On Inter-Theoretic Relations and Scientific Realism

Sebastian De Haro Ollé

Trinity College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

March 2020

Abstract

This thesis addresses three contemporary debates in the philosophy of science: namely, scientific realism, emergence, and theoretical equivalence. The thesis brings logico-semantic tools of the analytic tradition—about syntactic and semantic construals of theories, and about extensions and intensions—to bear on these debates. The thesis has two parts: Part I (Chapters 1-3) lays out the overall framework about scientific theories, scientific realism, and emergence. Part II (Chapters 4-6) develops more detailed themes.

Part I first gives a conception of a scientific theory (Chapter 1), using logico-semantic tools that will be used in the rest of the thesis.

Chapter 2 then brings these tools to bear on the debate about scientific realism, by construing the continuity of theories as a matter of extensions. The resulting position is a modest scientific realism, according to which one is justified in believing what confirmed theories say about extensions but not, in general, about intensions. I dub it 'extensional scientific realism'.

Chapter 3 proposes an account of the distinction between ontological and epistemic emergence, based on an explication of the notion of 'novel reference'. The ontological emergence of one theory from another is defined as the failure of an appropriate linkage map between the two theories to "mesh" with the two theories' interpretations.

In Part II, Chapter 4 first develops a notion of theoretical equivalence, and introduces duality in physics, as an appropriate isomorphism between theories. The Chapter discusses the relation between duality and theoretical equivalence in philosophy of science.

Chapter 5 discusses the heuristic roles of dualities in theory construction. It develops a distinction between the theoretical and heuristic functions of scientific theories, and illustrates the heuristic function of duality in theory construction.

Chapter 6 discusses how theories without a spacetime can lead to scientific understanding. To this end, the Chapter describes three theoretical tools that are often used in theory construction and which lead to understanding, both in cases with and cases without straightforward spacetime visualisation.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared below, and specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma, or other qualification at the University of Cambridge or any other University or similar institution. The length of this dissertation is approximately 80,000 words (excluding the Publications and References sections), and as such does not exceed the word limit of 80,000 words—as prescribed by the Degree Committee of the Department of History and Philosophy of Science, University of Cambridge.

- Chapter 1 is based on five published papers: De Haro (2019a, 2019b, 2020), De Haro and Butterfield (2018), De Haro and De Regt (2018); and on a paper under review, De Haro (2020c).

- Chapter 2 is based on a paper under review, De Haro (2020c).

- Chapter 3 is based on a published paper, De Haro (2019).

- Chapter 4 is based on three published papers, De Haro and Butterfield (2018) and De Haro (2019b, 2020).

- Chapter 5 is based on a published paper, De Haro (2019a).

- Chapter 6 is based on a published paper, De Haro and De Regt (2018).

Further work that I have published and is related to, but is not included in, this dissertation, is briefly discussed in the final Section of each Chapter.

Acknowledgements

I am greatly indebted to Jeremy Butterfield: for his guidance, close reading of texts, and encouragement throughout. These nouns hardly express one's privilege for having you as an advisor, collaborator, and friend. And it is rare to find another philosopher with whom one always ends up agreeing in matters philosophical, both general and of detail: or *almost* always! I also thank a number of other friends, with whom I had the pleasure and privilege of collaborating in various projects: Elena Castellani, Henk de Regt, Dennis Dieks, Nicholas Teh, Jeroen van Dongen, and Manus Visser.

I thank my advisor Hasok Chang, for guidance and philosophical discussion. I have learned so much from your philosophical temperament: especially about scientific realism, where you have been a major inspiration, even though our proposed views differ.

My philosophical thinking has been thoroughly influenced by discussions with Richard Dawid and Nicholas Huggett. I am grateful to the following philosophers for discussions about dualities and theoretical equivalence, often at workshops or symposiums that one of us organised: Doreen Fraser, Eleanor Knox, Keizo Matsubara, James Read, Dean Rickles, and James Weatherall. For discussions about emergence, again often at workshops that one of us organised: Jay Armas, Karen Crowther, Samuel Fletcher, Alexandre Guay, Stephan Hartmann, Klaas Landsman, Patricia Palacios, and Christian Wüthrich.

With scholars like you, I can foresee a future where philosophy will again be done, in addition to offices and lecture halls, also on walkways and public squares!

A very special thanks to Peter Galison for his invitation to join his group at the Black Hole Initiative at Harvard University during the winter semester of 2018-2019. This was a truly special time, not in the least thanks to two of my collaborators' joining the group during part of my visit. Thanks also to other members of that group: especially Abraham Loeb, Ramesh Narayan, Andrew Strominger; and Cumrun Vafa of the Department of Physics at Harvard University. Thanks to Shing-Tung Yau, from Harvard's Center of Mathematical Sciences and Applications, for his invitation to speak at two conferences in Hong-Kong and Hainan, China, in January 2019. To Alisa Bokulich for her invitation to speak at the Boston Colloquium in Philosophy of Science in October 2018. To Hans Halvorson for his invitation to speak at the workshop *Equivalent Theories in Physics and Metaphysics,* Princeton University, in March 2015.

Thanks for the philosophical discussion, both general and specialist (often holding a piece of chalk or a beer but always in earnest), to: Feraz Azhar, Guido Bacciagaluppi, Adam Caulton, Erik Curiel, Radin Dardashti, Silvia De Bianchi, Henrique Gomes, Carl Hoefer, Vincent Lam, Baptiste Le Bihan, Niels Linnemann, Olimpia Lombardi, Daniel Mayerson, F. A. Muller, John Norton, Daniele Oriti, Brian Pitts, Hans Radder, Miklós Rédei, Bryan Roberts, Olaf Tans, Rudi te Velde, and Bobby Vos.

This work was supported by: the Tarner Studentship in Philosophy of Science and History of Ideas of Trinity College, Cambridge; the Black Hole Initiative, Harvard University; the Dutch Institute for Emergent Phenomena; 'Space and Time After Quantum Gravity', University of Illinois at Chicago and University of Geneva; the Rouse Ball/Eddington Funds, Trinity College, Cambridge; a Faculty International Experience Grant of the Amsterdam University College; the Vossius Center for the History of Humanities and Sciences, University of Amsterdam; the Faculty Development Fund, Amsterdam University College.

Contents

Introduction

Since the logical positivists, the question ‘What is a Scientific Theory?’ has defined the philosophy of science, and has been seen as giving unity to the field’s otherwise rather disparate investigations. The *syntactic view* of scientific theories of the logical positivists holds that a theory is a set of sentences, appropriately structured, usually by deduction from a set of axioms. A subset of these sentences constitutes the *observation sentences*, which are suitable for empirical verification or confirmation. The syntactic view’s rival is the *semantic view* of theories, according to which a scientific theory is a set of models. Models contain substructures, the *empirical substructures,* that are suitable for empirical comparison.

