‘contextual identity’ in physics advocated by Ladyman, for example (Ladyman 2007). Granted the distinct differences between these two fields, perhaps such approaches offer the possibility of useful bridge-building being undertaken.

In biology, famously, one finds a dearth of laws, but a plethora of models (see Odenbaugh 2008). Bringing order to this diverse array requires the adoption of some meta-level framework, and the apparent lack of the kinds of laws one finds in physics, for example, has led many philosophers of biology to embrace the so-called model-theoretic approach to theories, particularly those versions that downplay the role of laws in general (see, for example, Giere 1999). Thus, to give just one example, one may find mathematical structures being devised which are then claimed to be similar to the spatial dynamics of butterfly metapopulations in California (Odenbaugh op. cit.). Such examples
naturally invite application of the model-theoretic account, but of course they do not constitute the entirety of model-building in biology. More so than in physics, say, biologists deploy physical models, as exemplified by Crick and Watson’s famous wire-and-tin model of the DNA helix. But more radically, they also use so-called model organisms such as fruit flies, flour beetles and so on. Can these be brought within the scope of the model-theoretic account? Some have suggested that this diversity of models should be accommodated by a sort of ‘tool-box’ philosophy of science, in which models are regarded as tools to be brought out and applied when needed (Cartwright et al. 1995). Others have responded by modifying the model-theoretic approach to capture this diversity, emphasizing the representational capacity that all such models have in common (da Costa and French 2003).

3.3 Theories and Things: Drawing on Art
The material nature of biological models also raises interesting issues when it comes to the ontology of the core elements that philosophers of science deal with, namely theories and models themselves. These are ascribed certain properties such as ‘being empirically successful’. Correspondingly, they are taken to be related to ‘the data’ as well as each other, and they are identified in various ways from current scientific practice as well as the history of science, and so on. But what kinds of entities are these? A straightforward answer is to say that insofar as theories are just sets of propositions, related via logical deduction, they are whatever propositions are. Some adherents of the model-theoretic approach have insisted that, on the contrary, theories are nothing more than set-theoretical models, and one well-known position maintains that theories and models are abstract entities (Giere 1988). Others have argued that they have a hybrid status, embracing both linguistic and model-theoretic elements (Hendry and Psillos 2007). A further suggestion is that these approaches should be seen as just representational tools for philosophers of science and that theories and models should not be identified with either propositions, sets or abstract entities (da Costa and French 2003). This leaves their ontological status open to debate (see Contessa 2010).

Given this diversity of views, how might we pin down an answer to our question above? It is here that interesting comparisons can be drawn with related themes in the philosophy of art. Consider musical works for example. These should not be identified with a particular example of the score, nor with an individual performance. But if they are regarded as abstract entities, how can we understand their creation? Likewise, it seems implausible to identify a given scientific theory with the paper in which it is first presented, or with its textbook representation and even more so to identify it with the scientist’s presentation of it at a conference (although this might well be regarded as a
kind of performance!). But if theories are taken to be abstract entities, how are we to understand their discovery? Are we to imagine scientists stumbling across them somehow? Such a picture hardly fits with what we know of scientific discovery, where specific heuristic moves can be identified. Some philosophers of art have suggested that abstract artworks can be created, since they depend on acts of human intention for both their creation and continued existence (Thomasson 2006). Could such a view be extended to theories? Again, the heuristics behind scientific discovery suggest that there is more than merely human intention involved, and it seems odd to say that quantum mechanics or the theory of evolution depend for their continued existence on human intention. Nevertheless, there is clearly scope for importing a range of moves and views from the philosophy of art into the philosophy of science.

This has already been exemplified in the case of representation, where examples from the world of art, and painting in particular, have been deployed to counter certain claims about theoretical representation, not always appositely (Suárez 2010). A more nuanced reflection on the similarities and differences can be found in van Fraassen’s recent work (2008), where connections are drawn between perspective in art and frames of reference in science, for example, and more generally between the kinds of distortions one finds in science, such as occlusion, and abstraction and idealization in science. In his essay for this volume, Contessa has delineated some of the issues involved in this recent shift to representation, and again, there is considerable further work to be done. Standing somewhere between musical works and paintings, theories and models may act as a kind of bridge between the philosophy of science and the philosophy of art, allowing for the fruitful exchange of views and approaches between both domains.

4. Conclusion

As we said at the beginning, this is a highly personal set of ‘projections’ of what we see as some of the emerging issues in the field. It is quite possible that in ten years’ time, the next generation of philosophers of science (perhaps even some of the readers of this book), will read this and shake their heads in bewilderment or pity! However, we hope we have conveyed some sense of the excitement we feel over these new developments, and we hope to have indicated some of the areas in which productive new research can be undertaken. As the field develops, new problems and issues will arise, and new areas will open up. In part, this will be driven by new developments in science itself, which philosophers of the special sciences can be expected to respond to, and partly through the new relations that are emerging with other fields within philosophy as well as with the history of science and science studies in
general. There are plenty of opportunities here and a lot of good work still to be done – far from degenerating, the philosophy of science has progressed enormously over the past 30 years or so, and we expect further progress in the future. And hopefully some of you now reading this collection will contribute to that progress.

Notes

1 In this volume, Walter and Eronen discuss reduction and multiple realizability. Craver and Kaplan look at the notion of scientific explanation (and related issues from general philosophy of science) in the context of neuroscience.

2 See Psillos’ essay in this volume for further discussion.

References


The Continuum Companion to the Philosophy of Science


A Brief Chronology of the Philosophy of Science

Peter Vickers

There can be no single, best way to write a chronology of the philosophy of science. There is the question of how one defines the discipline, and how much one focuses on which periods and which aspects of that discipline, so-defined. Thus, in addition to what follows, the reader is directed to Machamer (2002) and the introduction to Psillos and Curd’s edited volume (2008) as valuable alternative essay-length chronologies, and also to McGrew et al. (2009), which presents the history of the philosophy of science by reproducing select parts of a great wealth of important literature from the ancients through to the 1980s.

This chronology is structured around some of the most significant publications in the field, especially focusing on the most important century for the discipline: the twentieth century. Since it remains to be seen which of the more recent publications are the truly important ones, the year 2000 is chosen as a convenient stopping place. Further discussion of recent themes and trends can be found in the essay ‘Travelling in New Directions’ also in this volume.

Aristotle

There are good reasons to describe Aristotle (384–322 BC) as the first real philosopher of science. Relevant discussion can be found in his Posterior Analytics, his Physics, and his Metaphysics, for example. He has been described as a realist, in the sense that he thought one should seek out universal truths about the universe. He has been described as an empiricist, in the sense that he thought the best way to reach these truths was to observe nature, and discern universal generalizations from these observations. He has also been described as favouring a combination of inductive reasoning (moving from particular observations to generalizations) and deductive reasoning (following the model of geometry). Thus, in a sense, he presented for consideration what he took to be the goals of scientific enquiry, and the correct methodology for achieving these goals.
However, there are also good reasons to avoid referring to Aristotle as a philosopher of science in anything like the modern sense. Science as we know it today simply did not exist in Aristotle’s day. Indeed, the modern conception of science, combining theory, experiment (not just thought experiment), both prediction and explanation, and quantitative (not just qualitative) considerations, only surfaced in the seventeenth century. In Roman times and the Middle Ages, Aristotle’s philosophy continued to be discussed and debated by, among others, Proclus (410–485 AD), Philoponus (490–570 AD), Maimonides (1135–1204), Buridan (1300–1358), and Oresme (1323–1382). His influence continued to hold sway until the Copernican revolution, when philosophy of science as we know it today started to take shape.

From Bacon to Kant

In the seventeenth century, following the contributions of Copernicus, Galileo, Kepler and others, it started to become clear that previous conceptions of the structure, kinematics and dynamics of the universe had been dramatically mistaken. Francis Bacon was the first to argue that not only did Aristotle and those following him have false beliefs about the world, but also (crucially) flawed methods of forming such beliefs. In his main work of philosophy of science, *Novum Organum* (1620), Bacon developed criticisms along two main fronts. First, he urged thinkers to avoid Aristotle’s great mistake of relying too much on imagination in inventing theories (e.g. the crystalline spheres), without basing those theories on direct observation or experiment. Second, he urged thinkers not to go too far the other way, and merely describe observations without hypothesizing any cause or mechanism for what is observed (as many Ptolemaic astronomers had done). What is required, Bacon urged, is a middle way, where hypotheses are constructed and then tested. Commitment to the likelihood of such hypotheses should be withheld until appropriate tests are carried out.

However, the optimism of Bacon and his followers was soon to be countered by a realization that secure knowledge, beyond ‘mere’ hypothesis, is in many contexts impossible. Descartes (1641) was one of the first to face up to this difficulty with his search for indubitable truth. Locke took up a similar theme, although with a different focus (1690): he is sometimes called the first empiricist, since he argued that any claim going beyond experience is less than certain, however much evidence there is to support it. Hume (1748) carried on this theme in the eighteenth century, but went further than Locke in arguing that we don’t even have good reason for any beliefs going beyond experience. This is in part due to the ‘problem of induction’: we might say that it makes good sense to reason from particular experiences to a generalization. The reason this is justified, it is claimed, is because such an inference has
always been vindicated in the past. But this argument assumes the very thing it is supposed to support.

However, in the midst of such scepticism, Newton’s laws of motion, coupled with his law of gravitational attraction, were proving incredibly successful at explaining and quantitatively predicting various phenomena. Many scientists were fully convinced of the truth of the laws, and seemed to be justified in their convictions, apparently ignoring the warnings of the sceptics. Kant (1781) worked to address the tension: although still warning against the sort of unfounded speculation characteristic of Aristotelian science, Kant argued that certain ‘categories’ of human thought do apply directly to the world. For example, one can discover (a priori, synthetic) truths about space, time, substance and causality simply by pure reasoning. How we experience the world, according to Kant, is shaped by these categories: not everything derives from experience.

1840–1910s

Two major figures in the nineteenth century were Whewell and Mill. Whewell (1840) developed the so-called ‘hypothetico-deductive’ (HD) model of scientific method, according to which scientists should formulate hypotheses, deduce predictions, and then test these predictions. Under certain conditions, such correct predictions are said to be good evidence for the hypotheses. However, Whewell himself, along with Mill (1843), objected to the HD method, arguing that it ought to take more for something to count as evidence. These were major steps forward in theories of scientific methodology and what we would now call confirmation theory.

Mill also helped to develop economic theory and influenced Menger, who produced one of the first great works in the philosophy of economics (Menger 1883). Mill also argued against Kant’s ‘categories’ and claimed that everything – even logic and mathematics – ultimately derives from experience. This empiricist theme was developed further by Mach (1893), Poincaré (1902) and Duhem (1906). Despite the continuing success of Newtonian mechanics, Mach argued against making claims going beyond our current experience and (especially) our possible experience. For Mach, scientific theories are merely instruments to explain and predict phenomena, and should not be considered as candidates for ‘truth’. And Duhem furthered these sentiments, arguing that our theories should not be thought of as providing ‘true’ explanations and that they are not straightforwardly refuted by experiment. Poincaré’s philosophy did leave room for a realist interpretation, however: he argued (in a neo-Kantian spirit) that we can have confidence in the ‘structure’ (if not the material content) of our scientific theories.
Bertrand Russell shared certain views with the empiricists: for example, he urged that ‘cause’ – understood as a metaphysical concept – should be eliminated from science and philosophy of science, and he also argued against inductive reasoning. But Russell’s main contribution relevant to philosophy of science was his development of logic and the philosophy of mathematics. In *Principia Mathematica* (1910, 1912 and 1913), co-written with A. N. Whitehead, Russell tried to show that all mathematical truths can be derived from a finite set of logical axioms. Although ultimately unsuccessful in its aim, the emphasis on logic, combined with the empiricist trend, greatly influenced the next chapter in the philosophy of science.

1920s–1940s

In the first decades of the twentieth century, it had become clear that Kant had been mistaken to think that we can have a priori, synthetic knowledge of ‘categories’ such as space and time: the development of non-Euclidean geometry and of the elaboration of the special and general theories of relativity appeared to vindicate the empiricist stance. This, combined with developments in logic, led to two closely related movements emerging in the 1920s: ‘logical positivism’ and ‘logical empiricism’. Russell’s work, as well as Wittgenstein’s *Tractatus* (1921), influenced Carnap (1928, 1939), Ayer (1936), Reichenbach (1938) and others, who urged that in science (and epistemology generally) one should allow only hypotheses firmly based in experience, and observation in particular. A division between science and non-science was introduced, drawing a line between hypotheses that could be verified and those that couldn’t. The meaning of a statement was equated with its possible verification conditions; thus statements without an observational or experimental counterpart cannot be true or false, because they are meaningless.

This view presents a problem for certain sorts of – apparently perfectly legitimate – ampliative inferences, including inductive and probabilistic inferences. Such inferences go beyond pure logic and experience, and yet (it was felt) they cannot be dismissed as meaningless or unfounded. This led Carnap and Reichenbach to develop theories of probabilistic reasoning, with roots going back to the work of others in the 1920s. Keynes (1921) was a particular influence on Carnap, significantly developing the so-called ‘logical’ interpretation of probability, a view that regarded probability as a relation holding between propositions, meant to act as a guide to rational degree of belief. For Reichenbach, a bigger influence was von Mises (1928), who made significant contributions to the ‘frequency’ interpretation of probability. Here, probability is defined in terms of the frequency with which a given event occurs in some population.
A Brief Chronology of the Philosophy of Science

A further significant factor was the emergence of quantum mechanics circa 1925, which both incorporated positivistic elements (at least in Heisenberg’s ‘matrix’ formulation) and helped provide more fuel for the positivist programme. Its implications for notions of causality and determinism appeared to further undermine the Kantian approach to science, although neo-Kantians such as Cassirer (1935) argued that its more profound consequences had to do with our notion of object, and these consequences could be accommodated within a neo-Kantian understanding.

Keynes had also written on economic theory, an important test case for Carnap and Reichenbach’s philosophy of science. Given the place of ‘observable phenomena’ in economics, the question arose whether economics would have to be regarded as unscientific according to logical empiricism/positivism. Important developments in the 1930s, and major influences on what we would now call the ‘philosophy of economics’, are Robbins (1932) and Hutchinson (1938). Influenced by the philosophy of his day, the latter in particular called for a more ‘empiricist’ methodology of economics.

1950s

As Carnap (1950) developed his theory of probability, and Reichenbach (1951) set out his manifesto on ‘scientific philosophy’ two years before his death in 1953, cracks started to appear in the edifice of their programmes. In one of the most celebrated articles of the twentieth century, Quine (1951) presented serious problems for two of the most crucial assumptions of logical empiricism and positivism. First, he argued that the distinction between analytic and synthetic statements (and thus between the ‘logical’ and ‘empirical’ aspects of logical empiricism) dissolves on close inspection; second, he argued that ‘higher level’ scientific claims cannot be reduced to a combination of logic and observation statements.

Nevertheless, logical empiricism was to hold sway for a number of years yet: for one thing, at this stage there was no consensus on what philosophical framework might take its place. Braithwaite (1953), one of the early, key texts in the philosophy of scientific explanation, is thoroughly couched in the framework of logical empiricism. And Friedman (1953), drawing on the work of Robbins and Hutchinson before him, tried to develop a positivistic philosophy of economics.

But as the decade went on, further evidence that a new era was dawning in philosophy of science started to mount up. Kuhn (1957) was an important precursor to his massively influential work in the 1960s, showing how history of science can be a major part of philosophy of science (something dismissed by the logical positivists, who thought that philosophy should focus purely on
the ‘justification’ of theories, and not their ‘discovery’). And Popper (1959), although agreeing on the discovery–justification distinction, argued that the demarcation between science and non-science should be made in terms of falsifiability (or otherwise) of theories, not in terms of whether theoretical statements can be ‘verified’ by reduction to observation statements (something he had been arguing, with little success, at least since 1934). In this climate, by the end of the 1950s ‘pure’ analytic metaphysics could not be dismissed so easily, and Strawson (1959) breathed life into metaphysics as a serious philosophical pursuit.

With the decline of logical empiricism as a general framework, various different avenues of philosophy of science were to open up. Already in the 1950s there were signs of this. Wigner (1953) was a formative work on the relationship between mathematics and physics in particular. Suppes (1957) rejected the ‘syntactic’ approach of the positivists and presented an alternative way to represent scientific theories, drawing on advances in logic and model theory. This had a major influence on Stegmüller, van Fraassen and others who developed the so-called ‘structuralist’ and ‘model-theoretic’ approaches to scientific theories. And Oppenheim and Putnam (1958) argued for the ‘unity’ of science, something that accompanied the positivist emphasis on reduction, but that many felt could persist even if positivism itself did not survive the mounting criticism.

1960s

Building on his 1957 publication, Suppes published three further highly influential articles in the 1960s. Suppes (1960) argues for an understanding of ‘model’ in science in terms of the concept of ‘model’ in formal model theory, thus further supporting the development of the ‘model-theoretic’ or ‘semantic’ approach (Suppes 1967). Suppes (1962) is more famous for another reason: here Suppes argues that scientists attempt to explain not the raw data of experiments, but rather data models that have already been put together using some measure of theory. This went against extant work in the spirit of logical empiricism, such as Nagel (1961). However, the latter is remembered for being one of the first major publications on the reduction of scientific theories, laying the foundation (along with Oppenheim and Putnam 1958) for a debate which would continue to the present day.

The argument against logical empiricism continued with Putnam (1962), who built on Quine (1951) urging that a distinction between observation and theoretical terms cannot be maintained. At the same time, Popper (1962) – building on Popper (1959) – continued to develop his falsificationist programme, insisting that science should progress by means of conjectures which are then tested until refuted, to be replaced by new conjectures.
And that same year Kuhn (1962) represented a landmark in the decline of the positivist programme and presented a major challenge to the view of scientific progress as cumulative. Kuhn argued that, if one turns to the history of science, one sees that science is punctuated by a series of scientific revolutions, where scientific communities before and after a revolution see the world in such fundamentally different ways that their respective theories must be regarded as incommensurable. These communities were described as working within different ‘paradigms’ and the shift between one such paradigm and another was driven, at least in part, by social factors, raising concerns for standard notions of rationality in science. Furthermore, straightforward verification or falsification was typically unachievable due to the dependence of observation on theoretical presuppositions (the so-called ‘theory ladenness of observation’). However, perhaps the lasting legacy of this work was that it made clear (contra many positivists) that the history of science does indeed have its place alongside philosophical argument, and that there is philosophical value in a descriptive analysis of science as well as the traditional prescriptive analysis.

Meanwhile, progress was being made on a range of other issues in philosophy of science. Paneth (1962) was a major publication in the philosophy of chemistry, arguing for the unobservability of the elements, and asking what it means to be a realist about chemical theory. Smart (1963) promoted metaphysics as a viable philosophical endeavour (following Strawson 1959). This was also a seminal work in arguing for scientific realism in general, although it ignited a debate in the philosophy of biology by claiming that biology has no laws, and is thus fundamentally different from physics and chemistry. In the philosophy of physics, Reichenbach’s student Grünbaum (1963) reconfigured his teacher’s claim that space-time theory is underdetermined by empirical evidence as a metaphysical, rather than epistemological thesis. More significantly perhaps, Bell (1964) produced his famous theorem that appeared to rule out ‘hidden variable’ interpretations of quantum mechanics. Although it took some time for philosophers to appreciate its implications, much of the debate over the foundations of quantum physics through the 1970s and 1980s was motivated by Bell’s work.

Returning to general philosophy of science, Goodman (1965) centrally contributed to the debate on the logic of confirmation, with his famous ‘new’ problem of induction, and by arguing that Hume’s infamous ‘old’ problem of induction can be dissolved if we accept that methods of inference are justified by community acceptance of the results obtained. Soon after, Salmon (1967) made important headway on the use of deductive and probabilistic reasoning, and identified some problems for Bayesian confirmation theory even before Bayesianism really took off as a movement in philosophy of science.

The influence of the logical empiricist project did not wane immediately, of course. Hempel (1966) can be situated in this context, but this work
nevertheless represents a lasting contribution to the philosophy of scientific explanation and confirmation. He argued for a ‘deductive-nomological’ (DN) account of explanation, whereby it can be conceived in terms of a deduction from general laws and initial conditions to the explanandum. He also argued for a hypothetico-deductive (HD) approach to confirmation, whereby hypotheses are confirmed or disconfirmed by means of the predictions and explanations which can be deduced from them. In that same year, Hesse (1966) made a major contribution to the philosophy of scientific models, arguing that modelling is an integral part of scientific practice – contra Duhem (1906), who had argued that models lead only to confusion – and that modelling can be analysed in terms of positive, neutral and negative analogies between scientific models and the world. Drawing on his own work in physics, Wigner (1967) is notable for his contribution to our understanding of the role of symmetries and invariance principles in science, and the relationship between symmetries and laws of nature.

By the end of the 1960s it was clear that positivism could not stand up to the criticisms, and that a new philosophy of science was required. This was emphasized by Shapere (1969), who is also remembered for his discussion of ‘trans-theoretical’ terms: terms which can refer to the same entity across different theories. This idea came to play an important role in the realism debate, which was to follow in the 1970s and 1980s.

1970s

In an attempt to reconcile the opposing views of Kuhn and Popper, Lakatos (1970) provided a new and influential view of the structure and development of science. Instead of Kuhn’s monopolistic ‘paradigms’ he referred to ‘research programmes’ that can stand side by side and compete with each other. Contrary to Popper, these research programmes have a ‘hard core’ of scientific claims not open to falsification, and an ‘auxiliary belt’ of assumptions that can be rejected in the face of empirical evidence, in order to maintain the hard core. Thus, although conjectures need not be simply refuted in the face of conflicting evidence (as already emphasized by Duhem and Quine), a research programme will nevertheless be abandoned if it ceases to be progressive and becomes degenerative.

Toulmin (1972) also took issue with Kuhn’s view of science, in which ‘normal’ science is conducted within a paradigm, until it becomes ‘extraordinary’ in a scientific revolution. Toulmin suggested a novel theory of conceptual change based on evolution and natural selection, in which small and piecemeal changes take place, instead of sudden ‘revolutions’. This came at a time of important advancements in philosophy of biology itself: textbooks from
Ruse (1973) and Hull (1974) presented and discussed the most pertinent issues in the field, and set the scene for the debates to follow, including issues in genetics, taxonomy, teleology and reduction. For example, the form of reductionism presented by Nagel (1961) didn’t seem to apply to at least some biological theories. Causey (1977) took up this problem, criticizing and improving upon Nagel’s account, and arguing that even if not everything can be reduced (e.g. evolutionary biology) it is a good methodological principle to try to reduce as much as possible. Nickles (1973) was instructive here, introducing an important distinction between reduction as explanation, and reduction in the limit.

A belief in reduction was usually accompanied by a realist attitude to science. During the 1970s, realism emerged as a serious alternative to empiricism and instrumentalism. Early and influential expositions were published by Boyd (1972, 1973), to be followed by Putnam (1975a, 1978) and Leplin (1979), for example. Putnam (1975a) summed up the general spirit of the position with his famous statement that ‘realism is the only philosophy that does not make the success of science a miracle.’ This came to be articulated as the ‘no miracles argument’ in favour of realism.

However, at the same time, a number of publications emerged with a more anti-realist feel, even if they weren’t explicitly targeted against the ‘new’ realists. Quine (1975) – drawing in part on Duhem (1906) – is an important discussion of the problem of ‘underdetermination of theories by evidence’, which states that any evidence is always compatible with numerous possible theories (and thus, some claim, we have no grounds for accepting any one theory as true). Mackie’s (1974) analysis of causation is anti-realist in a more subtle way: he argued (with Hume) that causal necessity is really something that exists in the mind, rather than the world. This idea of the world being in some sense ‘constructed’ by us was taken further by Goodman (1978), who argued (following Kant) that conceptual schemes shape how things are seen. Latour and Woolgar (1979) took an even more radical stance and argued for an account of scientific knowledge as ‘socially constructed’: not only is our view of the world influenced by our conceptual scheme, it is also influenced by social factors.

This idea of a world view shaped by non-empirical, social factors was to some extent influenced by Kuhn’s account of incommensurability between theories existing in different scientific ‘paradigms’. Popper had argued against this, and for a ‘rational’ vision of science in which we are getting ‘closer to the truth’, with theories that are increasingly ‘verisimilar’. But this was dealt a blow in 1974, when Miller (1974) and Tichý (1974) showed, independently, that Popper’s definition of verisimilitude was fatally flawed, sparking an ongoing search for a working theory of truth-likeness.

Feyerabend’s *Against Method* (1975) adopted an even more extreme vision of science than Kuhn’s, developing a view which he described as ‘anarchistic’ and ‘dadaistic’. He argued against previous accounts of the ‘right’ methodology of...
science, claiming that there is no such thing as ‘scientific method’, and that many examples of scientific progress in the history of science conflicted with supposedly established accounts of proper scientific methodology. This threatened to pull the rug out from beneath philosophy of science: if this were true, it would be pointless to try to provide a general account of scientific methodology or progress.

1976 saw the publication of the social constructivist manifesto by Bloor (1976) which established the ‘strong programme’ in the sociology of science. This maintained that it was not just bias and error in science that should be accounted for in terms of social factors, but that both true and false theories should be treated alike in terms of the influence of the underlying cultural context.

However, criticisms of Feyerabend were quick to appear, with Suppe (1977) spelling out the most relevant, and the strong programme was dismissed by many philosophers of science. Kuhn himself had continued to develop his view in a series of articles since his 1962 work, and a selection of these were published together in a single volume in 1977 (Kuhn 1977). Here he elaborated on his account of paradigms and scientific revolutions, and explained that he did not want to accept Feyerabend’s claim that there is no general account of scientific progress, as some had assumed. But that same year Laudan (1977) criticized both Kuhn and Feyerabend on the grounds that their accounts of science don’t do justice to a more sophisticated understanding of the history of science: science is more rational and progressive than their accounts would suggest. (Nevertheless Laudan himself subsequently provided a major anti-realist challenge to the realist account of science as a convergent enterprise, as discussed below.)

The 1970s saw the establishment of a general consensus that logical empiricism had finally come to an end. Putnam (1975b) explained why the ‘received’ (syntactic) view of scientific theories couldn’t be tolerated. Suppe (1977) described the rise and fall of the movement, and concluded in an ‘Afterward’ that this era of philosophy of science was truly over. In this great edited volume – with contributions originating from a 1969 conference set up to explore new directions in the philosophy of science – the prospects of the field were debated by leading philosophers of science (including Putnam, Kuhn, Hempel, Nickles, Suppes and Toulmin). One option was presented by Stegmüller (1976), building on the work of Suppes (1957, 1962, 1967). This became known as the ‘structuralist’ view of scientific theories, a variation of the ‘model-theoretic’ approach focusing on highly formal representations of theories. Nevertheless, at the end of the 1970s there was no consensus on the way forward, and a fundamentally fractured philosophy of science was starting to emerge.
1980s

Widely regarded as one of the defining texts of the 1980s, van Fraassen’s *The Scientific Image* (1980) made important contributions to the major themes of the 1970s. First, he presented the ‘model-theoretic’ or ‘semantic’ approach as a successor to the syntactic approach to theories that had accompanied logical empiricism. Here the focus was on a representation of science by way of models, those of model theory in particular. But perhaps more importantly, van Fraassen presented an anti-realist, ‘constructive empiricist’ approach to science that replaced logical empiricism and managed to side-step the major objections which had been raised against the latter. In particular, he drew a sharp distinction between accepting a theory and believing it, where the former involves as belief only that the theory is empirically adequate. He then argued that the aim of science is empirical adequacy, and although the unobservable entities posited by theories might exist, we have no empirical grounds for asserting that they do. Thus, he accepted the underdetermination of theories by observation, and in that context developed a pragmatic account of explanation, showing how one could avoid realist commitments in capturing scientific practice.

Thus new life was breathed into anti-realism, and this was followed immediately by a serious attack on the realist stance, courtesy of Laudan (1981a, 1981b). The latter presented an argument which came to be known as the ‘pessimistic meta-induction’, an argument originating from Putnam (1978). Laudan claimed that the success of scientific theories is not a good indicator of truth, since many successful theories in the history of science have turned out to be radically false.

These challenges were responded to by numerous realists. Newton-Smith (1981) drew on the Putnam-Boyd-Leplin realism of the 1970s, comparing the ‘rational’ accounts of science due to Popper and Lakatos with the ‘social’ or ‘non-rational’ accounts associated with Feyerabend and Kuhn, and favouring the former. And in a collection of essays, Niiniluoto (1984) argued, against the Kuhnian vision, that science progresses and gets closer to ‘the truth’; this is exemplary of the significant body of work in the 1980s, aiming to improve on Popper’s definition of ‘verisimilitude’ in the face of the Miller/Tichý objection. Leplin (1984) is a classic edited collection which shows that at this time various different realist and anti-realist positions were emerging. In addition, Cartwright (1983) argued against realism with regard to ‘high-level’ theory and scientific laws, but retained a realist account of certain theoretical entities, such as electrons, that figure in causal explanations.

Cartwright’s account of explanation based on causation followed her rejection of the DN model of explanation, a move that was followed by several
other philosophers at this time. Kitcher (1981, 1989), drawing on Friedman (1974), developed a new theory of explanation based on the idea that science explains by unifying phenomena, and Achinstein (1983) presented a pragmatic account akin to van Fraassen (1980). Salmon (1984) developed an influential approach that combined elements of Kitcher’s unificationist account and Cartwright’s causation account. This analysis of ‘explanation’ is representative of an important tradition of analysing the key concepts in science and philosophy of science. In the 1970s there were major attempts to analyse ‘cause’ (e.g. Mackie) and ‘reduction’ (e.g. Nickles), for example. In the 1980s, in addition to ‘explanation’, philosophers focused on the concept of ‘(natural) law’, which was considered in detail in 1983 by both Cartwright (according to whom the laws of physics ‘lie’) and Armstrong (who looked at laws from a metaphysical point of view). Earman (1986) focused on ‘determinism’, a concept which had appeared straightforward, but which he showed to be more interesting and complex than had been assumed, even in the context of classical physics. Concerns with understanding confirmation also continued to be pursued, with Glymour (1980) presenting a ‘boot-strapping’ account which aimed to improve on hypothetico-deductive approaches. Howson and Urbach (1988) also pursued and extended the Bayesian approach, which turned into a major research programme in the 1980s.

Cartwright (1983) – together with Hacking (1983) – also helped lay the foundation for a new movement which came to be known as the ‘new experimentalism’. The core idea was to subject scientific experiment to more focused philosophical analysis, especially since (for Cartwright) whether one ought to believe in a theoretical entity depends on whether one can associate with the entity an explanatory causal regularity of the sort that we can independently access through a controlled experiment. Hacking (1983), along similar lines, based his ‘entity realism’ on the idea that we can manipulate some theoretical entities in an experiment, as captured in his slogan ‘If you can spray them, they’re real.’ Hacking also emphasized that ‘experiment may have a life of its own’, suggesting that philosophers of science should pay attention to the epistemology of experimental practice. The results of an experiment, he argued, are not the ‘phenomena’ to be explained, but are merely the data (cf. Suppes 1962). This distinction between phenomena and data was taken further and analysed in detail by Bogen and Woodward (1988). Also contributing to the new experimentalist theme, albeit in different fashions, Franklin (1986) and Galison (1987) presented several case studies that revealed the difficulties in deciding whether an experimental result is ‘real’ or is merely an artefact of the experiment.

The early 1980s also saw an important development in the philosophy of evolutionary theory: the ‘levels of selection’ debate. Sober (1984) was a major publication, asking the question whether the unit of selection in ‘natural selection’ is the gene, the individual, the group, the species or something else.
Although this seemed a long way from general philosophy of science, Hull (1988) forged a connection between the two, presenting an evolutionary account of the development of science, identifying ‘competition’, ‘selection’, ‘variation’ and so forth in the scientific enterprise (drawing on Toulmin 1972). This idea that the development of science could be influenced by apparently non-empirical factors such as ‘competition’, along with trends in anti-realism, was fuel to the fire of the social constructivists. Bringing the history of science to bear on this issue, Shapin and Shaffer (1985) argued that a social influence on science is evident in clashes between Hobbes and Boyle in the seventeenth century, for example.

An attempt to incorporate this ‘social’ perspective on science can be seen in the work of Giere (1988), who combined it with a ‘cognitive’ or ‘computational’ approach. Thagard (1988) also drew on advances in computing and artificial intelligence to draw lessons for our understanding of science. Giere focused more on cognitive psychology, and argued that the fact that human beings think in terms of ‘models’ (rather than deductions from axioms) is good evidence for the model-theoretic or semantic approach to scientific theories. The ‘structuralist’ programme also continued to be pursued (Balzer et al. 1987, building on Stegmüller 1976), but, perhaps because it was articulated in more formal terms, it tended to be overlooked by many philosophers of science.

The work of van Fraassen (1989) can also be viewed as further advancing the model-theoretic programme, but had more of an impact. Here he not only again deployed this approach to underpin his constructive empiricist stance, but also argued strongly against a realist approach to the ‘laws of nature’ (contra Armstrong 1983), emphasizing the critical role of symmetry principles instead. Suppe (1989), on the other hand, employed the model-theoretic approach to argue for a form of realism which incorporated the modal elements rejected by van Fraassen. Oddie (1986) and Niiniluoto (1987) also furthered the realist agenda, with accounts of ‘verisimilitude’ that were meant to accompany a vision of scientific theories gradually getting closer to ‘the truth’ (building on Niiniluoto 1984). Significantly, Worrall (1989) attempted to show how one could have the ‘best of both worlds’ with respect to Putnam’s ‘no miracles’ argument, on the one hand, and Laudan’s ‘pessimistic induction’, on the other. His ‘structural realism’ responded to the latter by emphasizing the retention of equations and (more generally) logico-mathematical ‘structure’ across theory change (drawing on Poincaré 1902).

In the philosophy of physics, however, the realist stance towards spacetime was undermined by Earman and Norton’s (1987) resurrection of Einstein’s ‘hole’ argument, which also showed how an acute knowledge of the relevant history of science could be put to good use in a modern philosophical context. Attempts to find an appropriate response to this argument continued
over the next decade, generating new positions in the debate over the nature of space-time.

1990s

As far as the philosophy of physics is concerned, in the 1990s the emphasis on the implications of Bell’s Theorem began to diminish, and the focus of attention shifted back to the measurement problem of quantum mechanics and interpretations of the latter. Saunders (1995) represents an important advance in bringing recent developments in quantum theory under philosophical scrutiny and contributing to renewed interest in the so-called ‘many worlds’ or Everett interpretation. Philosophers of physics were also increasingly venturing beyond non-relativistic quantum mechanics to consider the implications of relativistic quantum field theories, for example (e.g. Teller 1995).

General philosophy of science was clearly highly fractured by the 1990s, with no overarching framework akin to the logical empiricism which had dominated the first half of the century. Van Fraassen (1991) further applied his constructive empiricism to our understanding of quantum mechanics, again in the context of the model-theoretic approach, but although supporters of the latter continued to urge its adoption as the appropriate framework within which to analyse theories, models and their interrelationships, many remained critical.

The debate over the social influences on science was continued by Longino (1990) and Kitcher (1993) (drawing on Shapin and Shaffer 1985 etc.), with both arguing in different ways that objectivity and a cumulative view of scientific progress are compatible with those influences. This also fed into the debate over realism: realists could now draw on Longino and Kitcher to argue that social influences on science need not go against a belief in objectivity and a view of science as cumulative. Kitcher, in particular, argued that one can accept many of Kuhn’s claims about sociopolitical influences on science, while rejecting the Kuhnian view of science as fundamentally non-cumulative. A similar conclusion was reached by Thagard (1992), although couched in terms of the cognitive approach to philosophy of science (following Giere and his own work in the 1980s). He argued that scientific revolutions are rational, and that adopting a new conceptual system can be compared to learning a new language (contra Kuhn’s notion of ‘incommensurable worldviews’).

In the early 1990s, realism also received support from Lipton’s important analysis of ‘inference to the best explanation’ (Lipton 1991) which helped underpin the ‘No Miracles Argument’. In addition to providing a conceptual analysis of ‘explanation’ (in the spirit of Kitcher, Achinstein, Salmon and others in the 1980s), he argued against van Fraassen’s claim (1989) that the
‘best’ explanation needn’t be approximately true, because it could be the best of a bad lot. Critics of realism, however, could draw on Dupré’s (1993) arguments against reduction and the unity of science (although Dupré himself advocated a form of ‘promiscuous’ realism’).

Later in this period Leplin (1997) and Psillos (1999) also furthered the realist cause, both focusing on the importance of novel predictive success in science. By the 1990s it was increasingly popular to argue by drawing on detailed episodes from the history of science to make philosophical points. Leplin and Psillos did precisely this, and Psillos, in particular, was influential in responding to the case studies presented by Laudan’s ‘pessimistic induction’.

Nevertheless, the philosophy of science remained a fractured field, not tolerant of universal theories of science. Shifting attention to particular sciences, Hausman (1992) recommended caution, especially if we turn to economics. Economics is ‘different’ to other, more traditional sciences, he argued, and cannot be captured even by modified Kuhnian and Lakatosian visions of scientific practice. And a Popperian methodology of economics would destroy the discipline completely, he argued.

One candidate for a general framework, at least when it comes to the justification of scientific theories, was (and still is, according to some) Bayesianism. In the early 1990s Earman (1992) put it forward as the best hope for a unified account of scientific inference and confirmation, although he did acknowledge some fairly serious problems (such as the problems of old evidence, logical omniscience and subjective priors). Mayo (1996) presented a critical analysis (especially of ‘subjective’ Bayesianism), together with her own ‘error-statistical’ account of confirmation theory. In doing so she took the ‘new experimentalism’ movement forward, especially focusing on the importance of statistical analysis in experiment.

One area of particularly active interest in the 1990s concerned the nature of scientific representation, especially the question of how scientific models represent (e.g. Hughes 1997, Morgan and Morrison 1999). And, as ‘structural realism’ became a popular position in the realism debate (following Worrall 1989), ‘structure’ came under increasing scrutiny as its key concept, with important contributions from Psillos (1995), van Fraassen (1997) and Ladyman (1998). The latter, in particular, developed a form of structural realism that drew on developments in physics, specifically with regard to the role of symmetry. Connections began to be explored with the philosophy of mathematics, where ‘structure’ also increasingly became the focus of philosophical attention (e.g. Shapiro 1997).

Cartwright (1999) offered her own form of ‘dappled’ realism, which can be summed up in a modification of Hacking’s slogan: ‘When you can spray them, they’re real.’ She continued to emphasize that laws are ‘ceteris paribus’ (following her 1983 publication), but now she claimed that laws don’t ‘lie’,
because they do not carry representational content at all: they are just tools in the ‘toolbox of science’ (Cartwright et al. 1995). This vision of science as disunified (following Dupré 1993), with reduction rejected in favour of more complex relations between different domains and levels, had become a major focus by the end of the 1990s, with interesting work conducted by Teller 1992, Humphreys 1997, Kim 1999 and Sklar 1999, among others.

This, of course, represents only a snapshot of where the discipline stood at the end of the twentieth century. In the first decade of the new century, some of these developments have continued to flower, others have withered, and new programmes have been initiated. It is too soon to say which are significant, but some indication of the directions in which the field is starting to move can be found in the essay ‘Travelling in New Directions’ (this volume).

**Bibliography (by Chronological Period, in Chronological Order)**

**Pre-history**

**Bacon, Hume and Kant**

**Turn of the Century**
A Brief Chronology of the Philosophy of Science


1920s–1940s

1950s


1960s


1970s


A Brief Chronology of the Philosophy of Science


1980s
1990s–present day


379
Part IV

Resources
This page intentionally left blank
Here is a selection of books that we, the editors, regard as particularly significant, accessible or just plain interesting. It’s a highly personal selection, but we hope both the beginner and the advanced reader will find it helpful.

**General Philosophy of Science**


A technically demanding work, this book makes the case that a certain form of reasoning that is prevalent in many discussions of physical phenomena has been overlooked by philosophers. The central idea is that understanding these phenomena, which involve ‘universal’ behaviour, requires reducing the level of detail in the relevant explanations and adopting ‘asymptotic’ methods. Examples are drawn from optics, magnetism and thermodynamics, and the implications for our understanding of emergence and reduction are explored.


A great collection of papers on issues related to reduction and emergence, ranging from philosophical classics to articles by scientists that are less well known to philosophers. Overall, this is a very good introduction to the field with a good balance of philosophy and science. In particular, it covers the ‘contemporary’ issues that emerge from complexity theory, artificial life, physics and biology.


A brick of a book that contains more than 40 classic readings in the philosophy of science. Beginning with Schlick on ‘Positivism and Realism’, and proceeding through van Fraassen on saving the phenomena, Fox Keller on ‘Feminism and Science’, and Garfinkel on ‘Reductionism’, it also includes sections on the philosophy of physics, the philosophy of biology, and, unusually perhaps, the philosophy of psychology. With useful introductory essays, this book remains a handy resource, even if it is somewhat outdated.

Although this could be dismissed as an historical artefact, it actually offers a highly readable introductory account of a range of fundamental issues, from explanation and prediction, through experiment and measurement, to issues in the foundations of quantum physics and space-time theory, all from the perspective of one of the founders of logical positivism. Well-written and clearly argued, it includes an introduction from Martin Gardner.


This is a good introductory reader on laws of nature that covers many of the central topics, although there are some clear omissions as well, such as James Woodward’s conception of laws (to be found, for example, in his *Making Things Happen*, see below). The selection of original articles makes this a good companion to Psillos’s *Causation and Explanation* (see below).


Often regarded as a companion volume to Hacking’s *Representing and Intervening*, this provocative series of essays covers a range of topics from the nature of models to the reality of causes and the dissolution of the measurement problem in quantum mechanics. However, its most well-known and central thesis is that, insofar as they feature in explanations, the high level theoretical laws of science should not be regarded as true, whereas the low-level phenomenological laws offer true descriptions, but don’t explain very much. Richly illustrated with examples from physics (such as the working of lasers and the way in which particular Hamiltonians are taken down ‘off the shelf’ in quantum mechanics), it is this central claim that has provoked most controversy, although Cartwright’s analysis of ‘ceteris paribus’ conditions has now become a standard reference point in the field.


Here Cartwright argues for a disunified view of science, with laws understood as constituting a ‘patchwork’ rather than a ‘pyramid’, and the regimented behaviour that everyone thinks is characteristic of science is taken to actually be the result of good engineering. Radical in both its view of science and its realism, the book also offers detailed analyses of laws in physics and economics and examines the relationship between classical and quantum physics.

An accessible, if somewhat dry, introduction that begins with the common-sense view that science is based on ‘the facts’ and takes the reader through a useful discussion of observation and the role of experiment before discussing the standard array of topics on induction, falsificationism, Kuhn’s paradigm-centred philosophy of science, Lakatos’s methodology of scientific research programmes and Feyerabend’s ‘anything goes’ approach. Its somewhat outdated feel is offset in the latest edition by the inclusion of chapters on Bayesian confirmation theory and the ‘new experimentalism’ championed by the likes of Hacking.