While these two influential accounts focus on scientific *theories,* they are of course not limited to the theoretical aspects of science. For the accounts explicitly consider the empirical aspects—including observation, experimentation, etc.—to be important parts of a theory, essential for it to achieve its aims. And thus the description, prediction, and explanation of observable phenomena are seen, by most accounts, as important aims of science. Thus, on the traditional view in philosophy of science, ‘theory’ is a broad term that refers not just to the theoretical aspects of science, but also to experimental work and instrumentation.

These traditional accounts have been augmented by accounts that stress complementary aspects, especially in the face of the criticism that a focus on ‘theory’ tends to downplay scientific practice, and that to understand the nature of science one needs to study not just its end-products, but also the activities carried out by scientists.

In this thesis, I will re-examine the notion of a scientific theory, and its functions with respect to the aims of science, in the light of two developments, of which one is philosophical, and the other is partly philosophical and partly scientific:

(i) The development and application of logico-semantic tools in philosophy of science, and in particular the attention paid to the question of what elements contribute to fixing reference.

(ii) The recent focus, in both philosophy and in science, on inter-theoretic relations. Indeed, philosophers and scientists agree that the comparison between theories can contribute both to a better understanding of the phenomena, and to practical advantages such as increased computational power.

I will, in particular, consider three inter-theoretic relations that have recently received considerable attention:

(1) The general—and correspondingly somewhat vague, but nevertheless valuable—notion of 'correspondence' between theories. This has been used, in particular, in discussions of scientific realism and of heuristics.

(2) The notion of 'emergence', in both philosophy and in science, and its relation to the more traditional notion of 'reduction'.

(3) The notion of theoretical equivalence, and one of its well-studied instantiations in physics: namely, dualities between otherwise very different-looking theories. Developing such a notion then bears on the individuation of theories.

The thesis introduces a simple—indeed, almost elementary—conception of 'theory' that is apt for inter-theoretic comparison, and which it then brings to bear on the three topics above: namely, scientific realism, ontological emergence, and the individuation of theories.

I developed the account of theory that I will use, first on my own and then with Butterfield, in order to understand dualities in connection with the question of theoretical equivalence. The account then turned out to have a use as well for the question of emergence. Finally, the semantic distinction between intensions and extensions turns out to give a novel perspective on the problem of scientific realism—in particular, it allows for the construction of a new reply to the pessimistic meta-induction argument, and the development of a scientific realist position.

Thus the way in which the chapters are ordered is reverse relative to how they were written. This reversed order is natural, for the thesis will first present the more general question of scientific realism, and then move on to the more specific questions of emergence and theoretical equivalence—each of which has its own connections with the problem of scientific realism, although I do not have the space to explore those connections here.

The account of scientific theories explicitly takes into account various important aspects of scientific practice: such as the use of heuristics, the question of scientific understanding and visualisation, and of the material realisation of theories. Indeed the focus on inter-theoretic *relations,* rather than on single theories, is itself in part motivated by a parallel shift in the focus of scientific practices.[1]

The basic methodology that I will follow will be to apply classical semantics to scientific theories, to address the specific problems (1)-(3), thereby articulating my account of what a scientific theory is, and of how it achieves its aims.

And since my aim is to cast light on these three problems, which I believe are philosophically and scientifically urgent, rather than to give a definitive account of a 'scientific theory', the thesis does not embark in a systematic philosophical exploration of the notion of 'theory'. Rather, my approach will be opportunistic and, in some places, informal:

[1] Some Sections do consider relations within a single theory, e.g. between different models of it. These relations can often be seen as special cases of inter-theoretic relations between different theories.

namely, I will only include those aspects that help towards my three problems of scientific realism, ontological emergence, and theoretical equivalence.

Thus in Chapter 1 I will collect the elements—from semantics and from semi-formal mathematical philosophy—that I will need to define a theory, and to define inter-theoretic relations and theory-to-world relations. Chapter 2 then introduces an answer to the pessimistic meta-induction argument, based on an extensional version of scientific realism that the Chapter develops. Chapter 3 then uses the notion of a theory, from Chapter 1, to define ontological and epistemic emergence for the formal sciences, which it then applies to examples from physics. Chapter 4 gives a notion of theoretical equivalence, which it then contrasts with the phenomenon of duality in physics. The last two Chapters return to the aims of science: Chapter 5 discusses the role of heuristics in theory construction, in connection with dualities. Chapter 6 addresses the notion of scientific understanding and how it applies to theories with no spacetime.

The scope of the thesis and the exposition is detailed in some important places, but necessarily brief in many others. Thus several topics that would deserve a study of their own are, due to lack of space, not included in this work. In particular, each of the Chapters 3-6 ends with 'Discussion/Summary and Further Work', thus summarising other work that I have published on the topic. In particular, the thesis will not address in detail the following points:

(i) the formulation of a fully-fledged scientific realist position, especially the analysis a large number of historical examples and a general answer to the threat of empirical under-determination;

(ii) the comparison of my extensional scientific realism with other important forms of scientific realism;

(iii) fleshing out more details about, and various kinds of, epistemic (as opposed to ontological) emergence; and

(iv) analysing in detail various dualities in physics.

Part I. A Framework

Chapter 1

What is a Scientific Theory?

This Chapter will first address the aims of scientific theories (Section 1.1). Then, in Section 1.2, I will give a characterisation of scientific theories that will be useful for the various problems addressed in the rest of the thesis. In Section 1.3, I introduce some logico-semantic tools used in later Chapters.

1.1 Aims of Scientific Theories

In this thesis, we will be concerned with inter-theoretic relations and how they are used. Thus Chapter 2 addresses the relevance of relations of correspondence for the *truth* of scientific theories. In Chapters 3 and 4, the main motivation is the *predictive and explanatory* properties of emergence and equivalence. In Chapter 5, we will rather be interested in *theory construction* and in the *heuristic* function of equivalence relations.

The uses of inter-theoretic relations, to be discussed in the various Chapters, thus correspond to different *aims* of scientific theories: truth, prediction, explanation, heuristics, etc. In this Section, I discuss some of these aims, and how they can sometimes lead to tensions between the different *functions* of the elements comprising a theory. I will argue that some of these tensions are substantive; although they do not necessarily lead to contradictions, and they can be resolved.

A theory can be used to *describe* the world in detail and accurately, on a suitable interpretation. Other uses of theories are *instrumental*: for example, using the theory as a calculational tool to get "quick and dirty" results about a situation of interest, without paying attention to other contextual details that are irrelevant for, say, the aim of quantitative *prediction* in a specific situation—though prediction need, of course, not always be instrumental.

These two uses—the descriptive and the instrumental—are of course tuned to corresponding aims of theories: and so there is no real incompatibility here. Debate can then ensue about how to *prioritize* those goals, about how the goals are related, or about the conditions under which a given goal is worth pursuing. Laudan (1984: pp. 63-64) has argued that scientific theories can have several different goals. 'There is no single "right" goal for inquiry because it is evidently legitimate to engage in inquiry for a wide variety of reasons and with a wide variety of purposes'.[1]

However, Laudan makes here no distinction between the goals of *individual scientists* and the intrinsic aims of *science*—and I agree with van Fraassen (1994: pp. 181, 191-192) that one should not confuse those two sets of aims, since: (a) the aims of science are not by definition the same as the aims of all, or most, scientists with respect to theory construction; (b) the aims of all, or most, scientists are nevertheless surely *relevant* to the philosopher's understanding of these aims. However, I disagree with van Fraassen that science has a single aim.