Drawing on a detailed historical analysis of the development of temperature scales, Chang presents his vision of history and philosophy of science as both complementary to and the continuation of science by other means. Issues such as observability, measurement and the nature of scientific progress are also covered in this work that won the prestigious Lakatos Prize in the Philosophy of Science.


Put together in response to van Fraassen’s *The Scientific Image* (see below), this collection maps out the central issues behind the realist-antirealist debate. Noteworthy essays include Churchland on ‘The ontological status of observables’, Hacking on ‘Do we see through a microscope?’ and Musgrave on ‘Realism versus constructive empiricism’. It concludes with a series of responses from van Fraassen himself that further illuminates his anti-realist stance.


Extending a special ‘50th jubilee’ issue of the *British Journal for the Philosophy of Science* with some specially commissioned essays, this collection covers issues to do with theory-testing and success, the realism-antirealism debate, causation and the interpretation of theories, as well as the philosophies of biology, mathematics, mind, quantum mechanics, quantum field theory and space-time physics. A trenchant critique of sociological and feminist analyses of science adds further bite to this broad survey of the field as it stood around the turn of the millennium.

Although a bit of a mixed bag, this volume of essays from philosophers and scientists includes an excellent and eminently accessible introduction to the causal account of explanation from Peter Lipton, together with contributions from physicist Steven Weinberg on whether science can explain everything, Astronomer Royal Martin Rees on ‘explaining the universe’, Stephen Rose on explanation in biology and Jack Goody on social anthropology.


Another weighty tome, this book presents 50 classic papers and is divided into nine sections, each followed by a detailed commentary covering the main issues and core arguments of that topic. The sections are ‘Science and Pseudoscience’, which includes essays on an early U.S. court case on creation science; ‘Rationality, Objectivity and Values in Science’, which includes a discussion of gender and the biological sciences; ‘The Duhem–Quine Thesis and Underdetermination’; ‘Induction, Prediction and Evidence’; ‘Confirmation and Evidence: Bayesian Approaches’, which includes Glymour’s classic paper ‘Why I am not a Bayesian’; ‘Models of Explanation’; ‘Laws of Nature’; ‘Intertheoretic Reduction’; and ‘Empiricism and Scientific Realism’, which includes critical analyses of van Fraassen’s constructive empiricism and Hacking’s entity realism. Although it could use some updating, this is still one of the best anthologies on the market.


Sophisticated and balanced, this is a standard textbook on Bayesian confirmation theory in philosophy of science. Although it doesn’t attempt to avoid technicalities, it nevertheless remains accessible. A good accompaniment to the more pro-Bayesian *Scientific Reasoning: The Bayesian Approach*, by Howson, C. and Urbach, P. (Chicago, IL: Open Court, 2006, 3rd edn).


Some have suggested this book should come with a health warning! Often criticized for advocating the teaching of creationism and witchcraft alongside the theory of evolution and quantum physics, Feyerabend’s views are actually more nuanced than his Dadaist slogan of ‘anything goes’ suggests. A product of its times, the book nonetheless offers provocative, but always interesting, insights on scientific practice, drawing on a wide range of examples.

A standard work in the ‘new experimentalism’ movement that urged a shift in focus within the philosophy of science from theory to experiment. Less radical in its claims than Hacking’s *Representing and Intervening* (see below), it contains detailed (and we mean detailed) accounts of a range of experiments, from the discovery of parity non-conservation in high energy physics, to the observation of lattice structures in cell cytoplasm.


In this slim volume, Friedman defends his neo-Kantian view of a historicized philosophy of science (as discussed by Howard in his essay). In an attempt to balance scientific objectivity with a view of science as undergoing revolutionary changes, Friedman argues that rational consensus is preserved across ‘paradigm shifts’. The book covers important aspects of the history of the philosophy of science, but the more fundamental theme concerns the interrelationships between science and philosophy in general.


The aim of the book is to present a ‘cognitive’ theory of science in terms of the processes involved in representation and judgment. However, it also incorporates useful and interesting material on the nature and role of models in science, the limitations of anti-realist stances when faced with laboratory practices and two extended case studies: nuclear physics and, unusually, the revolution in geophysics that led to the theory of continental drift. Clearly written, it also includes useful summaries of logical positivism, Kuhn’s philosophy of science, sociological approaches and others.


Part introductory text, part defence of a shift in focus from theory to experiment, this is accessible, well written and full of interesting (if sometimes rather sketchily presented) examples from scientific practice. Although the discussion of the state of play in the debate over scientific realism is now out of date, it is here that one finds the articulation of entity realism, with its famous slogan ‘If you can spray them, then they’re real.’ Unusually even for a North America–based philosopher of science, Hacking urges reappraisal of the American pragmatist philosophers, particularly Dewey, and advocates a view of scientific experiment that emphasizes intervention, rather than passive observation.

Hacking covers all the basic notions and philosophical interpretations in probability theory in a rigorous but extremely accessible way, without unnecessary formalities. The book presents a balanced and well-informed introduction to topics such as Bayesian vs. frequentist theories of confirmation, the problem of induction and elementary decision theory. It also makes use of some clever examples and puzzles that make the reader think about the issues and fallacies involved in inductive reasoning and logic.


This is an introduction to eight disparate issues in philosophy of science, organized in a debate format, with two papers defending explicitly opposite positions on each issue. The topics covered are somewhat particular, and admittedly some of the debates are less engaging than others, but there is some very accessible philosophy here, and the debate format brings the issues alive.


One of the few philosophically rigorous accounts of the role of computers and simulation in science, Humphrey’s book discusses a broad range of modelling methods, drawing primarily on examples from physics. It spells out the consequences of such techniques for our understanding of scientific knowledge and concludes with a discussion of ‘further issues’, including the role of abstraction and idealization.


This book offers a robust defence of the notions of objectivity and progress in science. Kitcher argues that many criticisms of these notions have focused on the idealizing features of science, and he offers an account in terms of ‘real world’ decision making, complete with bias, the impact of social factors and so forth. A crucial concept here is that of a ‘scientific practice’, which is responsive both to the external world and to input from scientific colleagues. Scientific progress is then described in terms of an advancing consensus practice which deploys explanatory resources, framed in terms of ‘schemata’ that accumulate, along with truths. Challenging, but clearly written, the book also demonstrates the author’s expertise in the history of science, with regard to Darwinian evolution in particular, which is drawn upon to provide supporting case studies.

This is a helpful and accessible introduction that adopts a broadly historical approach, beginning with the positivist view of science and then proceeding through falsificationism, the Kuhnian view, and subsequent sociological and feminist accounts, as well as the problem of underdetermination, realism and explanation. It is particularly notable for deploying examples from immunology, in contrast with many books in this area that tend to draw only on examples from physics.


This offers another selection of classic papers covering a wide range of topics, including the theoretical/observational distinction, holism and underdetermination, realism-antirealism, the (inevitable) Kuhnian model of science and sociological approaches and feminism. It's obviously somewhat out of date, but still contains useful material.


Another classic, this work is widely regarded as having led to the demise of logical positivism (although, as always, the story is more complicated than that). Written in an accessible, non-technical manner, it is here that Kuhn introduces the notion of ‘scientific paradigm’, in terms of which normal science is conducted, and argues that scientific revolutions, characterized as paradigm shifts, have more to do with social factors than the standard epistemic virtues. More generally, Kuhn is credited with helping bring the philosophy of science back into positive engagement with the history of science.


This collection contains ten substantial chapters that review the current state of the art in general philosophy of science, ranging from theories of explanation (Psillos) to the demarcation problem (Mahner). There are many ‘focal issues’ that do not get covered, and the chapters are not exactly introductory, but this a good place to get broad up-to-date overviews on some of the central topics in philosophy of science.


This is an accessible and engaging introduction that covers all the main topics, beginning with induction and taking the reader through falsificationism,
Kuhn's paradigm-based approach, the realism debate and explanation. Given the author’s research interests, it should come as no surprise that the discussion of realism and alternatives such as van Fraassen's constructive empiricism is particularly good.


The book begins with a highly polemical critique of current metaphysics before offering an extended and detailed defence of structural realism. The authors take today's metaphysicians to task for failing to consider the implications of science for their highly elaborate and often quite abstract worldviews and urge the adoption of a naturalistic approach. The elaboration of their ‘rainbow realism’ goes beyond physics to include economics and covers further issues, such as understanding causality in this context.


This recent anthology in philosophy of science brings together a large selection of classic papers. It begins with Hempel and logical empiricism, and ends with Maudlin on the metaphysical implications of modern physics, covering all the standard topics from general philosophy of science in between. There’s a good deal of overlap with the other leading anthologies (for a reason), but this is one of the most up-to-date, and with its 38 articles, it is also very comprehensive.


Lipton's is the most detailed and thorough attempt to spell out the popular idea that explanatory virtues can guide inductive inferences: in a situation of underdetermination we (or scientists) infer the hypothesis (or the scientific theory) that best explains the observations. This second edition also covers some important related issues, such as IBE versus Bayesianism in confirmation theory, and the role and status of IBE in the scientific realism debate.


Perhaps our clearest competitor! After an introductory chapter on the history of the subject and another on the ‘classic’ debates and future prospects, this volume presents a series of specially written essays by leading philosophers of science on such standard issues as explanation, reduction, observation and so forth, as well as on more specialized topics like evolution and the philosophy of space-time. It concludes with a chapter on feminist approaches to the philosophy of science. Still, ours is better.

This recent anthology is quite special in its substantial historical dimension, stretching from the ancient and medieval periods (e.g. Zeno, Plato, Aristotle) to (almost) the current day (e.g. Boyd, Laudan, van Fraassen on scientific realism). While (of course) not the best collection on contemporary philosophy of science, the very large selection of historical readings, naturally organized into sections prefaced by the editors’ introductions, make this unique and very impressive collection highly useful for those who wish to tap into the history of science and the history of philosophy of science.


The central claim of Mayo’s important book is that we learn from error, but errors come in different forms. This forms the core of her ‘error-statistical’ view of the relationship between theory and experiment which eschews Bayesian accounts in favour of a piecemeal modelling approach. There are useful discussions of novel evidence, severe testing and the flaws in Bayesianism, and case studies include Brownian motion and the apparent observation of the deflection of starlight by the sun as predicted by Einstein, where Mayo famously argues that the observation was not as decisive as is typically thought.


Morrison analyses the nature and role of theoretical unification in science by focusing on general features of unified theories. Perhaps the most contentious aspect of the book concerns the claim that unification and explanation have little, if anything, to do with one another. The analysis is illuminated with case studies from both physics (such as Maxwell’s theory of electromagnetism and the theory of electroweak interactions), and biology (the evolutionary synthesis).


This disparate collection of essays on the different roles of models in science and economics has become most well known for the editors’ introduction, in which they present the view of scientific models as ‘mediators’ between high-level theory and the phenomena. It also contains some useful case studies, ranging from the Ising model and superconductivity in physics, to ‘paper tools’ in chemistry, and Marx’s ‘reproduction schema’.

A clear and systematic coverage of, and introduction to, the now somewhat outmoded debate between Popper, Lakatos, Kuhn and Feyerabend on the (ir) rationality of science. Newton-Smith defends a moderate realist and rationalist position.


With a list of topics from ‘Axiomatization’ to ‘Whewell’ and a range of contributors from Worrall to Achinstein, this work offers an encyclopaedic series of snapshots of issues in the philosophy of science written by leading researchers in the field. If some of the issues covered have moved from the core to the periphery (e.g. ‘Craig’s Theorem’; ‘Incommensurability’), this is still an instructive beginner’s guide and a handy reference work for the more advanced student.


An excellent and, as it says on the tin, short introduction to the major topics in the field. It begins with a discussion of the distinction between science and pseudo-science, before tackling induction and probability, explanation, the realism debate, Kuhn and sociological approaches, among other topics. Given the author’s specialism, the outline of the philosophy of biology is particularly good, and unusually for an introductory text, philosophical issues in linguistics are also covered. The book ends with consideration of the relationship between science and religion and the so-called ‘science wars’ between ‘rationalist’ and sociological approaches.


Primarily focused on the realism-antirealism debate and no longer at the cutting edge, this collection contains some still useful pieces, such as Worrall’s original paper on structural realism, Laudan’s presentation of the pessimistic meta-induction and van Fraassen on saving the phenomena. Papineau also provides a helpful introduction to the debate.


Drawing on developments in computer science, this work presents an advanced and technical analysis of complex and multidimensional probability distributions in terms of directed graphs. The aim is then to extract causal
models from such representations. Winner of the 2001 Lakatos Prize in Philosophy of Science, this book covers recent work in this field and clarifies a number of fundamental concepts. Although not easy, it is required reading for anyone with a serious interest in such matters.


Overhyped as ‘one of the most important documents of the twentieth century’, this is still a classic work in the philosophy of science. Originally written in German in 1934, it was rewritten in English by Popper himself and republished in 1959. The irony of the English title has not been lost on the generations of philosophy of science graduates who have opened it only to find that ‘discovery’ is relegated to the realms of psychology and social studies (the original title would have been better translated as *The Logic of Scientific Research*). The logic of science for Popper was not inductive, but involved a straightforward application of *modus tollens*: one draws a prediction from one’s theory; if that prediction fails to be supported by the appropriate observations, the theory must be deemed falsified, and one starts again with another. This underpins Popper’s falsifiability criterion for demarcating science from pseudo-science. Although seeming rather outdated, not surprisingly, *LSD* also contains useful discussions of simplicity and conventionalism. More contentious, perhaps, are Popper’s defence of the propensity interpretation of probability and his untenable interpretation of quantum mechanics.


This is the best book-length elaboration and defence of scientific realism currently available. After a historical opening, it develops the causal theory of reference in terms of which theoretical terms can be said to refer to unobservable entities and defends realism against the Pessimistic Meta-Induction from the history of science and the Underdetermination of Theories by empirical data. Although pitched at quite an advanced level, the book is clearly written and includes illuminating case studies, such as the caloric theory of heat and Maxwell’s theory of electromagnetism.


An accessible and helpful introduction to the major philosophical debates surrounding these concepts, as well as that of scientific law. It is also appropriately historically sensitive and draws upon the relevant metaphysics. The command of the literature is impressive, and although Psillos is even-handed in his presentation, he makes it clear that he favours a broadly Humean approach, while acknowledging the price that has to be paid in adopting such a position.

This contains 55 specially commissioned essays on a range of topics and issues in the field. Part 1 covers the ‘Historical and Philosophical Context’, with essays on metaphysics, the role of logic in the philosophy of science and its relationship with the history of philosophy. Part 2 is simply entitled ‘Debates’ and includes essays on Bayesianism, relativism, underdetermination, theory change and others. Part 3, ‘Concepts’, looks at causation, idealization, representation, reduction and symmetry, among others. The final part examines the philosophy of the ‘individual sciences’, such as biology, physics, cognitive science, economics and the social sciences. This is a comparatively up-to-date reference work that provides a useful entry point to the most pressing debates and issues in the subject.


Following its supposed demise at the hands of Kuhn and others, logical empiricism has recently been re-evaluated, both within its historical context and as a source of insight into how science works. This collection includes some of the results of this new scholarship and presents both the history of the Berlin and Vienna ‘circles’ together with essays on the relationship between logical empiricism and the special sciences, such as mathematics, physics and social science, as well as reassessments of the critics of the movement. Perhaps one for the seriously interested student only, this nevertheless contains much useful information on a movement whose views still resonate within the philosophy of science.


Steering a course between the Humean and non-Humean accounts of scientific laws, Roberts adopts a meta-theoretic analysis, which attempts to reveal the ‘law-role’ played by certain statements in scientific theories and to satisfy our intuitions about such statements. Although this is an advanced-level work, it offers a suggestive new approach to this issue and is vastly better than most philosophical accounts of laws when it comes to grasping the relevant features of scientific practice.


With the aim of showing that the problems of philosophy of science are among the most fundamental problems of philosophy in general, the book begins with a discussion of the relationship between these two fields, before
moving on to a detailed consideration of explanation, causation and laws, spanning two chapters. The structure of scientific theories, the problem of theoretical terms and the nature of theory testing are also covered, and the book ends with a discussion of relativism in the study of science. Helpful lists of further readings are included, together with study questions at the end of each chapter.


Salmon is widely credited for bringing the issue of explanation back onto centre stage, and his book contains an excellent overview of the state of play at the time. More importantly, it presents his causal account of explanation, in which the focus is on processes, rather than on events, and which remains one of the most important approaches to understanding scientific explanation. Widely cited, this is essential reading for anyone interested in the topic.


Although now somewhat out of date, this historically oriented and masterful survey of major views of explanation in science still has a lot to offer. Salmon argues, in particular, that unificationist and causal approaches to explanation are, in fact, complementary, rather than competitors and that, more generally, there can be no universal account of explanation, since it depends on contingent features of the world.


Unusually for an intervention in the realism-antirealism debate, Stanford draws extensively on examples from the history of biology, and in particular mid-to-late nineteenth century theories of inheritance and generation. His central focus is the ‘problem of unconceived alternatives’: if, as he claims, past scientists have overlooked serious alternatives to their theories, how can the realist insist that the theories we have ended up with are the best available and therefore worthy of our commitment? Instead of realism, Stanford offers a form of neo-instrumentalism according to which theories are mere, if powerful, tools for achieving our practical goals.


A long, detailed and sophisticated account of the causal view of explanation that both develops it further than has been attempted before and responds to criticisms of this view. Not for the philosophical faint of heart, the book is nevertheless clearly written and begins with a helpful critical survey of
previous accounts before plunging into the deep and demonstrating how fruitful the causal outlook can be.


At one time in the vanguard of new developments, now an interesting piece of history, this collection represents the state of play in the field at the end of the 1960s/beginning of the 1970s. Stand-out papers include Suppes on the analysis of data, Cohen on the relationship between the history of science and the philosophy of science and Shapere on the notion of ‘scientific domain’. However, it is perhaps most useful for the outstanding introduction and afterword by Suppe, which present a detailed history of the rise and fall of both logical positivism and the philosophies of Feyerabend and Kuhn – its critiques of the latter two remain some of the most powerful in print.


Blowing apart the realist consensus of the late 1970s, the constructive empiricism that van Fraassen presents in this book still represents, for many, the most robust and best developed form of scientific anti-realism available. Couched in terms of the so-called semantic view of theories, which presents them in terms of formal models, van Fraassen’s stance represents a major challenge to the scientific realist. Suitable for the advanced reader, the book also includes empiricist analyses of explanation and probability theory.


Here van Fraassen further develops his constructive empiricism and applies it to the notion of scientific law. In keeping with his rejection of metaphysics, he offers an updated version of the Humean ‘regularity’ view and, more significantly perhaps, urges a shift in focus to the notion of symmetry, which plays a fundamentally important role in modern physics.


A detailed and nuanced analysis of the way in which theories may be said to represent, this draws on multiple examples from the history of art as well as developments in modern physics. It also includes discussion of the notion of measurement and historically oriented chapters that feed into the presentation of van Fraassen’s recently elaborated view of ‘empiricist structuralism’.

This is a weighty, dense tome of almost 700 pages. It offers a highly original, if maverick, take on concepts and theories, turning heavily on a string of rather thought-provoking examples drawn from applied mathematics and engineering! A good example of philosophy of science that bravely crosses some traditional sub-disciplinary boundaries. Stylishly written, if somewhat long-winded at times.


The now classic exposition of the ‘interventionist theory’ of causation that holds that it is the possibility of intervention (in an idealized sense) that is essential to causation and explanation. Careful and well argued, the work draws on useful examples from scientific and medical practice as well as everyday life. Indeed, this illuminates one of the virtues of Woodward’s approach, which is its broad applicability to the biomedical and social sciences arising from its lack of dependence on specific views of scientific laws, for example. The downside is that there is no discussion of quantum physics, an area in which this kind of account faces obvious problems.

Philosophy of Biology


An accessible and well-written introduction, this focuses on the theory of evolution, beginning, of course, with Darwin and then taking the reader through the foundations of the theory, before discussing a range of philosophical issues that it generates, such as those to do with the notions of innateness, function, classification and species. The book concludes with the implications of the theory of evolution for religion, psychology and ethics.


Including commissioned contributions from both senior scholars and rising stars, this offers a broad selection of papers that tackle most of the central issues in the field, including, in particular, the role of information in biology, game theory, mechanisms and models, evolutionary psychology, as well as biodiversity and bioethics. Although some of the contributions are quite specialized (e.g. ‘The moral grammar of narratives in history of biology’), many both provide useful introductions to the central debates and take the reader to the leading edge of thought regarding the issues involved.

A significant contribution to the growing philosophical literature on biodiversity, this covers legal and ethical as well as biological and methodological issues. The authors argue against the view that biodiversity stands for a single property of biological systems and instead adopt a pluralistic approach according to which characterizations of the concept are context-dependent. The book highlights the problems faced by the claim that biodiversity must be valued and makes an important contribution to the philosophy of environmental science.


The product of collaboration between a biologist and philosopher of science, this book is a clearly written and highly accessible introduction. It begins with Darwin (of course), covers such issues as the lack of laws in biology, reduction and explanation, and progress in evolution and concludes with a discussion of biology, behaviour and ethics. The book very nicely sets out the central issues and arguments in the field and covers a lot of ground for its size.


With a helpful introduction summarizing the contributions, the aim of this collection is to both introduce the newcomer to the field and also challenge those who already have a basic knowledge of the relevant literature. The essays included cover a wide range of topics, including, in particular, the structure of evolutionary theory, teleology, species and taxonomy, the ‘evo-devo’ debate, animal behaviour, the concept of race in medicine and feminist approaches to the philosophy of biology.


Another major collection, this time providing clearly written overviews of philosophical issues which arise in a range of contexts within contemporary biology. In addition to the usual discussion of evolution and developmental biology, it also covers medicine and ecology, as well as addressing broad themes from the philosophy of science to do with the role of theories and models, reductionism and the nature of biological experimentation.


This remains one of the best textbooks under this heading on the topics it covers and is clear and philosophically sophisticated. It is a good companion
Sober places more emphasis on the logical structure of the debates, while *Sex and Death* brings to the fore more of the relevant biological detail.


This anthology provides a great selection of original articles for teaching and learning the central issues in the philosophy of evolutionary biology. The 3rd edition covers evolutionary psychology and laws of nature in evolutionary theory, for example, in addition to standard issues such as units of selection, adaptationism, reductionism, species and evolutionary ethics.


This book could be listed under ‘general philosophy of science’ as well for its provision of an accessible and thorough discussion of the concept of evidence in general. By comparing Bayesianism, frequentism and (Sober’s pet account) ‘likelihoodism’, Sober arrives at a kind of a pluralist position. Armed with the conceptual tools developed over the first 100 pages or so, the book goes on to provide a classy analysis of the creationism debate and other more specific debates within evolutionary biology.


This introductory text provides a wealth of exciting biological detail to accompany its philosophical analysis, nicely fleshing out and bringing alive topics that readers (non-philosophers, especially) may find a little dry. Written in a clear and lively manner, it is a good companion to Sober’s introductory monograph mentioned above.


Constructed around detailed case studies from cell biology, developmental genetics, neurophysiology and molecular genetics in general, Weber’s book goes beyond the topic in its title to consider a range of issues in the philosophy of science, from discovery and explanation to reduction and realism. Useful summaries and critiques of prominent views on these issues are presented, but the focus on experimental aspects of biological research is especially illuminating.
Philosophy of Chemistry


This collection begins with scene-setting introductory essays on the history of the development of the philosophy of chemistry, together with reflections on Aristotle's and Kant's views of chemistry. It then presents a diverse and somewhat disparate set of discussions on such topics as explanation and chemical bonds, causation in chemistry, pictures and chemical composition and, of course, the issue of chemical kinds.


Van Brakel makes a plea for the philosophy of chemistry as a distinct discipline, but also relates the discussion of discipline-specific issues to general topics in the philosophy of science, such as explanation, reduction and the nature of laws, as well as emergence and essentialism. Although it offers a gateway to discussions in contemporary philosophy of chemistry, this is not a book for the beginner.

Philosophy of Mathematics


A very readable little book defending and elaborating the so-called indispensability argument – originating from Quine and Putnam – for mathematical Platonism, according to which we should take abstract mathematical entities as existing in the world, due to the fact that mathematics is indispensable for our best scientific theories.


This is a very accessible introduction to the philosophy of mathematics, covering some of the history, the main philosophical positions of the twentieth century and the ‘contemporary scene’, including mathematical structuralism – Shapiro’s own position. It doesn’t cover some of the more recent debates regarding the applicability of mathematics, but it does provide a good initiation for all philosophy of mathematics.

Philosophy of Neuroscience

As it says in the title, this pioneering work offers a unified approach to the nature of the ‘mind-brain’. Drawing on research in neuroscience, it argues for a form of eliminativism in which mental processes are viewed as nothing but processes in the brain and, more generally, that bringing neuroscience and psychology together offers a richer and more successful framework than so-called folk psychology.


As becomes clear from his contribution to the present volume, Craver defends a mechanistic approach to neuroscience, and in this book he presents a detailed account of this approach. Covering issues of explanation, reduction and the integration of experimental results, this is a thorough, if demanding, work in the philosophy of the neurosciences.

**Philosophy of Physics**


A snappy and non-technical introduction to the central issues in the philosophy of quantum mechanics, this covers superposition, non-locality, measurement and the collapse of the wavefunction, as well as the Everett and Bohm interpretations. The book concludes with Albert’s own ‘many minds’ view, but even if one shies away from the latter, this still represents a useful entry point to the subject.


An excellent collection on a philosophically overlooked topic, this includes ‘classic’ papers from the likes of Leibniz, Kant, Curie and Weyl, together with commissioned essays from some of the leading philosophers of physics. Highlights include Ryckman on the connection between gauge symmetry and transcendental idealism, Pooley on parity violation, and Morrison on spontaneous symmetry breaking. Although many of these pieces are pitched at quite an advanced level, Brading and Castellani’s introduction is accessible and spans a broad range of issues.


Winner of the prestigious Lakatos Prize in Philosophy of Science in 2006, this advanced work offers a major reappraisal of special relativity in terms of the causal and dynamical understanding of length contraction and time dilation.
The Continuum Companion to the Philosophy of Science

Historically astute, the standard topics of simultaneity, covariance and consistency with quantum physics are considered from this rediscovered perspective.


This offers a classic treatment of determinism – a notion that pervades much of philosophy – revealing how complex and subtle the notion can be even at the level of classical (i.e. non-quantum) physics. Earman covers general relativity and quantum physics as well, and touches on freewill and many other topics to be expected under this heading, such as chaos and randomness.


Beginning with a historical introduction to the origins of the debate over absolute versus relational accounts of space and time, this takes the reader through the foundations of classical space-time physics before discussing the foundations of general relativity. The book contains a nice account of the problem of Kant’s concerns over incongruent counterparts, as well as presenting the infamous ‘hole argument’ against substantivalist views of space-time, although it is now out of date with regard to recent responses to this latter argument, in particular, and the development of so-called ‘sophisticated substantivalism’.


A classic in the field, and deservedly so, this offers a technically and philosophically sophisticated clarification of the foundations of space-time theories from Newton through to Einstein. Friedman begins with the neo-Kantian origins of conventionalism in the philosophy of space-time, before analysing the conceptual foundations of Newtonian kinematics, Newtonian gravitational theory, and the special and general theories of relativity.


For the more advanced reader, this collection includes essays on such standard topics as the foundations of quantum physics, including quantum information theory, and space-time physics, as well as less-well-covered subjects such as classical and quantum statistical mechanics, cosmology and quantum gravity. Written by some of the leading specialists in the field (as well as the odd Nobel prize winner), this lays out where to find the cutting edge.

A huge (and expensive) encyclopaedia of concepts, debates and issues in the foundations of quantum physics, this collection includes pieces by physicists as well as historians and philosophers. This is a very useful ‘first place to look’, before plunging further into the literature.


This is a good introductory text, and the initial chapters provide a useful grounding of the technical basics for those with little or no background in physics. Part II covers issues to do with the interpretation of quantum mechanics, including the EPR and Kochen-Specker arguments and, uncommonly perhaps, a chapter on quantum logic. The book concludes with Hughes’ own ‘quantum event interpretation’, which has not been widely adopted.


An engaging text that uses the notion of ‘locality’ as the hook on which to hang discussions of causality, unification, explanation and the nature of fields. The bulk of the book is taken up with classical field theory, with relativity and quantum mechanics covered in the final three chapters, but as Lange points out, there are philosophical puzzles aplenty within the classical realm. The price for this classical focus is a lack of discussion of Everett, decoherence and some of the more exciting recent topics in quantum mechanics.


As the title suggests, Maudlin focuses on the peculiar non-locality we find in quantum physics and examines its apparent conflict with the principles of special relativity. In doing so, he considers a number of issues in the philosophy of physics, and with the technical details kept to a minimum, this serves as a useful introductory text.


This is a volume of advanced essays that collectively defend the fundamental idea that metaphysics should content itself with reflecting upon physics. They cover the nature of scientific laws, causation and the passing of time and are linked by their rejection of the Humean ontological picture that takes what
there is to be nothing more than a collection of local qualities structured by spatio-temporal relations.


Written by a leading expert on the history and philosophy of relativity theory, this e-book is based on Norton’s course lectures that in turn are aimed at those who know little or no physics. It covers the basics of special relativity, general relativity and quantum mechanics, offers enlightening historical perspectives on these theories and takes the reader gently through such issues as the ‘Twins Paradox’, the import of E=mc², the nature of non-Euclidean geometry, big bang cosmology and the measurement problem in quantum physics, and ends with a handy overview of Einstein’s scientific career.


A reasonably accessible text that not only covers the basics, such as EPR, Bell, Bohm and so forth, but also gives a useful introduction to the coherent histories programme and, in particular, decoherence, which now plays such a significant role in our understanding of the foundations of quantum physics.


A great collection of papers nicely focused on the important question, is there causation in the (fundamental) physics, and if there isn’t, what’s the status of our causal notions? The title derives from Russell’s early scepticism regarding causal fundamentalism, but despite this venerable history, the issue remains very much unresolved: while philosophers often operate with the assumption that causation is a fundamental aspect of the world, it turns out to be difficult to find causation in the world as described by the fundamental physics.


Beginning with an introductory essay on the nature of the philosophy of physics, this includes contemporary papers on the philosophy of quantum mechanics, the foundations of statistical mechanics, philosophical aspects of quantum information theory and quantum gravity, written by some of the leading researchers in these areas. Technically demanding, this offers a state-of-the-art overview of the subject.
Annotated Bibliography


A classic introductory text that not only covers the standard topics in the foundations of space-time and quantum mechanics, but also looks at statistical mechanics, which has become a growing field of interest for philosophers. And all without equations, although at times that can seem more of a hindrance than a help. Nevertheless, this remains the standard by which other introductory efforts in this area are judged.


Statistical mechanics has long been the ‘Cinderella’ of topics within the philosophy of physics and, despite the efforts of Sklar in this work, it remained so until recently. Here he offers a useful introductory chapter on the foundations of statistical mechanics, followed by a detailed historical ‘sketch’ taking the reader through the work of Maxwell, Boltzmann, Gibbs and the Ehrenfests, before tackling a range of issues such as irreversibility and cosmology, the reduction of thermodynamics (of course) and the direction of time. There are also helpful chapters on the foundations of probability theory and the nature of statistical explanation. This is a lucid and clearly written classic in the field.
Research Resources

Societies

The three leading societies in the philosophy of science are:

British Society for the Philosophy of Science
www.thebsps.org

The BSPS publishes, through Oxford University Press, the British Journal for the Philosophy of Science, one of the world’s leading journals in the field. It holds an annual conference (at the same location as the ‘Joint Session’ of the Aristotelian Society and Mind Association) at which graduate students, in particular, are encouraged to present papers. It also sponsors a range of conferences and workshops around the United Kingdom on particular themes and awards a scholarship for PhD study in philosophy of science at a UK university.

European Society for the Philosophy of Science
www.epsa.ac.at

A relative newcomer, EPSA aims to bring together philosophers of science from across Europe (broadly defined). It holds a biennial conference (in odd years) with both contributed and invited papers. It also publishes, through Springer, the European Journal for the Philosophy of Science.

Philosophy of Science Association
www.philsci.org

Based in the United States, the PSA also includes non-US-based members and publishes, through Chicago University Press, the journal Philosophy of Science, another of the leading journals in the field (to publish in the journal you have to be a member of the PSA). It holds a biennial conference (in even years and usually in association with the History of Science Society – see below) with both contributed papers and symposia on significant topics. It also runs an email list that carries announcements of relevant conferences, grant awards, scholarships and so forth; see: http://philsci.org/mailinglist.html.
Other societies of possible interest include:

**Australasian Association for the History, Philosophy and Social Studies of Science**
www.usyd.edu.au/aaahpsss

This is the Australasian association of historians, philosophers and social scientists studying science, technology and medicine. It organizes an annual conference and publishes *Metascience*, a journal of book reviews in history and philosophy of science, through Springer:


**British Society for the History of Science**
http://bshs.org.uk/

The BSHS aims to ‘bring together’ people working in the histories of science, technology and medicine. It publishes (via Cambridge University Press) the *British Journal for the History of Science*, one of the ‘top’ journals in the subject:

http://journals.cambridge.org/action/displayJournal?jid=BJH.

It also publishes ‘Viewpoint’, a useful quarterly newsletter:

http://bshs.org.uk/publications/viewpoint

**Canadian Society for the History and Philosophy of Science**
www.yorku.ca/cshps1

This Canadian society brings historians and philosophers of science together, as well as sociologists and interdisciplinary scholars interested in all aspects of science. The society holds a conference annually.

**History of Science Society**
www.hssonline.org

Dedicated to promoting research in the history of science, the HSS holds its biennial conference in conjunction with the PSA. It publishes *Isis*, one of the leading journals in the history of science:

www.journals.uchicago.edu/loi/isis.
Apart from the papers, it also runs useful reviews of books in the philosophy of science and a newsletter:

http://www.hssonline.org/publications/newsletter.html

**HOPOS**
www.hopos.org
As the name suggests, the International Society for the History of the Philosophy of Science promotes research on the history of the subject. It holds a biennial conference (in even years) and has recently started to publish a journal in this area, through University of Chicago Press, *Hopos*:

www.journals.uchicago.edu/toc/hopos/current.

It also runs an active electronic list (‘Hopos-L’) that not only carries announcements of conferences and other events, but also promotes relevant discussion in this area; see:

www.hopos.org/listserv.html.

**International Union of History and Philosophy of Science**
This includes a Division of Logic, Methodology and Philosophy of Science that aims to promote interrelations between logicians, scientists and philosophers of science:

www.dlmps.org.

It organizes an international congress every four years and publishes the proceedings.

**Society for Philosophy of Science in Practice**
www.utwente.nl/gw/spsp

Launched in 2007, this is a relatively recently formed society that aims to shift the emphasis more towards detailed and systematic study of scientific practice. It also covers philosophy of technology and philosophy of medicine and has organized a conference biennially since its conception.

**Society for Social Studies of Science**
http://www.4sonline.org/
With a focus on the social context of science, the ‘4S’ holds an annual conference, and members receive the journal *Science, Society and Human Values* free of charge:

http://sth.sagepub.com/

**Conferences**

There are some recurring conferences *in addition* to the ones listed above in connection with the various societies.

**&HPS**

This conference on ‘Integrated History and Philosophy of Science’ has been held biennially since 2007, focusing on scholarship that is genuinely both historical and philosophical.

**Boston Colloquium for Philosophy of Science**

The Center for Philosophy and History of Science holds a well-known annual Colloquium. It’s free, open to the public, and in 2010 it celebrated its fiftieth anniversary!

**Boulder Conference on the History and Philosophy of Science**

The University of Colorado, Boulder, holds this well-established annual event.

**UK and European Meeting on the Foundations of Physics**

An (almost) annual foundations of physics conference usually held somewhere in the United Kingdom.

**UWO Foundations of Physics**

The University of Western Ontario holds an annual conference in the Foundations of Physics.

**Internet**

**EpistemeLinks**

A useful set of webpages presenting a range of philosophical resources on the internet, such as blogs, e-texts and even course materials from various departments – including a number in philosophy of science.

Internet Encyclopaedia of Philosophy
www.iep.utm.edu
While not as extensive as its Stanford cousin, this nevertheless includes some useful entries in the philosophy of science.

Intute: History and Philosophy of Science
www.intute.ac.uk/hps
The UK-based Resource Discovery Network (RDN) aims to provide an up-to-date directory of internet sites, containing a directory for HPS.

It’s Only a Theory Blog
http://itisonlyatheory.blogspot.com/
A group blog devoted to the general philosophy of science, with links to the philosophy-of-science-related blogosphere.

Mersenne
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=Mersenne
An electronic list that promotes informal discussion of topics in science studies, but tends to focus on the history of science.

Philos-l
http://listserv.liv.ac.uk/archives/philos-l.html
A UK/Europe-based electronic list that primarily posts announcements of awards, conferences and so forth, including those in philosophy of science.

Philosop
http://www.louisiana.edu/Academic/LiberalArts/PHIL/philosop.html
The US-based equivalent to Philos-l; again, it includes announcements of interest to philosophers of science.
PhilPapers
http://philpapers.org/
A comprehensive directory of online articles and books by academic philosophers, including (of course) philosophy of science. Monitors leading journals and holds archives and personal pages.

PhilSci Archive
http://philsci-archive.pitt.edu/
An electronic archive for preprints, serving specifically the philosophy of science community, in order to promote rapid dissemination of new work. This is the place to look for the most up-to-date work in the field.

PSA Listings
www.philsci.org/resources
A US-based subject directory, hosted by the Philosophy of Science Association, covering graduate programs, job openings, funding sources, journals and so forth.

PRS Subject Centre
www.prs.heacademy.ac.uk/hps
Run by the UK Higher Education Academy, the PRS Subject Centre offers resources for teaching support in Philosophical and Religious Studies in higher education in the United Kingdom. There is a specific section for history and philosophy of science.

Stanford Encyclopaedia of Philosophy
http://plato.stanford.edu/
An invaluable on-line resource with entries written by leading experts in their field. Topics included run the gamut from abduction to zombies (!), and many of the entries in both general philosophy of science and the philosophies of particular sciences have extensive reading lists.
ABDUCTION (a.k.a. inference to the best explanation)

A method of reasoning whereby one moves from a given phenomenon, or a set of phenomena, to the best explanation for those phenomena. Abduction moves in the opposite direction to deduction, where one starts with some assumption(s) and deduces what follows. With abduction, one starts with a certain explanandum (something to be explained), and infers a likely explanans (something explaining it). Usually it will be expected that one can then move in the other direction and deduce the explanandum (or a high probability for its occurrence) from the explanans.

Some even construe reasoning by straight induction as a type of abduction: for example, one can move from an observed regularity (e.g. all stones dropped thus far have fallen to earth) to a universal law (all stones fall to earth). The latter, if true, would (arguably) explain the former.

Precisely how to construe and model abductive reasoning is an open question. Attempts to formalize this have not been successful. Instead we are left to draw on paradigm examples and various rules of thumb. For an excellent overview of these issues, see Lipton, P. (2004), *Inference to the Best Explanation* (2nd edn). London: Routledge.

See also: EXPLANATION, HYPOTHETICO-DEDUCTIVE METHOD, INDUCTION.

BAYESIANISM

Bayesianism in philosophy of science is a research programme that attempts to capture concepts of evidence and confirmation in probabilistic terms. Subjective Bayesianism is based on the assumption that probability theory can give an appropriate formal representation of the notion of a degree of belief (in a theory or hypothesis) and its revision (or updating) according to new evidence. Of course one first has to accept that our attitude towards hypotheses can be described in terms of degrees of belief, and that it makes sense, and is
possible, to use probabilities to capture these degrees. With this given, the Bayesian makes use of the following theorem from probability theory:

\[ P(h/e) = \frac{P(e/h) P(h)}{P(e)} \]

Here \( P(h) \) stands for the probability of \( h \), \( P(h/e) \) stands for the conditional probability of \( h \) given \( e \), where \( e \) stands for a piece of evidence, and \( h \) stands for a particular hypothesis or theory. If \( P(h/e) > P(h) \) then the evidence is said to confirm the theory. The amount that the evidence confirms the theory can be quantified by looking at the difference between the prior probability of \( h \), \( P(h) \), and the posterior probability of \( h \) given the new evidence \( e \), \( P(h/e) \). Similarly, if \( P(h/e) < P(h) \) then the evidence is taken to disconfirm the theory.

One objection to the Bayesian programme is that it describes how logically omniscient agents should reason, and thus does not apply to human beings. Another well-known objection, the ‘problem of old evidence’, is that the Bayesian formula tells us that a theory’s ability to predict ‘old’ evidence – that is, observations known before the theory came about – should not increase our degree of belief in the theory at all, when it seems clear that it should. Bayesians have various responses to such challenges, and despite ongoing disputes, the Bayesian approach is widely accepted as the best way to formalize how scientific beliefs should be updated in the face of evidence. A classic text is: Earman, J. (1992), Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory. Cambridge, MA: MIT Press.

See also: CONFIRMATION, INDUCTION.

**CAUSATION**

The concept of causation is ubiquitous in science (although not all scientific disciplines operate in unequivocal causal terms). Causation is closely linked to explanation, since often a request for a scientific explanation of a given phenomenon can be read as the question ‘What caused the phenomenon?’ The fact that there is only natural, but no logical, necessity to causal connections, together with the fact that we have no ‘direct access’ to causation, but merely observe one event regularly followed by another, gives rise to empiricist scepticism about causation. One may retort that similarly we have no ‘direct access’ to many unobservable features of the world, say electrons, but we can nevertheless justify our theoretical conception of them.

There are various analyses of the concept of cause in the literature. For example, it is often assumed that \( A \) causes \( B \) if and only if it is the case that if \( A \) had not occurred, \( B \) would not have occurred. All counterfactual analyses of
`cause` from such starting points have been challenged. Indeed, even finding necessary conditions for causation has proved difficult. For example, it is debatable whether A needs to happen before B in order to cause B: there are examples of A occurring at the same time as B, or even after B in scenarios of backwards causation. In addition, the world does not appear to be fully deterministic (as the quantum theory tells us, for example) and there are imperfect regularities where it is not invariably the case that if A had not occurred, B would not have occurred. Such regularities can be accommodated as causal connections by probabilistic accounts of causation, according to which causes merely change the probability of their effects. For a recent textbook covering different theories of causation, see: Psillos, S. (2002), *Causation and Explanation*. Durham: Acumen.

See also: EXPLANATION, LAWS.