The reason why I am a pluralist about the intrinsic aims of science is that science is a human activity fully embedded in society. Thus its intrinsic aims are not only formed and measured by the aims of all, or most, scientists: but also by society's needs. Science's accountability and responsibility accrues society a relevant role for the determination of the aims of science, so that there cannot be just one aim: even though the aims *can* be prioritized, and search for truth about the natural world certainly *is* a distinguished aim.

This is of course a broad topic beyond the scope of this thesis, and the previous brief remarks are aimed only at defending my pluralism about aims.

In sum: a theory, and especially its *interpretation*, can—in addition to description—also be aimed at explanation or at understanding (Toulmin (1961: §2), De Haro and De Regt (2018: §1.1), De Regt (2017)). For example, Ruetsche (2011: pp. 3-4) has contrasted an 'ideal of pristine interpretation', which sees the business of interpretation as a 'lofty' affair that is only concerned with the general question of which worlds are possible according to the theory, and not with the application of the theory to actual systems. She argues for 'a less principled and more pragmatic approach to interpreting physical theories, one which allows 'geographical' considerations to influence theoretical content, and also allows the same theory to receive different interpretations in different contexts.' (p. 4).

A similar difference is sometimes seen between differing uses of the word 'model'. While the philosophy of physics literature tends to endorse the semantic conception of models, i.e. as the set of worlds that are possible according to the theory, in the general philosophy of science the notion of models involves both context and approximation—here, models mediate between the theory and concrete phenomena, which obtain under definite circumstances.

One must of course recognise that behind these disagreements in the literature, about what interpretations and models "really are", there can be—and there often are—larger philosophical differences: between theoretical and practical approaches to science, between realist and anti-realist positions, or between trust in the notion of laws of nature vs. belief in a 'dappled world'—to mention just some.

But one must also admit that, these larger differences aside, there is a clear way to go about resolving the disagreements about the philosophical notions of interpretation and model, i.e. about the notions which comprise a theory: namely, by recognising that they are mostly intended for diverse, but equally legitimate, aims for which the notions involved are used. Thus for example, the question of whether a model is a possible world in which the theory is true, or is a contextual and specific application of a theory to

[1] See also Toulmin (1961: §2).

describe a phenomenon, will receive different answers depending, to a large extent, on the kind of question one is asking. Indeed, it will depend on the specific purpose one wishes the word ‘model’ to fulfill—that of, say, expounding the descriptive capacities of a theory vs. that of expounding its applicability to specific cases. Again, the two uses are legitimate, though they will no doubt lead to different philosophical accounts.

1.2 Scientific Theories

Having briefly discussed the aims of scientific theories, this Section introduces the characterisation of scientific theories which I favour: because it is, I believe, close to how scientists tend to think about scientific theories.[2] The main feature of the characterisation is that it makes, without ever completely separating them, a conceptual distinction between the formal and interpretative aspects of scientific theories that in its turn allows us to use ideas from the philosophy of language.

1.2.1 On inter-theoretic relations, theories, and models

In this thesis we are interested in how the three inter-theoretic relations that I mentioned in the Introduction bear on what counts as a scientific theory, and thus how they bear on theory individuation. More specifically, Chapter 3 will study relations of *emergence* between theories, thus emphasising the *differences* between theories that may otherwise be closely related. And Chapter 4 will address the topic of the *equivalence* of theories. Roughly speaking, we will see that dualities in physics present novel cases of theoretical equivalence. Some of these dualities relate theories that look very different and that one would not at first sight have thought could be equivalent. Counting dualities as cases of theoretical equivalence—so that we can, in some cases, take very different-looking theories to be “mere rewritings of a single theory”—thus invites both the introduction of some new notions, and the modification of some of the standard jargon in philosophy of physics about ‘theories’ and ‘models’.

Specifically, the situation of very different-looking theories turning out to be, in an appropriate sense, equivalent, calls for rethinking what we mean by a ‘theory’. Indeed, if very different theories can be “one and the same theory”, then it would seem that one should generalise one’s notion of ‘theory’, so that two different ‘theories’ (old sense) can be the same ‘theory’ (new sense).

In view of this broadening of the terminology, I will argue that what we used to call a ‘theory’ (old sense) should now be seen as a ‘model’, i.e. a particular instantiation, realisation or representation of a theory (new sense). Two such models can then turn out to be equivalent, so that what we thought were distinct theories (old sense) are, on closer analysis, representations—models—of a single theory.

Thus I will propose that what we mean by a ‘model’ is not fixed, but is relative to ‘theory’:

[2] The characterisation of scientific theories, that follows in the next few pages, should not only apply to the physical sciences, but also to other fields whose theories are sufficiently formal.

A model is a realisation, instantiation, or representation (in the mathematical, not philosophical!, sense) of a theory.

But a theory can be more or less concrete, depending on the context. The relevant notion of 'theory' here will be that of a 'bare theory': see Section 1.2.2.

Thus my usage differs from the standard one in philosophy of physics, where the meaning of 'model' is more fixed. (It also differs from the "models as mediators view" discussed above, to which I will get back in Section 1.2.3 and in Chapter 6, when I discuss understanding). In philosophy of physics, a 'model' is standardly taken to be a solution of a set of equations of a theory (or perhaps a class of such solutions, e.g. when the model contains unspecified parameters). Hence a model thus understood represents a possible world, or a corresponding class of worlds.

The standard usage in philosophy of physics has the merit of its simplicity. On this usage, a model is a (set of) possible world(s), which is also an interpretation of the theory.

My proposal deviates from this specific and fixed usage. Indeed I propose to disentangle the meanings of 'model', 'possible world', and 'interpretation', for each has its own specific use. My main motivation for this comes from dualities, which give, as I will explain in Chapter 4, cases of theoretical equivalence between seemingly distinct theories. I will propose that we generalise the notion of a theory and, at the same time, keep the relation between theory and model fixed, in the general sense of one theory having many models that instantiate it. Namely, a model instantiates, realises or represents (I will also say: 'models') a theory. Thus, if the notion of 'theory' is generalised, then so is the notion of 'model'. What used to be a theory should, on this usage, be called a model. Thus I will advocate that we effectively push our usage of 'theory' and 'model' "one level up".

There are—apart from criticisms of the "pristine" view of interpretation, mentioned in Section 1.1, and the recognition of the pluralism of the uses of 'model' in physics—several other additional motivations for this usage. First, the notion of 'model as a solution of a set of equations' is rooted in classical physics, where the laws of the theory are almost always summarised by a set of differential equations. But quantum field theories are of course not always so formulated. For example, quantum field theories are often presented as a prescription to evaluate a set of path integrals that calculate transitions between states, and only in a second step is the equivalence with a Schrödinger-type of formulation shown.