**CETERIS PARIBUS CLAUSE**

Most law statements in science are not simple universal regularities without any exceptions. Rather, in science (outside fundamental physics, at least) most law statements are hedged by a ceteris paribus clause. This can be interpreted as ‘other things being equal’, or (perhaps better) ‘when nothing else is interfering’. By adding such a clause to a law statement, one can ignore the exceptions in referring to a regularity that isn’t universal. Alternatively, one can attempt to say explicitly what the exceptions are. But it has been emphasized that usually in science it is impossible to do the latter.

Many philosophical issues arise in this connection. For example, the content of a statement of a ceteris paribus law threatens to be vacuous, if it just says that ‘A regularly follows from B, if there’s nothing interfering that might stop B following from A.’ And how could we falsify it by performing an experiment, since we don’t know how to make everything else ‘equal’ in order to test the law in isolation, with its ceteris paribus clause removed? Can statements hedged by a ceteris paribus clause refer to genuine laws of nature at all, and if so, how should we make sense of the clause? There are various responses to such issues in the literature. For a useful selection of recent articles, see: Earman, J., Glymour, C. and Mitchell, S. (eds), (2003), *Ceteris Paribus Laws*. Dordrecht: Springer.

See also: CONFIRMATION, FALSIFICATION, LAWS.

**CONFIRMATION**

What does it take to have evidence for some theory or hypothesis? What is the relationship between a statement that describes the evidence, and a statement
that describes the hypothesis? According to a simple picture of confirmation, a theory can be confirmed (to a degree) as follows: a prediction is derived from the theory, and that prediction is corroborated by an experiment. It is intuitive to say that the experiment tests the theory, and our degree of belief in the theory should increase since the theory has passed the test.

This simple picture has generated a lot of debate in the twentieth century. Some classic paradoxes of confirmation (Goodman’s Paradox and Hempel’s ravens paradox) have challenged accounts of confirmation along these simplistic lines. There are also obvious outstanding questions: for example, how much should our degree of belief in a theory increase in a given case? It has been suggested that this will depend on how accurate the prediction is and how surprising the prediction is, for instance. And should a theory only be considered highly confirmed when it has provided successful predictions in a wide variety of contexts?

In an effort to avoid such difficulties, Popper argued that science is not a confirming activity, but a falsifying activity. For him, when a theory makes a successful prediction, that theory is corroborated, which only means that the theory has passed that test, and doesn’t mean that our degree of belief in it should increase. However, Popper’s views on this issue are not popular. Today, Bayesian confirmation theory is perhaps the most favoured approach.

See also: BAYESIANISM, EVIDENCE, FALSIFICATION, HYPOTHETICO-DEDUCTIVE METHOD, INDUCTION.

COVERING LAW MODEL

The covering law model is a now universally rejected account of scientific explanation that was popular throughout much of the second half of the twentieth century. The basic idea is that to explain a given phenomenon (e.g. a solar eclipse) is to provide a set of accepted assumptions (e.g. Newton’s inverse square law of gravitational attraction, etc.) and then to derive the phenomenon from those assumptions. In every such derivation there should be at least one ‘law of nature’, and the explanandum (thing to be explained) should follow from the premises by deduction or induction. When deduction is used, the account is usually called the deductive-nomological (DN) model of explanation.

There are many objections to this account, but the most crushing are the numerous counter-examples that can be found in the literature. These show there are clear cases where (a) the covering law model applies, but we clearly don’t have an explanation, because the model doesn’t track the direction of causation, for example; and (b) we clearly do have an explanation, but the covering law model doesn’t apply, because the explanation doesn’t make an
explicit appeal to a law of nature, for example. Hence the model is neither sufficient nor necessary for something to count as an explanation.

See also: EXPLANATION, LAWS.

**CRUCIAL EXPERIMENT**

A crucial experiment is an experiment that is said to *conclusively* decide between two competing theories. Think of Eddington’s observations of starlight ‘bending’ in the sun’s gravitational field, for example, as deciding in favour of Einstein’s general relativity, and against Newton’s theory of gravity. The idea is simply that when two theories make contrary predictions about the world – so that only one of them can be true – then a crucial experiment can be carried out in order to reject the unsuccessful theory and uphold the successful theory. Many accounts of science appeal to this idea of crucial experiment, originating from Francis Bacon’s *Novum Organum*. This simplistic picture has been widely criticized, however, starting with Duhem, who enunciated the most critical problem following from what came to be called the ‘Duhem–Quine thesis’: it is not possible to test scientific theories in isolation, since deriving predictions from a theory requires background (‘auxiliary’) assumptions. Hence, even if one prediction is very successful and the other wide of the mark, we should still be wary of proclaiming the successful theory true, or proclaiming the unsuccessful theory false, since it could be the auxiliary assumptions that are to be blamed for the discrepancy. For a thorough discussion of experimental methodology see: Franklin, A. (1989), *The Neglect of Experiment*. Cambridge: Cambridge University Press.

See also: DUHEM–QUINE THESIS, FALSIFICATION.

**DEDUCTIVE-NOMOLOGICAL MODEL**

See: COVERING LAW MODEL.

**DEMARCATION PROBLEM**

The ‘demarcation problem’ concerns the issue of how to distinguish genuine science from non-science or pseudo-science. Once a core concern in the philosophy of science, it is now widely regarded as either intractable or irrelevant. Early efforts to draw the line between, say, astronomy and astrology relied on rather simplistic criteria based on the verifiability or falsifiability of the
statements concerned. With the decline of methodological frameworks based on such criteria, attention shifted to families of desiderata that science should satisfy, including explanatory and predictive power, problem solving ability, as well as testability. However, these are contentious and still not sufficient to draw an appropriately clear demarcation. As a result, a number of philosophers of science regard the demarcation problem itself as a ‘pseudo-problem’ and urge that attention be focused on issues such as the nature of empirical success, the role of novel predictions and so forth.


See also: FALSIFICATION, METHODOLOGY.

**DUHEM–QUINE THESIS**

According to the Duhem–Quine thesis, a hypothesis cannot be tested in isolation, since such testing requires auxiliary hypotheses about, for example, the equipment used in the test, or background assumptions about the relevant context. This implies that we cannot straightforwardly falsify a hypothesis, for example, because it may be that it is one of the auxiliary hypotheses that is at fault and should be rejected. More generally, this suggests that as far as the logic of disconfirmation is concerned, there will always be a way to maintain that a given hypothesis is true, and that apparent falsification can be directed elsewhere in our web of beliefs.

Equally, verification cannot then accrue univocally to a given hypothesis, suggesting a holistic view of theory confirmation, according to which it is the entire nexus of hypotheses and auxiliaries that are confirmed by verifying evidence. When Quine, drawing on Duhem, re-emphasized the thesis in the 1950s, it was seen to cause difficulties for the logical positivists. The latter had argued that the meaning of a sentence consisted in whatever observational consequences were to be associated with that sentence. But the Duhem–Quine thesis tells us that observations cannot be seen as following from any single sentence, but only from a whole family of them. Although this view of meaning has now been abandoned, the thesis remains an issue within the philosophy of science.

Of course, even granted the central claim, in practice auxiliary hypotheses and background assumptions are regarded as such precisely because they are better confirmed than the hypothesis being tested. Thus, even though it is logically possible that it is one of these auxiliaries that should take the force of the falsification or verification (and indeed, there are cases where this
possibility has turned out to be an actuality), the testing is typically understood to be directed at the hypothesis in question. In particular, when different experiments are performed that draw on distinct, but overlapping, sets of assumptions, there is plenty of scope for making sensible decisions about which assumptions are more likely to be blameworthy for a false prediction, and which are less likely. Popper called it a matter of ‘scientific instinct’; Duhem called it ‘good sense’.

See also: EMPIRICISM, FALSIFICATION.

EMERGENCE

Emergence is said to occur when certain entities or phenomena arise from more fundamental ones, but in a way that renders them ontologically irreducible to, or epistemologically inexplicable in terms of, the latter. This is sometimes summed up in the slogan ‘The whole is greater than the sum of its parts.’ The classic example here is the relation between the brain and the mind: arguably the qualitative character of the colours we perceive cannot (possibly) be explained in terms of the underlying physical processes. Another possible example here comes from quantum physics: the behaviour of a pair of entangled particles cannot be explained purely in terms of the properties of the individual particles. Accounts of both epistemological and metaphysical types of emergence remain problematic, however. For useful discussions of these and related issues see: Bedau, M. and Humphreys, P. (eds) (2008), Emergence: Contemporary Readings in Philosophy and Science. Cambridge, MA: The MIT Press.

See also: REDUCTION, SUPERVENIENCE.

EMPIRICISM

Empiricists believe that scientific knowledge arises from evidence ultimately based on sense data. Observations, on this view, exclusively involve the use of the senses, and so instruments such as electron microscopes, scintillation screens, cloud chambers and the like, introduce theoretical features that undermine any claim to have ‘observed’ the entity concerned. On earlier forms of empiricism (as incorporated into logical empiricism) theories themselves were regarded either as reducible to sets of observations or as mere instruments, useful for making observable predictions. More recent versions, such as ‘constructive empiricism’, understand theorizing in terms of the construction of models that are empirically adequate, but cannot be regarded as true. On this view, there is nothing wrong with having a theory which talks about electrons, so long as one doesn’t thereby think that implies that electrons exist. They might exist, but the evidence might be, and typically is, compatible with a range of theoretical
alternatives. What is important, on either the older or more recent forms of empiricism, is the evidence, its grounding in sense data and, in turn, the way it grounds the empirical success of theories. For an advanced understanding of this set of positions, see: Richardson, A. and Uebel, T. (eds) (2007), The Cambridge Companion to Logical Empiricism. Cambridge: Cambridge University Press.

See also: LOGICAL POSITIVISM, EVIDENCE, REALISM.

EVIDENCE

Evidence is commonly taken to be the ‘bedrock’ of science. However, what counts as evidence is problematic: here we see not only differences between scientists and philosophers of science, but also between members of the latter community. For certain empiricists, evidence consists in sense data; for realists, it may embrace observations mediated by complex instruments; for some philosophers, it is expressed via observation statements; while for those of a Bayesian persuasion, evidence is identified with certain rational changes in beliefs. On the last view, E counts as evidence for a hypothesis or theory if and only if E increases our degree of belief in the theory. Here we can discern a further distinction between different conceptions of evidence: some hold that E can count as evidence for T for an agent, given their particular epistemic situation, even when others think that E in fact isn’t evidence for T, and if an agent thinks it is he or she is mistaken. On the former understanding, but not the latter, what counts as evidence is relative to one’s particular epistemic situation.

Most theories of evidence that have been put forward focus on probability measures. For example, some say that E is evidence for T if the probability of T is increased by the truth of E. Such theories have been criticized for being too weak, since E can count as evidence for T even if Ts probability is only marginally increased. But the question then becomes how much the probability has to increase for E to count as evidence, and how that increase is measured. This is very much an area of open debate.

See also: CONFIRMATION, EMPIRICISM, EXPLANATION.

EXPERIMENT

It has been argued that discussions of scientific evidence are too focused on observation, whereas for modern science it is experimentation – often under highly artificial conditions – that provides much of the evidence for our theories. An experiment is an intentional manipulation of the relevant environment in order to test the predictions of a theory or, sometimes, just to see what happens. What counts as a good experiment and how one deals,
philosophically, with the errors that come with experimentation are matters for debate.

In the second half of the twentieth century, it became clear that experiments cannot unambiguously test theories, but require careful analysis in order to ascertain what ought to be concluded. Two reasons for this can be distinguished. First, experiments make use of a good deal of theory in their construction, and so can only be trusted as far as this theoretical background can be trusted. Second, one needs to distinguish between the data produced by experiments, and the phenomena constructed from the data using some degree of theorizing. These and similar issues have given rise to an ongoing research programme within philosophy of science known as the New Experimentalism. A good representative example of the latter programme is: Hacking, I. (1983), Representing and Intervening. Cambridge: Cambridge University Press.

See also: CRUCIAL EXPERIMENT, DUHEM–QUINE THESIS.

EXPLANATION

An explanation tells us how or why something came to be, or why something is the way it is. The most popular account of explanation for many years was the covering law model: roughly speaking, $A$ explains $B$ if and only if $A$ contains a law and $B$ deductively follows from $A$. However, this allows for situations in which the intuitive explanatory order is reversed: consider the shadow cast by a flagpole, for example. Although with some simple measurements and trigonometry we can deduce the height of the flagpole from the length of the shadow, we typically balk at claiming that the length explains the height.

Such problems have led to alternative theories of explanation being proposed, including the statistical relevance model, the unificationist account, and, perhaps most well-known, the causal account. The core idea of this last view is that we explain an event by citing earlier events that caused it. Since the causal history of any event is potentially infinite – tracing all the way back to the Big Bang – it is typically furthermore required that an explanation cites only earlier events that made a difference to the occurrence of the event to be explained, where the latter is construed in suitable contrastive terms. For example, we don’t just explain ‘Why did Adam eat the apple?’ but rather (say) ‘Why did Adam eat the apple instead of a kiwi?’ Articulating the idea of difference-making, responding to counter-examples, and showing how it applies to the more complex cases that arise in scientific practice remain as ongoing projects, as does the development of alternative accounts.

See also: ABDUCTION, COVERING LAW MODEL.

FALSIFICATION

Following Popper, falsificationists believe that at the heart of scientific methodology is, and should be, the attempt to falsify and subsequently revise scientific hypotheses. Furthermore, it is by means of the notion of falsifiability that science (such as astronomy, for example) can be demarcated from pseudo-science (e.g. astrology): scientific claims are genuinely falsifiable, whereas the claims of pseudo-science are not. If a genuinely scientific theory makes a prediction, an experiment is carried out, and the prediction proves false, then the theory has been falsified and should be rejected. But pseudo-scientific hypotheses arguably either never make any genuine predictions, or the blame is shifted to something other than the theory that is intuitively responsible for the false prediction; in either case, the hypothesis is kept immune from falsification.

This notion is crucial for the Popperian methodology of ‘conjectures and refutations’: bold conjectures are made and thrown to the wolves of experiment; if they pass the test, they are tested again and again; if they fail, they are deemed to have been falsified and must be rejected; new conjectures are then proposed. In this way, it is claimed, science learns from its mistakes and progresses. As compelling as this picture might initially appear, in practice, science does not proceed in this way. In some cases, hypotheses that are apparently falsified are not rejected because other evidence in their favour is too compelling, or there are doubts about the falsifying evidence, or the ‘blame’ for the falsification is shifted to one or other of the auxiliary hypotheses used in the testing. (See Duhem–Quine thesis.) Popper’s most well-known work setting out the central tenets of falsificationism is: Popper, K. (1959/2002), The Logic of Scientific Discovery. London: Routledge.

See also: CONFIRMATION, DEMARCATION PROBLEM, EXPERIMENT, HYPOTHETICO-DEDUCTIVE METHOD, DUHEM–QUINE THESIS.

HYPOTHETICO-DEDUCTIVE METHOD

The hypothetico-deductive (HD) method was, for many years, a dominant account of scientific methodology. It holds that science proceeds by (a) observing (related) phenomena in the world, (b) formulating a hypothesis
which would explain those phenomena, (c) deriving predictions via logical
deduction from the hypothesis, and (d) testing those predictions, and thus
the hypothesis, by experiment. Each of these constituent steps can be
criticized. With regard to step (a), it has been argued that scientists do not
simply observe phenomena, but actively create and group together phenom-
ena, drawing on some theory in some measure. On step (b), how scientists
‘formulate’ a hypothesis remains open to debate. Some advocates of the
HD method insist that this issue and that of the nature of scientific discovery
in general, lies beyond the philosophy of science. Others draw on the ‘heuris-
tics’ of scientific discovery to articulate the moves that scientists may make
when they formulate hypotheses. Still other appeal to some form of abduction,
but it remains unclear whether scientists should make such an
abductive leap (which they expect to be falsified) from very little data, or
whether they should make a very careful abduction from masses of collected
observations. Moving on to step (c) it’s clear that we don’t make derivations
from individual hypotheses, but also draw on various auxiliary assumptions.
Hence whether the test is positive or negative, it may remain unclear
which hypothesis should get the credit or the blame. Finally, with respect to
step (d), there are many issues surrounding how experiments should be
conducted, how error should be handled, and how the relevant conclusions
should be drawn from them.

See also: ABDUCTION, DUHEM–QUINE PROBLEM, EXPERIMENT,
EXPLANATION, FALSIFICATION, METHODOLOGY.

IDEALIZATION

Idealization is a phenomenon ubiquitous in the sciences, whereby assump-
tions are made which are believed to be false, and sometimes not even approx-
imately true, for the sake of utility. Several different kinds of idealization have
been distinguished, and there are numerous reasons why it can be beneficial.
For example, the world is often too complicated for scientific description, so
that scientists focus on selected causal relations which will never actually
occur, but which would occur (it is claimed) if other factors could be elimi-
nated. Another motivation comes from the application of mathematics to
science. Sometimes a theory asks us to carry out mathematics that is either too
complicated in practice (even with modern computers), or is in fact impossible.
The solution is often to simplify one’s theoretical claims in one way or
another, and work on the assumption that the conclusions will still be approxi-
mately true. The extent to which such an assumption is justified has been an
area of particular philosophical focus in the past 30 years. It has been argued,
for example, that the presence of idealizations means that our best physical
theories are not even candidates for truth, or are not even false, because they
do not attempt to represent the world.

See also: INSTRUMENTALISM, REALISM.

INCOMMENSURABILITY

In his analysis of scientific revolutions and theory change in general, Kuhn
claimed that those living before and after a scientific revolution are not able to
objectively compare their scientific theories. Such theories are said to be ‘incom-
mensurable’. The primary reason has to do with the meaning of theoretical
terms: if this is fixed by the relevant theoretical context, as seems plausible,
then when that context changes, as it does in a ‘paradigm shift’, the same word
may acquire a different meaning. A well-known example of this is ‘mass’
which, it is argued, has a different meaning in classical and relativistic physics.
Even if translation between the two theories is possible, there may remain dif-
fferences in the accepted rules of reason and argument, methods of experiment,
theories of evidence and even scientific values. Under such circumstances, not
only is it not possible for scientists on either side of such a ‘revolution’ to com-
municate, but scientific progress itself cannot be said to be cumulative, since
one cannot say that a given theory is ‘better’ than the one it replaced.

There have been various responses to these claims. The most significant,
perhaps, involves dropping the ‘contextual’ theory of meaning in favour of
the causal theory of reference. Many have pointed out that a proper under-
standing of past science is precisely what historians of science (such as Kuhn
himself) regularly achieve. This has led to more nuanced accounts of exactly
what incommensurability amounts to, and to the weaker claim that there will
always remain some degree of bias or miscommunication between scientists
on opposite sides of the fence. See: Kuhn, T. S. (1970/1962), The Structure of

See also: EVIDENCE, EXPERIMENT, METHODOLOGY, RELATIVISM.

INDUCTION

Induction is a form of ‘ampliative’ reasoning which takes us beyond what’s
implicit or explicit in the premises. It can be contrasted with deduction as
follows: in a valid deductive argument, the premises logically entail the
conclusion; thus, if (and it may be a big if) the premises are true, the conclu-
sion must be true. In a good inductive argument, the premises do not entail
the conclusion, but they do provide support for it; thus if the premises are
true, the conclusion may be true, where this ‘may be’ can be formalized using
probability theory, and the stronger the support given by the premises, the higher is the probability that the conclusion is true. One of the most well-known forms of induction is ‘enumerative’ induction, whereby one moves from numerous occurrences of X (without exception) to the assumption ‘always X’. Clearly, this form of induction – as any form of induction – is fallible: it is easy to think of examples where it leads to a false conclusion: for example, if one has seen numerous white swans (and no exceptions), and one inductively infers that all swans are white, one will be mistaken. Nevertheless, it can be argued that under appropriate conditions such reasoning can be justified, and induction arguably plays an important role in science. Consider the formulation of empirical laws, for example. Thus, we may observe that every massive body measured so far obeys the inverse square law of gravitation, and conclude on the basis of these observations that all massive bodies obey this law.

_Falsificationists_ deny that induction plays any such role, insisting that how such laws are obtained is not a matter with which the philosopher of science should be concerned, and when it comes to the all-important testing of hypotheses, only valid deductive reasoning is required. Certainly, many would agree that scientists do not (and should not) spend much of their time simply collecting theory-independent observations from which to generalize in an enumerative fashion. Instead, the observations themselves are informed by underlying theoretical assumptions, and scientists do not simply generalize from observations, but devise (explanatory) theories that appeal to unobserved or unobservable entities. Nevertheless, it is also recognized that there is more to scientific practice than falsifying hypotheses, and various forms of inductive reasoning can be discerned to play a role in the confirmation and support of hypotheses, as captured by the Bayesian programme, for example.

See also: ABDUCTION, BAYESIANISM, CONFIRMATION.

**INFERENC TO THE BEST EXPLANATION**

See: ABDUCTION.

**INSTRUMENTALISM**

Instrumentalism is an attitude that may be adopted towards scientific theories. It is a form of anti-realism that insists that our scientific theories are not _truth-apt_ – viz. not even the kind of things that can be true or false. Instead, scientific theories are to be thought of merely as instruments or conceptual tools for classifying data and generating explanations and predictions.
Instrumentalism is most famously associated with logical positivism, which held that the meaning of a theoretical claim can be entirely given in terms of the observational consequences of that claim, or, more generally, that the knowledge conveyed by scientific theories can be completely understood as knowledge about the observable.

Instrumentalism has fallen out of favour, however, and even modern-day empiricists recognize the utility of theories for more than just observable predictions. In particular, it was recognized that theoretical terms are not eliminable from theories in the ways that logical positivism suggested and that their role goes beyond contributing to classification and prediction; for example, theories in some sense explain phenomena. How that further role should be understood remains a matter for debate.