Second, the notion of a 'theory' is itself vague both in physics and in philosophy of physics. In physics, a theory is paradigmatically fixed when a Lagrangian action is given—both in the classical and in the quantum cases. But there are exceptions: some systems that possess a set of equations of motion do not have a known action principle (and such an action principle is believed to not exist). Also, some systems can be described by different actions, which are nevertheless all thought to define the same theory for the same (set of) systems. Some dualities are paradigmatic examples of this, where the quantum theory has several classical formulations, each with its own action principle.

Thus I cannot see a basis in physics (even less so in any other natural science!) why we should, artificially, fix once and for all what we mean by a 'theory' (and, correspondingly, by a 'model'). Indeed I do not believe that a single agreed definition of this notion in science exists. My approach here will start with the question of inter-theoretic relations that distinguish and individuate theories. In the next Section, I will introduce conceptions of 'theory' and 'model' that: (i) are general and flexible enough that they can be used to study the question of theory individuation: we 'build our boat while at sea', to use Neurath's metaphor;[3] (ii) are concrete enough that they will allow us to address various philosophical questions.

1.2.2 Theories and models

In this Section, I discuss the notions of theory, model, and interpretation in more detail. The main step is to introduce *bare* notions of theory and model, which together with an interpretation form what we call a 'theory' or a 'model'.

Thus we start with the notion of a *bare theory*: a physically uninterpreted, but mathematically formulated, structure with a set of rules for forming sentences, i.e. an abstract calculus (we will refer to these rules as the 'language' of the theory: see below).[4] A bare theory could consist of a set of axioms or a set of equations. But, to be specific, I consider a bare theory as a triple, $T := \langle \mathcal{S}, \mathcal{Q}, \mathcal{D} \rangle$. It consists of:

a state-space endowed with appropriate structure, $\mathcal{S}$;

a set of quantities endowed with appropriate structure, $\mathcal{Q}$, and

a dynamics, $\mathcal{D}$, consistent with that structure.

'Appropriate structure' here refers: first, to symmetries which may act on the states and-or the quantities, e.g. as automorphisms of the state-space, $a : \mathcal{S} \to \mathcal{S}$. And second, 'structure' also refers to the set of rules for forming sentences, e.g. for assigning values to the quantities.[5]

[3] See Quine (1960: p. 3).

[4] One may object that theories in physics always come to us interpreted. But one should not take the conceptual analysis that I am doing here, namely dissecting a theory into its bare, uninterpreted, part and its interpretation, to be the temporal or methodological process of constructing new theories. Two reasons for this are as follows: (1) Although theories of physics usually come to us interpreted, it is sensible to conceptually distinguish between the formalism of a theory and its interpretation. (2) Bare theories often have *more than one interpretation,* and the bare theory captures the formal structure that underlies them: for example, the heat equation can be interpreted as describing the spread of heat in time in a given region, or as a diffusion equation for Brownian particles in a medium. My account of bare theories bears some similarities to the framework of van Fraassen (1970: pp. 328-329, 2014: pp. 281-282). What I will refer to as the 'sentences' of the bare theory is similar to van Fraassen's 'elementary statements'.

[5] 'Structure' here also refers to the rules for evaluating quantities. For the examples of quantum theories, which will illustrate this conception of theory, these are: (ia) the set of states will be a separable Hilbert space; (ib) the quantities will be elements (normally the self-adjoint, renormalisable elements) of an algebra; (ii) the rules for evaluating quantities are maps to the appropriate field: for most quantum theories, the inner product on the Hilbert space, and the usual rules for evaluating matrix elements; (iii) dynamical evolution will usually be a (unitary) map, satisfying appropriate commuting diagrams with the other maps in the theory; (iv) the group of symmetries will comprise the automorphisms of the algebra: and possibly additional symmetries, on the states and on the quantities. For classical theories, these comments get modified in familiar ways: e.g. (ia) would say that the set of states is a manifold, with structure appropriate to e.g. Lagrangian or Hamiltonian mechanics. These cases of 'structure' are of course not exhaustive. For example, state-spaces often come equipped with other structures: a topology, a symplectic form, etc. But we can take this in our stride—we will not need, nor is it I think meaningful, to try and specify all the relevant structure once and for all: it can be done case by case.

The word ‘uninterpreted’ here is perhaps best unpacked as ‘stripped of an explicit or concrete physical interpretation’. For we are doing physics, and not pure mathematics: thus a bare theory, even though it does not have an explicit or concrete interpretation, is a piece of mathematical physics that is still connected to physics—as suggested by the names ‘state’, ‘quantity’, and ‘dynamics’. Thus the point about the bare theory is simply that concrete interpretations are temporarily bracketed by our conceptual analysis: which, for example, allows us to conceptualise how a single theory can receive several different interpretations (see footnote 4).

For a *quantum* theory, which will be our main (though not our sole!) focus: we will take $\mathcal{H}$ to be a Hilbert space; $\mathcal{Q}$ will be a specific subset of operators on the Hilbert space; and $\mathcal{D}$ will be taken to be a choice of a unique (perhaps up to addition by a constant) Hamiltonian operator from the set $\mathcal{Q}$ of physical quantities. In a quantum theory, the appropriate structures are matrix elements of operators evaluated on states; and the symmetries are represented by unitary operators. But as mentioned: the present notion of a bare theory applies equally well to classical theories.[6]

A bare theory may contain many more quantities, but it is only after we have singled out the ones that have a physical significance that we have a *physical*, rather than a *mathematical*, theory or model. The quantities $\mathcal{Q}$, the states $\mathcal{H}$, and the dynamics $\mathcal{D}$ have a physical significance at a possible world W, and within it a domain of application D_W (which we can think of as a subset $D_W \subseteq W$), though it has not yet been specified what this significance may be, nor what the possible world “looks like”. To determine the physical significance of the triple, a physical interpretation needs to be provided: which I do in the next Section.

Recall that, as I announced in Section 1.2.1, a model is in this thesis *not* a particular solution of a theory (nor an approximate solution or set of solutions). Rather, a *model* M of a bare theory, T, is a realization, or mathematical instantiation: i.e. it is a mathematical entity having (a) the same structure as the theory, and usually (b) some specific structure of its own. More specifically, I will use a notion of model as a representation (in the mathematical sense, not the philosophical one!) of the theory, i.e. a homomorphism from the theory to some other known structure. Think, for example, how an abstract group can be represented by a set of $n \times n$ matrices, where the rank, n, is specific structure that goes beyond the definition of (the homomorphic copy of) the group.

Thus a model must not be a merely mathematical representation, since we need to make the following physical distinction within the homomorphism. A model of a physical theory naturally suggests the notions of:

(a) A *model root*: the realization of the theory, usually its homomorphic copy (though, in some cases, an isomorphic copy).

(b) The *specific structure*: that structure which goes into building the model root, which is not part of the theory’s defined structure, and which gives the model its specificity. Specific structure may well be physically significant, depending on the context of application of the theory: and part of it is normally used for calculations within the

[6] The ‘dynamics’ here is not necessarily a time evolution, but can also be any dynamical condition, like the specification of a constraint on the Hamiltonian’s value (e.g. in cases where the Hamiltonian vanishes).

model—but calculations can of course be done in different ways, using different specific structure. For examples, see: Chapter 4, De Haro (2020: §2.1), and De Haro and Butterfield (2017: §5.2).

It is helpful to have a schematic notation for models that exhibits how a model augments the homomorph of the structure of a bare theory, T, with its own specific structure:

$$M = \langle m; \bar{M} \rangle \ . \tag{1.1}$$

Here, m is the model root, and $\bar{M}$ is the specific structure which goes into building m. In cases where T is a triple, m must itself be a triple with properties that are homomorphic to those of T. When the model root, m, is itself a triple, then we call it the *model triple*. Notice that the distinction between the model root and the specific structure, even if formalised into the definition of the model as in Eq. (1.1), is conceptual rather than strictly mathematical. This means, firstly, that it can only be made on a case-by-case basis; and second, that we should not think of m and $\bar{M}$ as being given independently and together defining M, i.e. Eq. (1.1) is not an ordered pair of two "pre-given" items. Rather, $\bar{M}$ is the specific structure from which the model root m is built.

Like bare theories, models (and model triples) are, at this stage, *uninterpreted*; and an interpretation can again be added as a set of partial maps (see Section 1.2.4).

Finally, it will also be convenient, in Chapter 4, to have a notation for a model considered in itself, not by comparison with the bare theory of which it is a model. A model is of course *itself* also a triple of a set of states, quantities and a dynamics: i.e. its own states etc., not that of the bare theory. And we will again use the overbar to indicate what is specific to the model. So we write:

$$M = \langle \bar{\mathcal{S}}, \bar{\mathcal{Q}}, \bar{\mathcal{D}} \rangle \ . \tag{1.2}$$

1.2.3 What is an interpretation?

So far I have only spoken about bare, i.e. uninterpreted, theories and models. In this Section, I will sketch some of the recent literature on interpretation, and place my preferred conception of interpretation against this background.

If we want to know what a scientific theory says about the world, or if we want to use a theory for predicting, explaining or understanding empirical phenomena, we need to *interpret* that theory. Abstract theories can only be put to concrete use if they are interpreted. As Bas van Fraassen (1989, p. 226) states in *Laws and Symmetry*, 'any question about content [of a scientific theory] is, in actuality, met with an interpretation.' So, interpretation is crucial: but what is it and what does it involve? According to the standard view—or the 'ideal view', as Hoefer and Smeenk (2016) call it—the interpretation of a physical theory 'should characterize the physical possibilities allowed by the theory as well as specifying how the mathematical structures used by the theory acquire empirical content' (Hoefer and Smeenk (2016: p. 118)). This standard account of interpretation is widely accepted and is implicit in many philosophical accounts of science. A recent statement and critical analysis of it can be found in Laura Ruetsche's *Interpreting Quantum Theories* (2011). In line with the characterisations given above, she describes the standard account as asserting that 'the content of a theory is given by the set of worlds of which that theory is true' (p. 6), and 'to interpret a physical theory is to characterize the worlds possible according to that theory' (p. 7; for my own treatment, see Section 1.3.1).

In this thesis, I will adopt a conception of interpretation that aims to capture both elements hinted at in Ruetsche (2011), i.e. the theoretical and the practical (for more details, see De Haro (2020) and De Haro and Butterfield (2017)). Theory and practice seem too often to be presented as antagonistic, while they are in fact complement and require each other (see the discussion Section 1.2.1). Thus I characterise an interpretation as follows:

Interpretation: an interpretation of a theory, T, is a (partial) map, i, preserving appropriate structure, from the theory to a domain of application, D, within a possible world, W, i.e. a map $i : T \to D$.[7]

There are different kinds of interpretations, tuned to different purposes (see below; I will give more details about the semantics in Section 1.3.1).

Let me say a bit more about the 'worlds' that are assumed by standard semantic analyses, as the latter apply to scientific theories.[8] In philosophy of science, 'possible worlds' are often construed as the various instantiations of a theory, or sets of solutions of the theory's dynamical equations (sometimes also called 'models'). Some of these instantiations or sets of solutions (worlds, or models) describe a part of the actual world (for example, a solution of Newtonian gravitational theory will describe the solar system, while another solution will describe a system in another galaxy, even though there might not be a single solution of Newtonian theory that accurately describes both). Thus I will distinguish 'possible worlds', or simply 'worlds' (instantiations, models, or solutions, whether they describe some corner of our universe or not) from 'our world' or 'the actual world' (our actual universe).

This usage of the phrase 'world' for scientific theories is standard, but there are three important aspects to keep in mind: (1) Most scientific theories do not describe entire worlds, but rather *aspects* of a world:[9] a scientific theory usually applies to specific classes of experiments, observations, entities or properties. For example, Maxwell's theory describes the electromagnetic forces but not the gravitational forces; genetics studies the genetic aspects of organisms, such as the content and variation of genetic information and heredity, but not directly the organism's morphology. Thus these are treated as 'worlds' in the sense of their being complete idealisations.[10] (2) Many solutions of the theories, which are called 'worlds', have pathologies such as singularities, i.e. parts of the putative

[7] If the theory is presented as a triple of states, quantities, and dynamics, then i denotes a triple of maps, one for each component.

[8] In philosophy of physics, a distinction is sometimes made between 'kinematically possible worlds' and 'dynamically possible worlds'. Since I will take a theory to also include a dynamics, we will be mostly interested in the dynamically possible worlds.

[9] We could also call these worlds 'states of affairs' or 'state-descriptions', but I will stick to the possible worlds jargon, even though these worlds are only partial.

[10] For a discussion of some critical questions about such idealisations (for example, whether we can just *conceive* of such worlds, or whether they are *physically* possible, given what we know about modern physics), see French (2018: pp. 398-400).

world that are not described by the theory. This often means that these worlds are not complete: they have boundaries and other defects. Nevertheless, this is not a problem for the usage 'worlds'.[11] (3) Many theories, like for example general relativity, have solutions that describe worlds that do not appear to be approximations to the actual world, even though they also contain solutions that do describe aspects of the actual world rather accurately. This is entirely consistent with the interest in 'possible worlds'.