See also: EMPIRICISM, REALISM, THEORY.

LAWS

The presence of ‘laws’ is typically taken to be a crucial characteristic of science. However, how these are to be delineated from ‘principles’, ‘theorems’, ‘equations’ and so forth, that one also finds associated with scientific theories, remains a live issue in the philosophy of science. As a starting point, laws are standardly characterized in terms of universal generalizations that go beyond mere empirical regularities. However, some laws appear to admit exceptions. (Cf. ceteris paribus laws.) Some generalizations that look like laws, and have no exceptions, are not counted as laws. Some laws are deterministic, while others are statistical. And so on.

Also, there are distinctions to be drawn between laws of coexistence, laws of succession, and laws of interaction, for example. Obtaining a taxonomy of the different types of laws remains a challenge, as is the issue of providing a general account of their nature and role. With regard to the latter, there is considerable overlap with metaphysics as the debate has centred on whether laws involve an element of physical necessity (that is, they tell us what must happen or be the case), or whether laws should be understood as mere redescriptions of certain contingent empirical regularities.

A further concern is that philosophical discussion of laws has tended to focus on examples from physics, whereas laws feature less prominently in other sciences, such as biology, for example. Arguably, those ‘laws’ that are found in biology admit of exceptions – and hence more attention has been paid to the role of models, for example – whereas exceptions to physical laws are usually forbidden (at least at the fundamental level). However, even within physics, doubts have arisen as to the nature and status of laws, with some claiming that they are either false (because too simplistic, idealized or abstract),
or cannot be satisfactorily articulated and have to be understood as functioning with ceteris paribus clauses attached. Such difficulties have led some to suggest that the concept of ‘law of nature’ should be eliminated altogether, at least in the sense of a ‘physically necessary generalization’. For a recent account, see: Roberts, J. T. (2008), *The Law-Governed Universe*. Oxford: Oxford University Press.

See also: CETERIS PARIBUS CLAUSE, EXPLANATION, HYPOTHETICO-DEDUCTIVE METHOD.

**LOGICAL POSITIVISM**

Logical positivism was a dominant movement within the philosophy of science in the first half of the twentieth century, drawing heavily on empiricist thinking (especially as articulated by Mach and Duhem). Such an attitude gained momentum around the turn of the twentieth century, because scientific developments showed many of the claims – to do with space, time, causality and so forth – going beyond the ‘positively given’ experience of the senses to be mistaken. In response, the positivists attempted to provide secure grounds for science by broadly combining this ‘positivistic’ attitude with recent advances in logic to develop a picture of science as divided into observation statements, theoretical statements and ‘correspondence rules’ connecting these two. The main positivist ingredient was the idea that the theoretical statements were to be interpreted as nothing more than shorthand ways of referring to observation statements. Indeed, according to the verifiability theory of meaning that was widely adopted within the movement, the very meaning of theoretical statements, and the theoretical terms they included, was to be understood as entirely given in terms of observation statements. The correspondence rules had the job of relating the theoretical statements with the observation statements.

This programme held sway for many years, until cracks started to show in the 1950s and 1960s. A major problem concerned the distinction between observation and theoretical statements, and critics insisted that no such clear distinction could, in fact, be drawn. Indeed, it was argued that observations are ‘theory laden’: theory helps determine what is to be taken as relevant or salient about an observation; theory may also determine, at least in part, the meaning of observation statements; and theory, or more broadly, conceptual commitments, may shape or influence what it is that is taken to be observed. Furthermore, the above positivistic picture of science was recognized to be inadequate and in need of replacement by a more nuanced understanding of the relationship between theories and data in general. Finally, the association of logical positivism with the verifiability theory of meaning proved to be a major problem as the latter came under increased critical fire.
A consensus gradually emerged that logical positivism was inadequate for understanding how science both is and should be done. However, recent work in the history of the philosophy of science has both revealed a more nuanced perspective on this movement and helped to articulate those of its ideas that still have lasting value. A useful summary of the rise and fall of logical positivism can be found in: Suppe, F. (ed.) (1977), The Structure of Scientific Theories. Urbana, IL: University of Illinois Press.

See also: EMPIRICISM.

**METHODOLOGY**

Articulating the fundamentals of scientific methodology can be divided into *descriptive* and *normative* enterprises. The former considers the question of how science is done; the latter asks how it should be done. Although they took note of scientific practice, many philosophers of science in the first half of the twentieth century – such as those associated with logical positivism and falsificationism – were primarily concerned with the normative project. With the collapse of these movements in the 1960s, attention shifted to the issue of arriving at an adequate description of methodology as it actually features in scientific practice.

Numerous theories of scientific methodology have been put forward, the most prominent being the inductivist and verificationist proposals associated with logical positivism, the hypothetico-deductive method, Popper’s falsificationism, and Kuhn’s paradigm-based theory of scientific revolutions. Today, it is widely agreed that science is too complex to be easily captured in some simple philosophical slogan, and attention has become focused instead on issues to do with explanation, model-building, theory-data relationships and so on. Nevertheless, the philosophy of science has come under criticism for its lack of a normative stance and its consequent failure to engage with more public issues to do with the wider impact of science, ethics, the science-technology relationship and so on.

See also: DEMARCATION PROBLEM, FALSIFICATION, HYPOTHETICO-DEDUCTIVE METHOD, INCOMMENSURABILITY, INDUCTION, LOGICAL POSITIVISM, RELATIVISM.

**MODEL**

It has long been recognized that much of science involves building, testing and revising models, rather than theories themselves. What scientific models are and how they are best characterized are themselves philosophical issues.
Broadly speaking, they may be conceived as collections of theoretical assumptions which are intentionally simplified or idealized in some way, either to aid utility or because it isn't clear how to de-simplify them (or both). The aim is to make these simplifications and idealizations in such a way that they will not completely prevent the model from serving important scientific ends (such as explanation and prediction). Models are often seen as ‘mediating’ between high-level theory and descriptions of the phenomena, and as constituting one of the principal loci of scientific activity.

There are also many different types of model that feature in scientific practice. In addition to theoretical models (such as models of the atomic nucleus) there are mathematical models (e.g. sets of mathematical claims) and physical models (such as Crick and Watson’s famous tinplate and wire model of the structure of DNA). In addition, the term ‘model’ is used in logic, to describe a structure that makes a given set of statements true. This latter conception forms the basis of the view that scientific theories themselves should be conceived of in terms of families of such models, rather than as sets of statements. This engenders a move away from questions about truth and reference (language-questions) and towards questions of representation, for example.

It is claimed that not only can both theory–phenomena and theory–theory relationships be better captured via this ‘model-theoretic’ view, but so can all the different forms of model that can be found in science. However, it has also been argued that this approach affords too narrow a scope and that a broader conception of models and their role must be adopted. An accessible account that emphasizes the role of models can be found in: Giere, R. (1988), *Explaining Science: A Cognitive Approach*. Chicago, IL: University of Chicago Press.

See also: IDEALIZATION, THEORY.

**NATURALISM**

Naturalism is a very general movement within philosophy that broadly asserts the authority of the scientific method (whatever that is!) when it comes to investigating and understanding reality. In slogan form, it holds that there is no higher tribunal for truth than science itself. A weak version and a strong version can be distinguished: in the former, philosophical questions remain sensible questions, and one merely uses science to help answer them, whereas in the latter case the only sensible questions are scientific questions (and there is little or nothing for the philosopher to do). Not surprisingly perhaps, the strong form has not remained popular among philosophers, although many would describe themselves as naturalists of one form or another.

Within the philosophy of science, naturalism is associated with the view that the best way to both characterize and understand science is to use the
methods, techniques and frameworks of science itself. So, for example, instead of using logic to characterize theories as deductively closed sets of statements, we should use mathematical models, of the kind that scientists themselves use. Naturalism, in its strong form at least, can be criticized on the grounds of circularity: before the investigation begins, assumptions have been made about the very thing that is being investigated. However, many philosophers of science would respond that there is no place for a ‘first philosophy’ in the sense of some a priori methodology in terms of which scientific practice can be justified, and many would insist that many questions of the philosophy of science can be answered by making use of science in a non-circular way.

**NATURAL KIND**

Much of science seeks to make general claims about different kinds of thing, be it electrons, gold atoms, molecules, genes, primates or cases of multiple-personality disorder (MPD). Any such general claims implicitly suggest that the type of thing being talked about is a ‘natural kind’: that the specific individuals falling under that kind have an underlying similarity that is to some extent *objective* (in the sense of reflecting the way the world is, as opposed to being purely a matter of our subjective classificatory practices, say). In certain disciplines, we appear to find clear examples of such kinds: in chemistry, for example, there are the chemical elements as well as compounds such as water. In other areas, serious controversies arise: is there a real, objective similarity between all different cases of MPD, or are these cases a collection of quite different things we have merely grouped together with the term ‘MPD’? There are many genuine cases of the latter in the history of science: for example, ‘hysteria’ was a word used in the nineteenth century for a psychological condition which is now thought never to have existed (in any objective sense). Even chemical kinds have been scrutinized, with the complex composition of water suggesting that classifying it as a ‘kind’ may be too simplistic. Within biology, it has been claimed that the theory of evolution creates difficulties for those who view organisms as falling into natural kinds.

Perhaps the only uncontentious examples of kinds lie in fundamental physics, as represented by the elementary particle ‘zoo’. However, philosophical concerns remain, having to do with the connection between the notion of kind and various forms of the view that things have ‘essences’ of some kind, how these kind terms come to refer, and so on. Given these concerns and difficulties, some philosophers maintain that, in fact, there are no natural kinds, and that such divisions (but not necessarily the things being divided) are nothing more than human constructs. For a recent collection of essays, see: Beebee, H.
The Continuum Companion to the Philosophy of Science


**NO MIRACLES ARGUMENT**

The so-called ‘(no) miracles argument’ is an argument for scientific realism originating with Boyd and Putnam in the 1970s. It claims that our scientific theories are at least approximately true, because otherwise their ‘empirical success’ would be an inexplicable ‘miracle’. An immediate question arises as to how ‘empirical success’ is to be understood; nowadays much weight is placed on novel predictions, in particular. Although many find this intuitively pulling, critics have argued that it is not a good inductive (never mind a deductively valid) argument, and that in its characterization of empirical success, for example, it begs the question against the anti-realist. Despite various such attacks over the years, the general thrust of the argument is still widely accepted, if only as a motivation for realism.

The ‘pessimistic (meta-)induction’ attempts to undermine the foundations of the argument by pointing out that many theories we now believe to be false were nevertheless successful at some previous time in the history of science (a much discussed example is the caloric theory of heat).


See also: PESSIMISTIC META-INDUCTION, REALISM.

**PESSIMISTIC META-INDUCTION**

The pessimistic (meta-)induction is an argument against realism, to the effect that the history of science speaks against a realist attitude towards our current best theories. According to the ‘no miracles argument’ (see above), the main justification for realism is due to the empirical success of current science: our current scientific theories do an impressive job in explaining and predicting phenomena (particularly novel phenomena) in their target domains. But there are (apparently) numerous examples of past scientific theories that have been extremely successful and have turned out to be radically false. Lists of such examples have been given, but a prominent case is that of the caloric theory of heat, which apparently did a good job of predicting and explaining a range of thermal phenomena, including the regularities associated with the ideal gas law and the speed of sound.

Over the past 30 years two main strands of defence have emerged from the realist camp. First, it is claimed that many of the theories from the history
of science that are typically cited in this context were not, in fact, nearly as successful as our current scientific theories are, so they have no ‘inductive’ relevance for our attitude to current science. Second, it is claimed that when it comes to those theories that were sufficiently successful, one can separate off those elements or features of these theories responsible for this success and thus adopt a selective realist attitude towards them.

See also: INDUCTION, INSTRUMENTALISM, NO MIRACLES ARGUMENT, REALISM.

PSEUDO-SCIENCE

See: DEMARCATION PROBLEM.

REALISM

‘Scientific realism’, in its most general characterization, is the attitude that successful scientific theories represent the world more or less correctly in both its observable and unobservable aspects. If theories are regarded as sets of linguistic statements containing theoretical and observational terms, these terms may be described as referring to the relevant entities, processes and so forth in the world. If theories are characterized in terms of families of models, for example, a broader notion of representation may be adopted to describe their ‘faithful’ relationship with the world. In either case, such successful theories may be described as approximately or partially true, where ‘truth’ is typically understood in terms of some form of correspondence between the theory and the world.

A realist stance can be adopted towards different parts of our scientific theories: for example, the realist might say that the posited laws of nature are the true laws of nature (e.g. the inverse square law of gravity), or she might say that the entities referred to (e.g. electrons) really exist, and then she might say that the properties attributed to those entities (e.g. charge) really are instantiated in the world. However, many have suggested that scientific theories come as a package, and that there is no justification for picking and choosing which parts of a theory to be a realist about.

The form of realism adopted may also depend on the attitude adopted towards other issues, such as the nature of truth, or reference, or representation more generally. Realists may differ as to which features of theories one should be realist about: typically, realists have focused on the entities posited by theories, such as electrons, electromagnetic fields and so forth; structural realists, however, have suggested that it is the relevant structures – as characterized by
Maxwell’s equations of electromagnetism, for example – that we should be realists about; others have urged a minimal kind of realism that embraces a particular view of the properties that feature in scientific theories. Realists may also differ on their attitude to other issues that arise within the philosophy of science, such as explanation or reduction, and although many of the examples that are given are drawn from physics, interesting forms of realism can be adopted with regard to biology, chemistry, neuroscience and so on.

In recent decades, the debate between scientific realists and anti-realists has revolved around issues to do with the no miracles argument, on the one hand, and the pessimistic meta-induction and the underdetermination of theories by evidence, on the other. With the elaboration of various forms of realism, attention has also turned to their metaphysical underpinning and implications as well as associated epistemological issues. Perhaps the best articulation of the realist stance can be found in: Psillos, S. (1999), Scientific Realism: How Science Tracks Truth. London: Routledge.

See also: EMPIRICISM, INSTRUMENTALISM, NO MIRACLES ARGUMENT, PESSIMISTIC META-INDUCTION, REDUCTION, UNDERDETERMINATION.

REDUCTION

New theories are often said to relate to their predecessors via reduction. How reduction is understood typically differs between philosophers and scientists. The former take the older theory to reduce to the newer, whereas the latter conceive of the relation going the other way, with the new reducing to the old in some limit. On the philosophers’ conception, an older theory is said to ‘reduce to’ a newer theory when the latter explains why the former was successful, and at the same time goes beyond it. Thus (a) classical mechanics is said to reduce to special relativity, (b) Kepler’s laws of planetary motion reduce to Newton’s theory of gravity, and (c) thermodynamics reduces to statistical mechanics. On the scientists’ view, however, the newer theory reduces to the older one, in the limit: thus, special relativity is said to reduce to classical mechanics in the limit of velocities that are low compared to the speed of light.

Characterizing these reductive relations is problematic. Sometimes the older theory can be derived from the newer theory (case b) and sometimes not (cases a and c). And sometimes the newer theory genuinely replaces the older theory (case b), whereas sometimes this definitely is not the case (case c). And according to certain philosophies of science, notions of reduction and replacement are themselves problematic when it comes to scientific revolutions or ‘paradigm shifts’, because the relevant theories are ‘incommensurable’ with one another.
Furthermore, the apparent reduction of theories from one domain to those of another has fuelled claims that the relevant domains or more generally scientific disciplines themselves stand in such a relationship of reduction: thus the (purported) explanation of chemical bonding in terms of quantum physics led to claims that chemistry itself could be reduced to physics. However, the problems of characterizing and explicating such relationships between theories have led many philosophers of science to resist such claims between domains, disciplines, or ‘levels’ of scientific practice in general.

Nevertheless, many would agree that there are no chemical entities distinct from, or existing in addition to, those found in physics. Here reduction as a relation between theories can be contrasted with ontological reduction. This is a relation between objects or processes: the properties and behaviour of one entity (the chemical bond, say) is said to be fully determined by the properties and behaviour of its constituent entities and their relations (electrons and their quantum entanglement). Alternatively, some argue for cases of emergence for which, it is claimed, no such determination of the ‘higher’ level by the ‘lower’ level holds.

See also: EMERGENCE, EXPLANATION, INCOMMENSURABILITY, REALISM, THEORY.

RELATIVISM

Relativism is a family of views that hold that beliefs, moral stances, judgments and so forth are completely, or in part, relative to a particular context, which might be socio-economic, political or broadly cultural. In the context of the philosophy of science, relativists oppose what they see as naïve assumptions about the ‘objectivity’ of science, and they take the justification of scientific theories, their success and scientific progress in general to be determined more by historical, social, political, cultural or psychological factors than by how the world ‘really is’. Thus, certain feminist philosophers of science argue, for example, that gender is one such factor and it plays an important role in the articulation and justification of at least some theories in science. Consider the field of primatology, for example: it has been claimed that observations of chimp behaviour have been biased by the scientists’ projection of certain gender-based roles, leading to a distorted and inaccurate understanding of this behaviour. More generally, it has been argued that since science is predominantly a male enterprise, the conclusions are inevitably shaped by an underlying ‘male bias’, and must thus be considered to be relative to the gendered context.

Kuhn’s vision of science in terms of scientific revolutions as paradigm shifts between incommensurable theories has often been taken to support a
relativist view of science. The underlying idea is that scientists are inevitably entrenched in the paradigm in which they have been trained, and are unable to appreciate that their perspective is merely different, and not better, than other perspectives. Relativists also draw on the Duhem–Quine thesis to make their case, emphasizing that it is possible to hold onto any theory (or paradigm) one wants to, if one is willing to make changes elsewhere in one’s web of beliefs. The theory-ladenness of observation is also brought to bear: relativists claim that there is no unbiased way to make observations, since what is observed is always relative to the agent’s psychology and the context of observation.

Although such views have found support among sociologists and certain historians of science, they have been strongly criticized within the philosophy of science. It has been argued that even if such bias as indicated above exists in certain fields, it is difficult to see how it could shape the justification and development of theories in chemistry and physics, for example. Furthermore, within primatology itself, the bias of certain observers was subsequently revealed by further observations, and the shift in focus to characterizing and understanding experimental science has helped clarify how such bias, and error in general, can be handled. Nevertheless, efforts continue within the philosophy of science to accommodate social factors where relevant, while retaining an appropriately robust sense of objectivity.

See also: DUHEM–QUINE THESIS, INCOMMENSURABILITY, METHODOLOGY, UNDERDETERMINATION.

**REPRESENTATION**

With the claim that scientific theories and models represent the world – in the sense that the ‘picture’ of the world provided by our best theories is a close match to or resembles the world in itself – there has been considerable attention paid recently to the notion of representation itself. Much has been said on what might be meant by a ‘close match’ or ‘resembles’. An obvious way of articulating this claim is to say that theories are ‘similar’, in various respects and degrees, to the world. A more formal characterization can be given of the representation relation as an isomorphism (or some other morphism) holding between the structure of the theory or model and the ‘structure of the world’. This latter notion is often translated to mean the structure of the data-models devised from observation and experiment, in which case it is acknowledged that representation is no longer between theory and world, but between theory and more theory.

The philosophy of art has been drawn upon to both defend and criticize such views. In an effort to undermine similarity- or isomorphism-based accounts, examples of artworks have been given that do not appear to bear any
similarity to relevant aspects of the world, or for which the relations of representation are multiple or ambiguous. However, it is not clear how relevant such examples are when it comes to shedding light on representation in science. It has also been suggested that the representation relation can hold when there is no ‘close match’ at all, and that it is also possible to have a ‘close match’ without providing a representation. For example, a triangle on an Ordnance Survey map represents, but doesn’t resemble, a youth hostel, and two identical twins don’t represent each other. Those who are sympathetic to anti-realist accounts of science, in particular, have concluded that scientific theories and models may well represent the world without resembling it in any way. A recent account of representation in science is given in: van Fraassen, B. C. (2008), Scientific Representation: Paradox of Perspective. Oxford: Oxford University Press.

See also: EMPIRICISM, MODEL, THEORY.

SUPERVENIENCE

A general characterization of supervenience can be given roughly as follows: X is said to supervene on Y if there cannot be a change in X without a change in Y. More precise characterizations depend on how the modal ‘cannot be’ in the above is understood. So, for example, it is often said that the mind supervenes on the brain, because it is believed that there can be no change in one’s mental state without an underlying change in one’s brain state. The concept is sometimes invoked to help underpin a form of reductionism: the chemical properties of a system, for example, are said to supervene on the physical properties (there can be no change in the chemical properties without a change in the physical properties), and in this sense chemistry can be said to reduce to physics. This notion of reduction is rather weak, however, since supervenience – as characterized above – is a matter of covariance between groups of properties.

A topic of much discussion in recent years has been the status of ‘Humean supervenience’, sitting as it does at the intersection between metaphysics and the philosophy of physics. It was long assumed to be true, but seems very problematic in the face of modern physics (quantum mechanics, gauge theories, relativity theories). Humean supervenience says that, for any system, the properties of the whole supervene on the intrinsic properties of the fundamental constituents of the system and their spatio-temporal relations. However, quantum mechanics, for example, tells us that when two or more particles are in an ‘entangled’ state, the overall system exhibits properties which do not supervene on the properties of the particles taken separately. These ‘non-supervenient’ properties have been described as not reducing to the intrinsic properties of the individual particles and their spatio-temporal relations.
Whether Humean supervenience can be sustained in some form, and how the apparent non-supervenience exhibited in modern physics might be understood, remain as ongoing issues at the intersection between metaphysics and the philosophy of physics.

See also: EMERGENGE, REDUCTION.

THEORY

The question of what a scientific theory is, or (according to some) how it should be represented, took centre stage in philosophy of science in the mid-twentieth century. The logical positivists provided an unambiguous answer: a theory is a set of statements written in first order logic, and closed under logical deduction, with theoretical terms interpreted (via correspondence rules) in terms of an observation language. This formed a central feature of what became known as the ‘syntactic view’ of scientific theories. This picture was rejected in the 1960s and 1970s, and has since been replaced by alternative conceptions. Perhaps the most prominent is the so-called ‘semantic’ or ‘model-theoretic’ approach, which conceives of a theory as a family of models. Some philosophers of science prefer to focus on scientific models, whereas others focus on a more formal conception of ‘model’ borrowed from set theory: in the end, ‘the semantic approach’ covers a number of different approaches. Some have taken theories to consist of or be identified with such families of models; others insist that these models merely provide a convenient way to represent those things called theories that feature in scientific practice.

How theories are understood and represented bears, in turn, on a number of other relevant debates in the field, to do with the relationship between theories and models, how theories themselves represent (if at all), and the comparison between theories and cultural artefacts in general, such as paintings and pieces of music.

See also: EMPIRICISM, MODEL, LOGICAL POSITIVISM, REPRESENTATION.

UNDERDETERMINATION

The underdetermination of theories by evidence offers one of the most serious challenges to scientific realism, and it underpins forms of anti-realism, such as constructive empiricism. The challenge runs as follows: however much we observe and catalogue phenomena in the world, there will always be different theories compatible with those phenomena; on the basis of observations we will never be able to pick out one single theory as definitely correct or true.
One response has been to say that this is true only as a purely logical thesis, and in practice it is often clear that even if some theories ‘save’ the phenomena, they can be rejected by virtue of being clearly contrived, for example. Take the Ptolemaic view of the solar system – with planets orbiting the earth on epicycles – for instance: arguably it was able to capture the astronomers’ observations, but it surely looks contrived and implausible when compared with the simple, elegant solution provided by the heliocentric model.

A stronger version of the underdetermination objection states that there can always be alternative theories that also explain the phenomena and are equally attractive in terms of other theoretical virtues, such as simplicity and elegance. Or, alternatively, it is sometimes claimed that what looks simple and elegant can change over time, and depends on factors such as social and cultural context. Typically, realists respond to these stronger claims by pointing out that, in practice, it is incredibly hard to come up with even one theory that explains the phenomena, and that there are no serious alternatives to the so-called ‘Standard Model’ of elementary particle physics, for example. Of course, this doesn’t mean that there won’t be serious alternatives in the future, but empirical evidence will determine whether or not they are accepted over current theories, and hence, there will be no underdetermination. Nevertheless, anti-realists maintain that no matter how hard it is to come up with plausible cases of underdetermination in practice, the possibility of underdetermination raises serious concerns for the realist.

See also: CONSTRUCTIVE EMPIRICISM, INFERENCE TO THE BEST EXPLANATION, PESSIMISTIC META-INDUCTION, REALISM, RELATIVISM.

**UNIFICATION**

Unification refers to the bringing together, in one way or another, originally disparate scientific theories, equations, processes or entities. It has many different faces: for example, a very strong form of unification was Maxwell’s discovery in the nineteenth century that light is an electromagnetic wave. Thereafter, studies of light and electromagnetism could be seen as studies of the same thing. Maxwell’s theory of electromagnetism in itself was a triumph of unification: studies of electricity and magnetism were unified by the intimate relations between them. Since electricity and magnetism are intimately related but not identical, this is a slightly weaker case of unification, although still quite strong. A still weaker example of unification comes with recognizing that the solar wind and the aurora borealis are related, although many would not like to refer to this as a ‘proper’ case of unification, because the relation between them is quite weak.
Most scientists and philosophers see unification as a form of progress. For example, the ultimate goal for many theoretical physicists is to unify all of science in a ‘final theory of everything’. This, of course, assumes that all of science can, in principle, if not in practice, be reduced to fundamental physics. In recent years, there has been criticism of this goal, and of the notion of unification in general, with a consequent recognition of its limitations. Certain philosophers of science have even adopted a more extreme view that rejects the possibility of unification entirely, offering a vision of science as a mixed bag of different theories, suitable for a ‘dappled world’, where successful unification is not to be expected or pursued. A useful account of unification that draws on case studies from physics is: Morrison, M. (2000), Unifying Scientific Theories. Cambridge: Cambridge University Press.

See also: REDUCTION.
# Index

Abbott, L. F. 271, 288  
abduction 254, 312, 411–12, 422 See also inference to the best explanation  
Abrahamsen, A. 166, 176, 272, 287  
acceptance 86, 185, 255, 365  
adaptationism 247, 399  
Ainsworth, P. 83, 94  
Albert, D. Z. 236, 239, 244, 401  
Alexander’s dictum 338, 339  
analogy 28, 134, 141–2  
Anderson, C. H. 280, 289  
Anderson, E. 342, 355  
Anscombe, G. E. M. 117n 8, 118  
anti-realism 5, 6, 8, 10, 18, 27, 85, 87, 132–3, 330, 345, 367–8, 369–71, 385, 387, 396, 424, 430–1, 434, 436  
approximate truth, see truth, approximate  
approximation 97–8, 116, 189, 303, 304, 307, 311, 313, 349  
Aristotle 1, 68, 101, 125, 359–60, 374, 391, 400  
Aspray, W. 323, 324, 331  
Atkins, P. W. 304, 310  
atomism 84, 295, 313  
axiomatization 322, 392  
Azzouni, J. 327, 331  
Bacon, F. 15, 360–1, 374, 416  
Bachelard, G. 60, 68  
Bader, R. 302, 303, 310  
Bailer-Jones, D. 135  
Baker, A. 328, 331  
Balaguer, M. 322, 331  
Bandyopadhyay, P. S. 195  
Bangu, S. 328, 331, 351, 355  
Barker, M. J. 292  
Bartels, A. 129, 135  
Batterman, R. 135, 139, 144, 154, 329, 331, 340, 355, 383  
Bayes’ theorem 19, 22, 23, 179–81, 197, 202–208, 212  
and degrees of belief 5, 19–21, 181, 200, 213–24, 412, 413, 414, 419  
and prior probabilities 20–1, 180–1, 185, 202–5, 213, 215, 373, 413  
and problem of old evidence 213, 373, 413  
realistic interpretation of 205  
Subjectivist interpretation of 20, 200, 213, 214  
Beaney, M. 331  
Beatty, J. 174, 252, 263  
Bechtel, W. 143, 148, 149, 150, 152, 154, 166, 167, 176, 177, 260, 263, 269, 272, 283, 275, 278, 284, 286, 287  
Bedau, M. 138, 154, 383, 418  
belief  
degree of, see Bayesianism and degrees of belief  
justified 3, 20, 22–5, 27, 183–4  
web of 417, 434  

439
Bell, J. S. 234, 235–6, 244, 365, 372, 376, 404
Bell's theorem, see quantum mechanics
Benacerraf, P. 319, 320, 321, 322, 331
Bennett, K. 41, 52
Berson, J. 303, 310
Beurton, P. J. 253, 263
Bickle, J. 141, 144, 149, 150, 154, 284, 285, 286n, 287
Bird, A. 5, 7, 15, 28, 29, 30, 91, 248, 263, 296, 310, 342
Blackwell, R. J. 244
Blacksee, S. 290
Bloom, P. 275, 288
Bohr, N. 48, 233, 344
Boooos, G. 321, 331
Boyd, R. 84, 171, 176, 248, 249, 250, 251, 264, 287, 288, 367, 369, 376, 383, 391, 430
Boyer, P. 275, 288
Boyle’s law 159
Braddon-Mitchell, D. 35, 51n, 52
Brading, K. 351, 355, 401
Brandon, R. N. 260, 264
Bressler, S. L. 288
bridge laws 140–1, 143, 144, 145 see also reduction
Brigandt, I. 9, 139, 154, 246, 250, 251, 253, 254, 258, 259, 260, 264, 267, 292, 338, 345, 352
Brock, W. 301, 310
Brouwer, L. E. J. 317
Brown, H. R. 228, 244, 391, 401
Brunschvicg, L. 59
Bub, J. 277, 288
Buchdahl, G. 55, 56, 68
Bueno, O. 135, 330, 331, 358
Burgess, J. P. 322, 328, 331
Burian, R. M. 2, 11, 252, 263, 264
Buskes, C. J. J. 177
Butterfield, J. 347, 355, 356
Byron, J. M. 246, 264
Callender, C. 4, 5, 6, 10, 33, 45, 53, 134n, 135, 340, 341, 343, 345–6, 355
caloric 393, 430
Campbell, J. 110
Canguilhem, G. 56, 60, 68
Cao, T. 346, 355
Carroll, J. 138, 154, 351, 355, 384
Carruthers, P. 275, 288
Cassirer, E. 363, 375
Castellani, E. 351, 355, 401
Catterall, W. A. 272, 292
causal exclusion problem 338
causation 7, 96–116, 367, 385, 393, 400, 404, 413–14
causal mechanisms 100, 163, 173, 272, 279, 309
causal processes 161–2, 174, 280, 281
and counterfactuals 107–13, 162, 338, 350
efficient 101
and neuron diagrams 98–9
mental causation 399
interventionist account of 150, 153, 162, 338, 397
and powers 91, 92, 101, 138, 308, 339, 350, 351
and structural equations 96, 111–12, 118
Causey, R. L. 141, 154, 367, 377
ceteris paribus laws, see laws, ceteris paribus
Chakravartty, A. 88, 92, 94, 132, 135, 351, 355
Chalmers, D. 3, 11, 35, 41, 52, 53, 54, 357
Chang, H. 3, 11, 343, 385
Chemero, A. 273
440
Index

chemical bond 9, 300–3, 322, 352, 400, 433
Chihara, C. 322, 327, 331
Churchland, P. M. 136, 141, 151, 154
Churchland, P. S. 270, 277, 282, 288, 291, 385, 400
Ciufolini, I. 68n
Clark, P. 331
classical mechanics 42, 120, 122, 238, 347, 432
Coffa, A. 65, 68
Cohen, J. 45, 53, 134n, 135
Cohen, L. G. 288
Cohen, R. S. 31, 70, 94, 265, 266, 396
Collingwood, R. 167, 176
Collins, J. 118
Colyvan, M. 326–8, 330, 331, 332, 400
commitment, see acceptance
concepts 36, 39, 100, 147, 251–4, 261–2, 316, 339, 341, 397
Conceptual change 251–4
Conee, E. 40, 53
confirmation 16, 76, 77, 361, 414–15, 417–18
Bayesian theory of, see Bayesianism
bootstrap theory of 52n, 370
and discovery 254–7, 261
error-statistical theory of 255, 373, 391
holism 75–7, 326–7, 350 See also Duhem thesis
conservation laws, see laws, conservation
Contessa, G. 8, 120, 126, 130, 134n, 349, 353, 354
context of discovery, see discovery, distinction from justification
contradiction, see inconsistency
Corfield, D. 324, 332
correspondence rules 426, 432
Corry, R. 339, 356, 404
covering-law model, see explanation
Craig’s theorem 80, 392
creationism 46, 386, 399
Cummins, R. 165–6, 176, 280, 288
Cushing, J. 344, 356
da Costa, N. C. A. 93, 94, 133n, 134n, 135, 330, 332, 353, 356
Darden, L. 155, 166, 167, 176, 177, 256, 260, 264, 266, 279, 284, 288, 290
Daston, L. 56, 68
data
model of 123, 364, 434
and phenomena 370, 420
theory-ladenness of 17, 42, 434
Dayan, P. 271, 288
deductive-nomological – see explanation, deductive-nomological
De Finetti, B. 216n
Democritus 138
Demopoulos, W. 86, 94, 331
Denbigh, K. G. 298, 299, 310, 311
De Regt, H. 6, 9, 157, 162, 170–1, 176, 177
Descartes 25, 52n, 64, 68, 360, 374
Devitt, M. 251, 264
Dewey, J. 16, 64, 387
Dickson, M. 68n, 68, 69
Dieks, D. 162, 170–1, 176
DiSalle, R. 228, 244
discovery
and confirmation 254–7, 261
distinction from justification 55, 62–5, 364, 393, 422
disunity of science 259–60
Domski, M. 68n, 68
Dowe, P. 162, 176
Downes, S. 262, 264
Doyle, D. A. 272, 288
Dray, W. H. 167, 174, 176, 272, 288
Dretske, F. 281, 288
Duhem, P. 27, 30, 57–60, 68, 77, 84, 90, 94, 121, 136, 295, 311, 361, 366, 367, 374, 426
Duhem thesis 58–9, 61, 416, 417–18, 421, 434 See also confirmation holism
Dummett, M. 317, 331n, 332
Dunbar, K. 28, 31
Dunn, J. C. 276
Index

Dupré, J. 250, 251, 252, 259, 264, 311, 352, 356, 373, 374, 379
Dutch book argument 22
dynamicism 273–4
Earman, J. 115, 118, 228, 239, 244, 347, 355, 356, 370, 371, 373, 378, 379, 386, 402, 413, 414
eco 250, 398
economics 48, 51n, 61–2, 250, 294, 314, 325, 361–3, 373, 375, 379, 384, 390
Edgington, D. 44, 53
Edwards, A. W. F. 182, 195, 216n
Egan, F. 289
Einstein, A. 3, 16, 33, 53, 66, 68n, 235, 237, 244, 371, 391, 402, 404
Elga, A. 104, 118
Eliasmith, C. 280, 289
Ellis, B. 91
Ellis, G. F. R. 46, 53
emergence 138, 258, 383, 400, 418, 433
empirical adequacy 5, 46, 86, 169, 195n, 369
empirical success 89, 132, 187–8, 191–2, 341, 419, 430
constructive 85–7, 169, 372, 385, 386, 396, 436
logical 4, 18, 60, 133, 362–4, 369, 372, 390, 394, 418
and two dogmas 77, 79
Engels, F. 61, 69
epistemic dichotomy 87–9
EPR (Einstein-Podolsky-Rosen) correlations 162, 403
Ereshefsky, M. 249, 251, 252, 264
essentialism 36, 93, 248, 295, 305, 400
dispositional 91, 92, 339
ether 304
evidence
empirical 59, 182, 195n, 215, 319, 346, 366, 437
observational 183, 189
phenomenal conception of 183–5
and unification 186–91
evidential support 180, 182, 194, 197–8, 201–8
evolution, theory of 166, 247, 251–2, 256, 260, 295, 344, 347, 354, 367, 388, 398
exclusion principle 241–2
exemplar 17, 28
explanation 6, 9, 83–5, 90–3, 157–78, 300–6, 420–1
and understanding,
see understanding
causal 6, 10, 11, 150, 160–4, 173, 247, 282, 338, 369, 386, 395, 397, 413, 420
computational 280–3
covering law model 157–9, 173, 192, 415, 420
Deductive-Nomological 140, 158–9, 173, 192, 257–8, 366, 369, 415
functional 164–9, 173, 174, 247
inductive-statistical 159
how-possibly vs. how-actually 272, 274
kairetic 163–4
mathematical 325, 328, 350
pragmatics of 168–170, 171
reductive 140, 148–50, 173, 257–60, 262, 367
statistical relevance model 257, 420
simulacrum account of 172
unificationist 159–60, 175, 258, 325, 370, 391, 395, 420
externalism 21, 22–7, 342
Falk, R. 253, 263, 264
falsificationism 18, 364, 365, 366, 376, 385, 389, 393, 415, 417, 421, 424, 427
Faye, J. 170, 176
Feigl, H. 77–9, 85, 90, 94
feminist philosophy of science 342, 385, 389, 390, 398, 433
Feyerabend, P. 141, 154, 251, 264, 267, 368, 369, 385, 386, 396
Feynmann, R. 39, 116, 345
Index

Field, H. 321, 322, 327, 328, 332
Fine, A. 38, 53, 251, 265
Fine, K. 322, 332
Fitts, P. M. 287n, 289
Fodor, F. A. 262, 275, 281, 289
Fodor, J. 143, 144, 154
Forster, M. R. 7, 179, 182, 184, 187, 190, 192, 193, 195n, 195, 196
Frank, P. 61, 80, 94
Fraser, D. 346, 356
Frege, G. 315–17, 321, 331n, 332
French, S. 1, 93, 94, 129, 133n, 134n, 135, 330, 332, 337, 351, 353, 356, 357
Friedman, M. 4, 50, 53, 56, 65, 66–7, 68n, 69, 160, 177, 192, 196, 228, 244, 343, 370, 377, 387, 402
Frigg, R. 129, 134n, 136, 346, 355
functionalism 146
Galison, P. 56, 69, 370, 378
gauge symmetry 244, 401
Geertz, C. 175, 177
general relativity, see relativity
geometry
 Euclidean 49, 76, 221–3, 225–6, 325
 Minkowski 225–8
 non-Euclidean 49, 76, 223, 315, 362, 404
Georgopoulos, A. P. 289
Ghiselin, M. T. 248, 249, 265
Gillespie, R. 302, 311
Gillett, C. 150, 155
Glennan, S. 166, 177
Glymour, C. 52n, 53, 216n, 277–8, 289, 370, 377, 386
Gödel, K. 228, 318, 319, 320
Gödel’s incompleteness theorems 318, 331n
Godfrey-Smith, P. 134n, 136, 350, 356
Goldman, A. 23, 31
Goodman, N. 365, 367, 376, 377, 415
Goodwin, W. 301, 311
Graziano, M. S. A. 379, 389
Griffiths, D. J. 232, 245
Griffiths, P. E. 247, 250, 253, 265, 399
Grush, R. 279–82, 288, 289, 290
Haaland, A. D. 303, 311
Hacking, I. 58, 69, 182, 196, 256, 265, 370, 373, 378, 384, 385, 387, 388, 420
Haken, H. 273, 274, 289
Hale, B. 321, 331n
Hall, N. 7, 10, 96, 106, 112, 114, 117n, 118
Hanson, N. 18, 55–6, 85
Hardcastle, V. G. 278
Hardwig, J. 24
Harman, G. 21, 31
Hartmann, S. 20, 30, 176n
Hawley, K. 341, 356, 385
Hawthorne, J. 7, 197, 216
Heisenberg, W. 235, 363
Hellman, G. 322, 327, 332
Hempel, C. 77, 80–1, 139, 157–61, 164–5, 168, 172, 173, 190, 191–2, 257, 269, 272, 365, 368, 390, 415
Hesse, M. 121, 366
heuristics 256, 354
 and fertility/fruitfulness 256, 341
hidden variable (in QM) 235, 365
Hilbert, D. 314, 317–18
Hilbert space 229–35
Hille, B. 271, 289
Hirsch, E. 38, 41, 53
Hitchcock, C. 95, 117n, 175n, 388
Hodge, A. L. 270–2, 285, 289
Hofmann, J. R. 307, 311
Hohwy, J. 339, 356
hole argument 371, 402
holism
 theory 61–2
 and confirmation, see confirmation, holism
quantum mechanical 306
Holton, G. 33, 53
Hooker, C. A. 141, 155, 385
Howard, D. 4, 55, 57, 65, 67n, 68n, 69, 342, 387
Howson, C. 20, 31, 196, 216n, 217, 370, 386
Hubel, D. H. 280, 289
Hudson, H. 51n, 52n, 53
Huggett, N. 221
Hughes, R. I. G. 122, 126, 136, 373, 379, 403
Hume, D. 6, 11n, 32n, 360, 367, 374
Humean metaphysics 45, 90–3, 105–7, 117n, 393, 394, 396, 403
supervenience 435–6
Hume's principle 316
Hume's problem (of induction) 17, 18, 20, 360, 365
Humphreys, P. 138, 154, 192, 196, 374, 379, 383, 388
Huxley, A. F. 270–2, 289
Huygens, C. 226, 244
hypothetico-deductive (HD-) method 201, 254, 361, 366, 370, 421–2, 426
idealization 10, 30, 130–2, 135n, 306, 329, 331n, 349, 354, 388, 422–3
Galilean account of 130–1, 329
identity of indiscernibles 8, 242
incommensurability 66, 251, 253, 261, 267, 392, 423
inconsistency 349
indeterminacy 103, 109, 117n
quantum 233–4, 244
referential 296
indispensability argument 319, 326, 328, 349–50, 400
induction 249, 254–5, 412, 415, 423–4
and analogy 255–6
enumerative 16, 424
eliminative 207–8
forms of 423–4
justification of 255
logic of 254 see also logic of confirmation
material theory of 255
problem of see Hume's problem (of induction)
and support 159 see also evidential support
inference to the best explanation 5, 21–2, 91, 171, 342, 372, 412 See also abduction
instrumentalism 80–4, 188, 192, 233, 367, 424–5
interfield theories 260
internalism 3, 21, 22–7, 342
invariance 174, 366
Jackson, F. 139, 155
Jantzen, K. J. 275, 289
Jaworski, W. 144, 155
Jin, X. 347, 356
Jirsa, V. K. 375
Jones, M. 135n
justification 4, 15, 19–27, 190, 255–6, 319, 321, 326, 342, 364, 433
of belief, see belief, justified
context of 55, 60, 62–5, 256
Kallestrup, J. 339
Kandel, E. R. 285, 291
Kant, I. 4, 34, 41, 42, 65, 66, 314–15, 330n, 360–1, 362, 374
Kelley, T. 179
Kelso, J. A. S. 274, 275, 286n
Kepler, J. 42–3, 52n, 184, 360
Ketland, J. 86
Keynes, J. M. 362–3
Kim, J. 112, 143–8, 151, 308, 374
Kirsner, K. 276
Kleinhans, M. G. 174, 175n
Kline, M. 330n
Knobe, J. 37
Koch, C. 279–80
Kornblith, H. 25, 144, 287n
Koyrë, A. 56, 59, 66
Kripke, S. 35, 36, 44, 140, 145, 248, 294–5, 300, 338