It is worth clarifying, here at the outset, some possible misleading connotations of the above conception of interpretation, which I reject:

(a) *Not necessarily a model-theoretic conception of interpretation:* While my notion of interpretation bears a resemblance with the model-theoretic notion of interpretation (of which Ruetsche presents the somewhat bleak version that she calls the 'pristine ideal'), it aims to be more general than a model-theoretic conception as usually presented. According to the latter, a theory is often identified with an entire class of models, which are each, individually, a consistent interpretation of the language of the theory (alternatively, an interpretation is the set of possible worlds in which the theory is true). On this view, a family of models constitutes the truth conditions for the theory, that is, all the possible interpretations that provide the theory with a truth value. My notion differs from this in two ways:

(i) As I already announced in Section 1.2.1, I disentangle the notions of 'model', 'possible world/domain' and 'interpretation'. Furthermore, in my conception, a theory is not a collection of models. Thus the domain of application, D, should not be confused with a 'model' of the theory, in the formal sense. Rather, the domain of application, D, is a part of the world: it is not "more theory", but rather the range of the map i is to be straightforwardly interpreted as being "in the world".[12] In other words, the elements of the domain of application, D, are not formal objects—even if it is very often useful to *present them* as formalised, specifically: as elements of sets, and as relations between the elements.

(ii) My conception of interpretation does not require truth-values or truth-conditions, because some of the interpretations I will consider (especially in De Haro and De Regt (2018)) are not literally 'true', or not true in a way that is relevant for the natural sciences. Indeed, my conception of interpretation does not always use the notion of 'truth': whether this notion is used depends on the aim for which an interpretation is used: description and explanation often appeal to truth, while understanding sometimes does not.

[11] For a discussion of singularities in connection with possible world-talk in physics, see De Haro (2019b: §2.3.2). Frisch (2005: pp. 4, 7) notices that *consistency* is built into both the semantic and the syntactic conceptions of theories, and that any account of theories in terms of possible worlds assumes that the theory's models include structures 'rich enough to represent possible worlds as complex as ours' (ibid: p. 8). For related discussions, see Wilson (2006) and Vickers (2013).

[12] Depending on the kind of interpretation one is considering, the items in this domain are natural objects, or observational or experimental procedures or activities carried out by scientists, material artefacts constructed by scientists, etc. See Chapter 6.

(b) *Applicability beyond the mathematical aspects of theories:* A partial map is a flexible notion,[13] which maps not only between mathematical objects, but also between other objects: and this is something we want. Indeed, the map has a domain (the bare theory, T) and a range (the domain of application, D), which are both sets. As already mentioned under (a), the domain of application, D, can contain such diverse items as natural entities, numbers on a scale, results of experiments, and even human actions. As such, this conception of interpretation is closer to the "mediator" conception of a model, where a model is seen as mediating between the theory and the phenomena. A map can indeed provide such a mediation, between the theory (or a particular instantiation, or sector, of the theory) and the domain of application, which often consists of the particular phenomena the theory aims to describe. And the map can be as complex as you want. These kinds of contextual details belong to interpretations that are extensions (see the paragraph just below). Similar remarks can be made about the theory, T, which need not be a formal theory, but admits diverse degrees of formalisation, adapted to the branch of science in question. Needless to say: in this thesis, I will concentrate on mathematically formulated theories, but I do not see any bar to applying this conception of interpretation to other fields.

It is important to clarify that the interpretation maps here discussed need not be taken in a rigorous mathematical sense, but allow a degree of imprecision. The degree of precision obviously depends both on the accuracy of the theory, and on the level of detail of the domain. The more formal the theory, the more precise the interpretation map can be, and thus the more accurate the representation of reality that it can provide. But an interpretation is of course never merely formal (see the points (a) and (b) above).

This conception is logically weak, in that there is very little that the existence of such a map requires. And there are various ways in which an interpretation can be made more restrictive (by imposing additional conditions on the map), and various kinds of interpretations that can be described—by further specifying the kinds of domains of application and worlds that are admissible (cf. De Haro (2020)).

(c) *Interpretation and the aims of scientific theories.* Interpretations, as aspects of scientific theories, have the same aims as the theories themselves. Usually, as noted above, interpretation is related to scientific realism: What are we committed to believe if we interpret a scientific theory realistically, i.e. as a description of the world? This goal will be explored in Chapter 2. In addition to description, however, central aims of science are also prediction, explanation, and understanding. Interpretation serves these goals as well: indeed, it is a precondition for achieving them. But it is not self-evident that these goals require the same considerations about interpretation as descriptive goals do. As to the aim of understanding, some realists will typically be inclined to believe that there is a unique interpretation of the theory that provides understanding of the physical world by virtue of presenting the set of possible worlds of which the theory is true. Such realists

[13] Although flexibility can also have some disadvantages, I submit that my conception does not have these disadvantages, and that the kind of flexibility discussed here is something we want: see my comments, below, on the logical weakness of the conception of interpretation, and on the aims of an interpretation. This conception of interpretation is flexible, but appropriate strengthenings can be adopted: depending on the aim of the interpretation.

may specify additional constraints; for example, they might require that the interpretation is in accordance with particular metaphysical presuppositions. Similar considerations will be endorsed for the predictive and the explanatory power of the theory. What such a priori considerations sometimes ignore, however, is the question of how an actual physical theory achieves these goals in practice. As Ruetsche (2011, p. 5) observes:

'The lofty debate is conducted in terms that obscure how real theories possess the virtues they possess. It is often a theory *under an interpretation* that predicts, explains, and promotes understanding. To the disappointment of the realist, there may not be a single interpretation under which a given theory accomplishes all those things.'

We will later get back to the question of whether scientific realism is necessarily committed to the idea that each (bare) theory must have a single interpretation (I do not believe it is).

In any case, the framework of intensions and extensions that I will advocate in Section 1.2.4 accounts for the fact that a single theory may admit of many (and even mutually exclusive) interpretations. The kinds of interpretations, required for the other goals mentioned (prediction, explanation, and understanding), indeed do not need to be realistic, or accurate, scientific representation: but the kind of reference picked out by the interpretation may be tuned to these aims of prediction, explanation, and understanding. Thus we must carefully distinguish the *representation of a given system* (i.e. the accurate scientific description of a *given* system, according to appropriate standards of empirical adequacy) from the theory's *reference*, with which the above conception of interpretation is concerned. And the question of representation still differs from the question of the truth of the theory: for both realists and constructive empiricists aim at representation, though their construals of these representations differ.

Thus the above definition of interpretation as a partial map may be strengthened, depending on the aim for which a theory is used. For example, if the aim is accurate description of a system (which will be our aim in Sections 2, 3, and 4), the interpretation is required to be empirically adequate for that system (see Section 3.2.1 for a discussion).