444
Index

Ladyman, J. 3, 34, 52n, 53, 93, 134n, 351–2, 373, 389, 390
Lakatos, I. 17–8, 19, 31, 323–4, 346, 366, 369, 373, 376, 385, 392
Lambon R. M. A. 277
Landry, E. 136
LaPorte, J. 250, 265, 294, 295–6, 311
Latour, B. 367, 377
Laudan L. 263, 265, 345, 357, 368, 369, 371, 373, 377, 391, 392
and biology 257–9, 287n, 365, 398, 399, 425
ceteris paribus 373, 414 see also laws, of special science
conservation 3, 25, 59, 66, 162, 244, 339, 351
and counterfactuals 7, 52n, 102–7, 116, 163, 350–1
constructive vs. principle 237
and explanation 157–65, 172–4, 176n, 257–8, 272, 284–5, 366
fundamental 97, 102, 104–7, 113, 117n, 237, 308–9, 329
Humean account of 45, 90–3, 117n, 394, 396 see also Humean metaphysics
of metaphysics 46–8
phenomenological 384
and reduction 9, 138–143, 258–9, 308–9 see also reduction
of special science 110–11, 116, 143, 145
Laymon, R. 331n
Leeds, S. 52n
Leng, M. 350
Leplin, J. 367, 373
Le Poidevin, R. 307–8
Le Roy, E. 59
Leucippus 138
Lewis, G. N. 301–2
levels of organization 138, 151–3, 262–3, 275, 277, 284–6, 287n
likelihood 7, 179–82, 185, 189–90, 191, 194, 197–216
Lipton, P. 21–2, 171–2, 372, 386, 390, 412
Locke, J. 49, 360, 374
Locke’s Thesis 49, 53
Loewer, B. 106, 118, 239, 245
logic 116n, 142, 198–9, 256, 314, 321, 329, 361, 362, 363, 394, 426
deductive 198–9, 201, 213, 254–5, 256
first-order (predicate) 142, 319, 329
inductive 17, 19, 254 see also induction
modal 36
of confirmation 197–215, 365 see also confirmation
second-order 321
logical positivism, see positivism
Longino, H. E. 262, 372, 379
Looren de Jong, H. 173
Love, A. 139, 250, 251, 259, 260
Lowe, J. 36
MacBride, F. 327, 332
Mach, E. 57–8, 59, 60, 61, 66, 70, 80, 157, 361, 374, 426
Machamer, P. 148–9, 155, 166, 177, 195, 257, 266, 269, 271, 284, 286n, 289, 290, 357, 359, 380, 390
Machery, E. 339–40, 357
Maddy, P. 34, 36, 37, 51, 53, 324, 327, 332
McLaughlin, B. 309, 312
McLaughlin, P. 289, 290
McClelland, J. L. 290
McMullin, E. 56, 70, 130, 135n, 136, 329, 332
McShea, D. 345, 357, 398
Malament, D. 327, 332
Mancosu, P. 324–5, 332
Mandik, P. 287
Manley, D. 11, 34, 37, 53, 357
Marr, D. 282, 290
Marxism 60–1
Index

mathematics 100, 116n, 130, 198–9, 215n, 310, 314–33, 339, 341, 349, 350, 361, 362, 364, 373, 385, 394, 397, 400, 422
applied to 130, 339, 397, 422
and empiricism 315, 323–4, 361
and fictionalism 322
and formalism 317–8, 331n
and indispensability argument, see indispensability argument 10, 319, 326–8, 349–50, 400
and intuitionism 317–8, 331n
and logicism 315–7, 331n
and nominalism 322–3, 324, 327–8
philosophy of 10, 314–33, 362, 373, 385, 394, 400
and platonism 319–20, 321, 322–3, 327–8, 349–50, 400
and structuralism 320–1, 322, 400
and synthetic apriori knowledge 314–5
Maudlin, T. 34, 53, 103, 118, 235, 239, 245, 390, 403–4
Maudn, N. 260, 264
Maxwell, G. 18, 31, 156, 177, 375
Maxwell’s electrodynamics 159, 391, 393, 405, 432, 437
Mayden, R. L. 252, 266
Mayo, D. G. 196, 255, 266, 373, 379, 391
Mayr, E. 345, 357
measurement problem, see quantum mechanics 33, 235–6, 309, 372, 384, 404
Meier, R. 252, 267
Melia, J. 328, 332
Menzies, P. 355, 356
mereology 2, 49, 51n, 151, 249, 322, 357
metaphysical equivalence 42, 48
meta-metaphysics 2–3, 11, 34, 41, 53, 357
Metzger, H. 59, 66, 70
Meyer, E. 59–60, 66, 70
Mink, L. O. 167, 177
Mitchell, S. D. 259, 266, 414
modality 2, 4–5, 36, 37, 40, 44–5, 49, 52n, 93, 106, 111, 326, 327, 332, 344, 358, 371, 435
and analogy, see analogy
autonomous 121, 136, 177
computer 166, 269, 282–3
contextualist account of 172
exemplars, see exemplar
fictional 122
hybrid view 121, 353
mathematical 122, 136, 269, 270, 272, 352, 428–9
material 122, 353
mediating 121, 133, 136, 177, 329–30, 333, 380, 391, 428
partial 94, 129, 135–6, 332, 349, 356
phenomenal 271, 274
representational 8, 121–3, 132–3, 135n, 172, 329, 353
scale 166
simulacrum 172
theoretical 122–3, 428
model theoretic approach 86, 135, 136, 352–3, 364, 368–9, 371, 372, 428, 436 see also semantic approach
model theory 364, 369
Molnar, G. 138, 155
Morgan, M. 133, 136, 177, 329, 333, 373, 380, 391
Morrison, M. 132, 133, 136, 172, 175, 177, 329, 333, 373, 380, 391, 401, 438
Moss, L. 253, 266
Muller, F. A. 242, 245
multiple realizability 142–7, 153, 154, 155, 156, 340, 355
Mumford, S. 138, 155, 351, 357
Mundale, J. 143, 154, 287
Index

Nagel, E. 140–1, 143, 144, 146, 155, 164–5, 173, 176, 177, 258, 266, 364, 367, 376

Newton’s Laws of Motion 66, 70, 38, 233


and homeostatic property cluster 249–51, 261, 287, 352

Needham, J. 60, 64–5, 70

Needham, P. 295, 296, 297, 298, 299, 307, 308, 311, 312

Nersessian, N. 253, 266

Neurath, O. 61–2, 63–4, 65, 67, 68, 70, 71

neuroscience 1, 9–10, 139, 142, 143–4, 149, 150, 152, 154, 155, 165, 174, 175, 176, 264, 268–92, 340, 355, 400–1, 432


new experimentalism 256, 370, 373, 385, 387, 420

Newton, I. 66, 70, 138, 233

Newtonian gravity 142, 176, 179, 189–91, 244, 327, 402, 415, 416, 432

Newton’s Laws of Motion 105–6, 223–4, 361

Newtonian mechanics 58, 76, 105–6, 110, 141, 159, 172, 176, 195, 226, 237, 244, 361, 402

Newton-Smith, W. 369, 377, 392

Nichols, S. 37, 53

Nickles, T. 142, 155, 256, 266, 357, 367, 368, 370, 377

Niiniluoto, I. 369, 371, 378, 379

Nolan, R. 35

no miracles argument 84, 328, 342, 367, 371, 372, 430, 432

non-Euclidean geometry, see geometry, non-Euclidean

normativity 18, 61, 66, 67, 166, 174, 255, 262–3, 265, 279, 348, 427

Norris, C. 58, 70

Norton, J. 69, 116, 118, 182, 196, 228, 245, 254–5, 266, 352, 357, 371, 378, 404


observational terms 61, 62, 80, 191–2, 426, 431, 436

Odenbaugh, J. 352, 357

Ojemann, G. A. 278, 290

Okasha, S. 250, 266, 347, 357, 392

Olton, D. S. 276, 290

O’Malley, M. 352, 356

Oppenheim, P. 140, 141, 151, 152, 155, 257, 265, 279, 290, 364, 375

Ostwald, W. 297, 298, 312

Pais, A. 229, 245

Papineau, D. 36, 53, 260, 266, 358, 392

paradigm 17, 30, 66, 85, 251, 268, 365, 366, 367, 368, 385, 387, 389, 390, 423, 427, 432, 433–4

Park, B. S. 312

Parsons, C. 323, 333

partial interpretation 50

partial isomorphism 129

partial meaning 80

partial structures 135, 136

partial truth 94, 194, 332, 349, 356, 431

partons 39–40

Paul, L. A. 114, 118, 341, 357

Pearl, J. 96, 117, 118, 392

Pereboom, D. 144, 155

pessimistic (meta-)induction 29, 87, 89, 345, 369, 371, 373, 392, 393, 430–1, 432

447
Index

and emergence 418
and inference to the best explanation 192
phenomenalism 62, 94, 148, 183–5
physicalism 52, 62, 112–3, 153, 155, 258, 261, 312, 339, 358, 379
Piccinini, G. 279, 290
Pincock, C. 10, 327, 328, 333, 349
Pinker, S. 275, 290
Place, U. 143, 156
Plaut, D. C. 276, 290
pluralism 10, 35, 107, 154, 252, 256, 260, 264, 266, 342, 350, 352, 356, 357, 398, 399
explanatory 6, 148–50, 153
methodological 26, 259
taxonomic 108
Poater, J. 303, 312
Poincaré, H. 29, 31, 77, 223, 245, 361, 371, 374
Pooley, O. 228, 243, 245, 401
Popperlié, P. 302, 311
novel 40, 46, 84, 89, 351, 373, 417, 430
predictivism 269–72, 274
Price H. 51, 54, 339, 352, 356, 404
Primas, H. 306, 312
frequency interpretation of 236, 362, 388, 399
logical 213–4, 362
prior 20, 180, 181, 182, 185, 194, 202, 203–4, 205, 207, 208, 210–11, 213, 413
posterior 181, 193, 202, 203–4, 206, 207, 208, 211, 214, 216, 413
propensity interpretation of 47, 393
protocol sentences 62, 67, 71, 77
quantum chemistry 301, 303, 306, 311, 312
quantum field theory 244, 245, 346, 347, 355, 379
quantum gravity 38, 50, 245, 345–6, 357, 402, 404
quantum information theory 346–7, 358, 402, 404
quantum mechanics 33, 48, 50, 69, 93, 162, 229–36, 244, 245, 300, 301–5, 307, 308–9, 312, 313, 344, 354, 356,
Index

363, 372, 379, 384, 385, 393, 401, 403, 404, 405, 435
and Bell's Theorem 235–6, 244, 365, 372, 404
and Born rule 230, 232
Copenhagen interpretation of 344
and hidden variables 235, 365
and indeterminacy 117, 233–5
and indiscernibility 242–3
and locality 235
and the measurement problem 33, 235–6, 309, 372, 384, 404
quarks 39, 47, 350
Radick, G. 344, 357
Ramachandran, V. S. 285, 290
Ramberg, P. 301, 312
Ramsey, F. 216, 217
Ramsey, J. L. 167, 177, 305, 312
Ramsey sentence 51, 80–3, 86–7
contextual 305
convergent 89, 357, 368, 378
constructive 136
dappled 118, 135, 373–4, 380, 384, 438
emph{empirical} 75
entity 88, 370, 386, 387, 431
neo-Aristotelian 6, 91–3
patchwork 135, 384
promiscuous 373
semantic 79
semi-6, 88, 92
structural 6, 82, 87, 90, 93, 136, 245, 351–2, 356, 358, 371, 373, 378, 379, 390, 392, 431–2
received view
of theories 121, 368
Redhead, M. 242, 243, 244, 299, 311, 346, 357
and biology 9, 142, 154, 155, 247, 250, 257–62, 264, 266, 398, 399
and chemistry 9, 154, 293, 305, 306–10, 311, 400
eliminative 142, 147
explanatory 258–9
functional 146–8, 153, 156, 173
inter-theoretic 9, 139–42, 156, 176, 258–60, 265, 306–8, 377, 386, 405, 432, 433
local 143, 144–6
micro-140
Nagelian 140–1, 143, 144, 146, 173, 176, 177, 258, 364, 367
and neuro-science 143–4, 154, 155, 257, 262, 284–5, 286, 287, 401
‘new wave’ 141–2
ontological 9, 139, 140, 150, 257, 307–9, 312, 432, 433
translational 140
reference 77, 81, 132, 251, 252–4, 261, 262, 263, 377
causal theory of 248, 251, 262, 295, 321, 393, 423, 428, 431
‘floating’ 253
Reisch, G. 67, 70
relativity 11, 45, 52, 54, 64, 66, 176, 235, 244, 245, 402, 435
general theory of 3, 8, 24, 29, 42, 69, 76, 142, 179, 195, 196, 221–8, 245, 345, 362, 402, 404, 416
special theory of 223–6, 245, 362, 401–2, 403, 404, 432
relativism 4, 17, 66, 394, 395, 433–4
reliabilism 23–4

449
Index


artistic 8, 354
denotational account of 124
diagrammatic 166, 313
epistemic 123–30, 134
faithful 8, 121, 124, 127–130, 131, 134
geometric 226
and idealization 130–2, 135, 349
inferential account of 125–6, 134
interpretational account of 126–7, 134
linguistic 166
mathematical 110, 314, 329–30
mental 262, 283
and realism 132–3
similarity account of 124, 127–8, 130
structural account of 124, 128–30, 134
and surrogative reasoning 8, 123, 126, 134
of theories 172, 353–4, 358
Resnik, M. D. 320, 328, 333
Richardson, A. 56, 69, 70, 394, 419
Richardson, R. C. 141, 148, 150, 154, 156, 264, 269, 275, 287
Rieke, F. 279, 290
Rickles, D. 245, 346, 357, 358, 404
Rieppel, O. 249, 266
rigid designation 146–8, 248
Ritchie, J. 36, 54
Roberts, J. 118, 351, 357, 394, 426
Robinson, D. A. 282, 290
Rocke, A. J. 301, 313
Rosenberg, A. 259, 266, 345, 357, 394–5, 398
Ross, D. 3, 11, 34, 52, 53, 93, 94, 352, 357, 390
Royall, R. M. 182, 184–5, 196, 216, 217
Rueger, A. 133, 136, 148, 156
Rumelhart, D. E. 280, 290
Ruse, M. 247, 265, 367, 377, 397, 398
Russell, B. 96, 118, 315, 317, 352, 356, 362, 375, 404
Russell’s paradox 317
Saatsi, J. 134, 136, 245, 357
Sabbarton-Leary, N. 310, 311, 312, 338, 355, 430
Sadato, N. 278, 290
Salmon, M. H. 255, 266
Salzmann, C. D. 279, 290
Sankey, H. 91, 95, 196, 251, 266
de Santillana, G. 60, 68
Sarkar, S. 142, 150, 156, 357, 398
Saunders, S. 242, 245, 372, 379
Savage, L. J. 196, 216, 217
Scrier, E. 307, 312, 313, 400
Schaffner, K. 139, 141, 142, 148, 156, 256, 258, 266, 268, 271, 287, 291, 378
Schlick, M. 16, 31, 57, 60, 61, 62, 70, 75, 95, 383
Schöner, G. 275, 291
Schummer, J. 297, 310, 313
Schurz, G. 160, 177
Scriven, M. 161, 168, 177, 375
Searle, J. 279, 291
Seevinck, M. P. 242, 245
Sejnowski, T. J. 151, 154, 279, 280, 282, 288, 291
Sellars, W. 79, 81, 95, 155
semantic approach, see model theoretic approach
Shagrir, O. 280, 282, 283, 291
Shapere, D. 366, 376, 396
Shapiro, L. 144, 156
Shapiro, S. 320, 328, 330, 331, 333, 373, 379, 400
Shomar, T. 135, 355, 379
Shostak, S. 254, 266
Sider, T. 40, 41, 48, 53, 54, 340, 357
Silberstein, M. 273, 288, 380, 390
Index

Simon, H. A. 285, 291
Simples 38–9, 41, 51, 53, 54
Sintonen, M. 170, 177
Sipser, M. 282, 287, 291
Sklar, L. 140, 141, 156, 374, 379, 405
Smart, J. 83–4, 95, 365, 376
Smith, P. 331, 333
Sober, E. 180, 193, 195, 196, 247, 248, 256, 266, 370, 378, 398–9
social sciences 61, 113–4, 136, 139, 157, 162, 163, 164, 167, 173, 174, 175, 178, 249, 250, 263, 291, 333, 391, 394, 397
sociology of science 55, 65, 70, 256, 262, 368
species (biological) 143, 145, 146, 147, 148, 211, 246–52, 254, 263–7, 294, 370–1, 397, 398, 399
species (chemical) 296–7, 299, 303, 306, 313
Spencer, J. 38, 54
Spirtes, P. 96, 119
Squire, L. R. 285, 291
Stadler, F. 67, 71
Stalnaker, R. 102, 119
Stanford, K. 88, 95, 345, 357, 395
statistical mechanics 8, 118, 236–9, 245, 402, 404, 405
and thermodynamics 8, 141, 156, 176, 236, 379, 432
Steel, D. 285, 291
Stegmüller, W. 364, 368, 371, 377
Stephan, A. 138, 150, 156, 273, 291
Sterelny, K. 247, 265, 397, 399
Stewart, C. M. 278, 289
Stich, S. 37, 54
Stotz, K. 253, 265, 266
Strawson, P. F. 35, 51, 364, 365, 376
string theory 38–9, 48, 51–2, 346, 355
Strauss, D. 58, 71
structuralism 81–3, 86, 93, 135, 320–3, 356, 364, 368, 371, 378, 396, 400
micro- 294–7, 300, 306
structural realism, see realism, structural
Stueber, K. R. 167, 178
Suárez, M. 134, 135, 137, 311, 329, 333, 354, 355, 357, 379
substance 9, 16, 21, 101, 293–313, 352, 361
microstructuralist account of 293, 294–7, 300
and phase rule 299
and separability 297–8
Sullivan, J. A. 287, 291
Sumner, J. 343, 358
supervenience 118, 151, 155, 156, 308, 435–6
Suppe, F. 2, 5, 8, 11, 132, 133, 137, 368, 371, 377, 378, 396, 427
Suppes, P. 133, 134, 137, 364, 368, 370, 375, 376, 396
surrogative reasoning 8, 123–7, 134, 135, 137
Sutcliffe, B. T. 304, 313
Swoyer, C. 123, 137
syntactic approach 121, 133, 213, 258, 368–9, 436
to computation 281–2
Taper, M. 195, 196
Teller, P. 22, 32, 127, 137, 243, 244, 245, 372, 374, 379
theoretical terms 77, 80–1, 192, 295, 364, 366, 393, 395, 423, 425, 426, 436
theoretician’s dilemma 80, 94
Thomasson, A. 354, 358
Thurlow, K. J. 294, 313
Timmermans, J. 295, 297, 298, 313
Timpson, C. 347, 358
Tobin, E. 248, 263
Toulmin, S. 55, 366, 368, 371, 388
Index

approximate 89, 132, 171, 187, 195, 349, 373, 430, 431

underdetermination 61, 93, 356, 367, 369, 377, 386, 389, 390, 393, 394, 432, 436–7
understanding 4–5, 6, 48, 97, 101, 121, 149, 163, 168, 170–1, 172, 175, 176, 177, 227, 286, 297, 311, 329, 343, 350, 352, 377, 388
Urbach, P. 20, 31, 182, 196, 216, 217, 370, 378, 386
Uzan, J-P. 46, 53

Vaidya, A. 344, 358
van Brakel, J. 293, 295–7, 307, 310, 313, 400
Van Gelder, T. 286, 291

verificationism 42, 61, 62, 75–8, 139, 362, 364, 365, 416–7, 426–7
verisimilitude 29, 349, 367, 369, 371, 377
Vickers, P. 346, 358
Vienna circle 4, 57–8, 60–3, 65, 67, 68, 69, 70, 71, 394
Von Neumann, J. 280, 291

Wald, R. M. 224, 228, 245
Wallace, D. 346, 358
Walter, S. 9–10, 144, 148, 156, 263, 355
Ward, B. 111, 119
Waters, C. K. 253, 267
Weinberg, S. 42, 173, 178, 345, 358, 386
Weisberg, M. 135, 137, 302, 313, 331, 333
Wheeler, J. A. 68
Wheeler, Q. D. 252, 267
Whewell, W. 15–6, 17, 32, 57, 71, 183–4, 194, 195, 196, 361, 374, 392
Wiesel, T. N. 280, 289
Will, C. M. 195, 196
Williamson, T. 29, 32, 183, 196
Wilson, J. 340, 358
Wilson, M. 254, 267, 339, 358, 396–7
Wilson, R. A. 249, 267, 287, 291, 340, 358
Wimsatt, W. C. 139, 142, 148, 150, 151–2, 153, 156, 256, 258, 261, 267, 268, 285, 292
Winch, P. 175, 178
Winters, E. 239, 245
Woody, A. I. 310, 311, 313
Woolley, R. G. 303–5, 307, 313
Worrall, J. 90, 95, 351, 358, 373, 378, 392
Wright, C. 321, 331n, 332

Yablo, S. 41, 54, 322, 333
Yu, F. H. 272, 292

Zemansky, M. 298, 313
Zilsel, E. 60, 68, 71
Zimmerman, H. 67, 71