1.2.4 Interpretations of theories and models

The position about interpretation introduced in the previous Section uses a mainstream view of meaning: namely, classical referential semantics. Thus we will assign references in the world to the elements of a theory—viz. to the states and the quantities, and to the formulas built from them—and likewise for models.[14] One advantage of the framework of referential semantics is that both realists and constructive empiricists can agree about the interpretation of a theory or model, in other words about its basic ontology ('the picture of the world drawn by the theory', to use van Fraassen's (1980: pp. 14, 43, 57) words),

[14] Referential semantics, at least in its mainstream presentation, may admittedly have some limitations in that it sometimes does not pay sufficient attention to scientific practices, norms, and non-linguistic skills (as witnessed by works, such as Montague (1970), that are supposedly about "pragmatics" in the philosophy of language, but remain largely formal). However, there is *no incompatibility* between using a referential semantics for theories, and other lines of work that emphasise complementary aspects such as practices, norms, and non-linguistic skills. For a defence of this irenic perspective, see Lewis (1975: p. 35), De Haro and Butterfield (2018: Section 2.1.3), and De Haro and De Regt (2018: pp. 633-637).

even though they have different degrees of belief in the entities that the ontology of the theory postulates—and indeed they will have a different conception of what such a map entails. We will see illustrations of this point below, and in Section 4.2.4.

Thus I will model referential semantics by interpretation maps, developing the conception introduced in the previous Section. Recall that an *interpretation* is a set of partial maps, preserving appropriate structure, from the theory to the world.[15] The interpretation fixes the reference of the terms in the theory. If the theory is a triple of states, quantities, and dynamics, viz. $T = \langle \mathcal{S}, \mathcal{Q}, \mathcal{D} \rangle$, then the interpretation maps are also a triple, one for each item in the triple: however, we will not often need to make this explicit. Using different interpretation maps, the same theory can describe different domains of the world, and even different possible worlds.

Likewise for models, an interpretation is a partial map, $i : M \to D$, preserving appropriate structure, from the model to the world. (Since a model M is a representation of a *bare theory,* T, rather than of an interpreted theory: and so, the model requires its own interpretation).

Notice that the choice of 'the part of the structure that is common to all of the theory's models' goes into the definition of a model, and singles out the core physics that is described by the theory, because it is represented in all its models: namely, it is in the distinction between the *model root* and the *specific structure* from Section 1.2.2: and this choice constrains the kinds of systems the model is able to describe.

The way to determine the relevant structure is interpretative not formal. For example, only experiments can tell us that massless vector fields correctly describe photons. Once that question is experimentally settled, one can take a model root including a massless vector field to describe a photon. But once we have that model root with its massless vector field, we can strip it of its photon interpretation and use the same model root to describe whichever other particle exhibits degrees of freedom of the same kind.

Interpretations thus defined are very general. To specify them further, I endorse, as I announced before, a more specific framework, viz. intensional semantics (cf. Lewis (1980), Carnap (1947: pp. 177-182; 1963: pp. 889-908)). In this framework, the notion of 'linguistic meaning' (for us, in the context of scientific theories: an 'interpretation') is taken to be ambiguous between what Frege (1892) called 'sense' and what he called 'reference', here called 'intension' and 'extension' respectively. The intension is the linguistic meaning of a term, while an extension is the worldly reference of the term, relative to a single possible world (with all of its contingent details). Thus Lewis (1987: pp. 22-27) defines intensions as maps from n-tuples of sequences of items—he calls such a sequence an 'index'—to extensions, e.g. the truth-values of sentences. And, as Lewis remarks, the framework also applies if the indices are construed as models consisting of states representing possible worlds (ibid, p. 23): see also Carnap (1947: pp. 177-182; 1963: pp. 889-908). Here, I will

[15] I will occasionally use the phrase 'physical interpretation', in order to distinguish this interpretation (i.e. what the items of the theory refer to in the world) from the linguistic interpretation (i.e. the language of—an abstract form of—mathematical physics in which the theory is formulated). The condition that the map only needs to be partial means that there need not be a reference for all its arguments, i.e. for all the terms in the bare theory. This is because some interpretations may map to worlds with less structure than envisaged by the bare theory. See De Haro and Butterfield (2018: §2.4). However, this point will not be very relevant in this thesis.

discuss a variation of this framework that involves: (a) mapping from scientific theories and models (usually presented as triples, so that there are three such maps) rather than from sequences of general linguistic items; (b) a simplification: namely, modelling both intensions and extensions by interpretation maps, rather than defining an extension as the (set of) objects/truth value(s) and an intension as a map to extensions.

Thus both intensions and extensions are structure-preserving partial maps, which I will dub i_{Int} and i_{Ext}, from a bare theory or model to a domain of application relative to a possible world. The difference between the intension and the extension is in the kind of domain of application: explicitly, for states: an intension maps a state to a generic property (or physical arrangement) of a system mirroring the defining properties of the mapped state (and likewise for quantities and dynamics). Thus the image of the state and the domain of application abstract from contingencies such as the detailed arrangement of the system and how the system is measured (so that the interpretation applies to all possible worlds that are described by the theory or model). By contrast, in the case of an extension, the image and domain of application are a fully concrete physical system: usually including also a specific context of experiment or description, and all the contingent details that are involved in applying a scientific theory to a concrete system (further discussion follows, in Section 1.3.1). For example, 'the voltmeter which I have in my hand reads 220 V'. This sentence will clearly have different meanings when uttered by different scientists, and even with different voltmeters because of measurement errors; and so, it is to be interpreted differently—i.e. to receive a different extension—in different contexts.

Let me illustrate this notion of interpretation in the simple case of Maxwell's theory in vacuum, and on $\mathbb{R}^4$, mentioned in Section 1.2.2: where the state-space $\mathcal{S}$ is a suitable set of 2-forms F, the quantities $\mathcal{Q}$ contain a distinguished quantity, the stress-energy tensor, with components $T_{\mu\nu}[F]$, and the dynamics is the condition that the 2-forms F are both closed and co-closed (more details in De Haro (2019b)). Under the standard electromagnetic interpretation, the Faraday tensor F is the coordinate-free presentation of a tensor whose components are interpreted as electric and magnetic fields, so that the intensional interpretation map is:

$$\begin{aligned} i(F) \;&=\; \text{`a coordinate-free specification of an electric and magnetic configuration} \\ &\qquad\qquad\qquad\qquad \text{in vacuum'}\;. \end{aligned} \tag{1.3}$$

The key interpretative words are here 'coordinate-free specification' and 'electric and magnetic configurations', which we (uncontroversially) assume already have established meanings, for example as correlating with certain experimental procedures that we use to measure the fields, or as corresponding to certain properties of fields that we are familiar with in our world (such as the polarisation of light waves, and how they interact with other entities).

This meaning has come about through centuries of experimentation, instrumentation, and theorising about electro-magnetic fields. Different historical epochs may of course have different ways of manipulating and measuring these fields, but referential semantics assumes such reference to be clear, the more because the theory is sufficiently well established, as Maxwell's electromagnetic theory indeed is—and this reference is construed as saying that there are such things as electric and magnetic fields in the world (if one is a realist) or that there are indeed appearances of phenomena of electric and magnetic fields (if one is an empiricist). Likewise for the phrase 'coordinate-free specification': it summarises the properties of the fields under changes of the frame of reference, i.e. properties like 'the electric field is augmented in the directions perpendicular to the motion by an amount given by the Lorentz factor'. Such properties again uncontroversially correlate with particular phenomena in the world.[16]

Notice the following two properties of the interpretation map Eq. (1.3). First, recall that I defined interpretation maps $i : M \to D$ as appropriately structure-preserving. Therefore the phrase 'coordinate-free specification' in Eq. (1.3) implements this appropriately structure-preserving character of the Poincaré symmetries. Namely, Poincaré symmetries, which leave the 2-tensor F invariant and transform its components according to the standard Poincaré transformations, have a "shadow" in the domain of the world, D: this shadow contains the ordinary effects, of Lorentz expansions and contractions of the fields, that I referred to in the previous paragraph. Thus the interpretation map Eq. (1.3) correctly preserves symmetries.

Second, the interpretation Eq. (1.3) is an intension: for it is valid at any possible world that instantiates electric and magnetic fields whose corresponding Faraday tensor satisfies the definitions of our states (viz. being a square-integrable, smooth 2-form). In other words, it does not depend on a specification of F—as being, for example, a collection of travelling waves with certain polarizations. Such "generic" interpretations hold at every possible world that is described by the theory. In order to get an extension, we should specify the particular 2-form F that we are mapping. This would entail giving the details of its functional form: which then corresponds, under the interpretation map, to further specifications in a particular world, like 'a wave travelling in such and such direction, with polarizations at such and such angles'. But, as I mentioned above, this way of modelling intensional semantics accommodates for boths kinds of interpretations.

There is also an interpretation map for quantities, which for example maps the 00-component of the stress-energy tensor:
$i(T_{00})$ = 'the electromagnetic energy density of the system'. Again, this map is an intension—it holds at all possible worlds at which Maxwell's theory applies, and in which a frame of reference has been specified, relative to which the 00-components of the tensor are taken. To specify the extension of T_{00}, we should provide additional details such as where the electromagnetic energy is produced from, how it interacts with its environment and is measured, etc.

The idea of theoretical equivalence in Chapter 4 will be, roughly, that two models are isomorphic as regards their formal structure—as in their model roots, m—and also match

[16] Any sensible theory can of course be formulated in generally covariant form, and so it is not the coordinate-free formulation of the theory *per se* that secures the preservation of the symmetries. Rather, the dynamical symmetries are preserved because the interpretation map $i : M \to D$ correctly maps the electric and magnetic variables to the electric and magnetic fields, for any inertial frame, whereas it would not in general do so for non-inertial frames. Thus the components of F also have their own interpretations, once a system of coordinates is chosen. For a discussion of the contrast between general covariance and invariance under a subgroup of the diffeomorphism group, see e.g. Pooley (2017: pp. 114-118).

as regards their interpretation. Thus I will now define an interpretation that depends only on the *model root, and not the specific structure*:

Internal interpretation: an interpretation that maps all of and only the model root, regardless of the specific structure of the model. Since internal interpretations map the model roots (and they may map specific structure only in so far as it appears in the model root), it will often be clearer to restrict the internal interpretation map of models to the model root, and write: $i : m \to D$. Internal interpretations obviously also apply to bare theories, T. (This formal notion will be further characterised, in the context of theory construction in physics, in Section 4.3.2).

I shall contrast this with other interpretations that also map the specific structure (or that involve coupling to other theories): which I will dub *external interpretations* (a fuller exposition is given in Chapter 4).

I use 'internal interpretation' in the singular here because we will normally consider a single interpretation. However, a given model can have multiple internal interpretations.[17] For example, take a model whose dynamics is the heat equation for a real, non-negative function over space and time that we denote as: $\rho : \mathbb{R}^4 \to \mathbb{R}_{\geq 0}$, and is given by: $(t, \mathbf{x}) \mapsto \rho\,(t, \mathbf{x})$. Take the state-space of the model to be the space of configurations of the function $\rho(t, \mathbf{x})$ (and let the quantities of the model also be constructed from ρ, its powers, and its derivatives and integrals over space and time—so that we now have a triple of states, quantities, and dynamics). This model can be interpreted as describing the spread of heat in time in a given region of space, or as a diffusion equation for Brownian particles in a medium. With these definitions, there is here a single model (in fact, a theory), with no specific structure. Both interpretations are internal, and yet they differ.[18]

I also assume that the model root is rich enough that an internal interpretation allows the specification of a set of possible worlds that instantiate the solutions of the equations of the model (or, at least, the interpretation specifies suitable domains of application within such worlds). Nevertheless, for simplicity, I will talk about 'the possible world specified by the model' rather than about the set of such worlds.

[17] The following example is also mentioned by van Fraassen (2014: p. 279).

[18] This formulation of the internal interpretation maps only the model root: i.e. the interpretation map, which is a partial map, is *not defined* on the parts of the specific structure which are not part of the model root. This gives a clear-cut criterion for when an interpretation is internal. The formulation in Dieks et al. (2015: pp. 208-210) might, on the other hand, give the (mistaken) impression that an 'internal viewpoint' somehow requires an interpretation that is constructed with no other information except for the theory. But this is of course not how interpretations come about in physics (as I also emphasised for the example of Maxwell's theory): they are always constructed against the background of past theories. This methodological point is expounded in Chapter 6, which gives three interpretative tools that can be used to construct *internal* interpretations, i.e. without interpreting the specific structure and without coupling the theory to an already interpreted theory (see also Dewar (2017: §6) for some excellent comments on related issues). One of the tools is indeed internal to the *theory*, but three others relate the theory (conceptually) to already existing theories, while still being 'internal' in my sense. Thus, just as dualities are used to internally interpret theories, other inter-theoretic relations are also used.

1.3 Meaning and Continuity: between Extensions and Intensions

One of the main jobs that referential semantics is supposed to do for the debate over scientific realism that I will review in the next Chapter is securing the continuity—or the *dis*continuity, depending on which side of the debate you are on—between theories from different historical epochs. Thus this Section will apply some further—indeed, elementary—aspects of classical intensional semantics to scientific theories.

1.3.1 Meaning as a matter of intension and extension

This Section introduces the distinction between intensions and extensions that our analysis in Chapter 2 will need. Then it introduces a notion of extensional equivalence that will play a central role in what follows, and briefly addresses the question of how extensions are determined—which Chapter 2 addresses in more detail.

Recall that, quite generally, the *intension* of a *term* is the term's linguistic meaning (cf. the Fregean sense), as described by the theory, and thus relative to all possible worlds described by the theory.[19] The term's *extension* is its worldly reference (cf. the Fregean 'reference'), relative to a single possible world, i.e. the (set of) thing(s) or entity(ies) being referred to, relative to a given possible world (with all of its contingent details). Thus 'Evening star' and 'Morning star' have different intensions (i.e. diferent linguistic meanings, and the stars they refer to are different in some possible worlds), but they have the same extension in our world, viz. the planet Venus—thus it is a contingent fact about our world that, while their intensions differ, their extension is the same.

In the case of scientific theories, by 'terms' we mean the individual expressions, be they simple or composite, i.e. bits of scientific language (e.g. mathematical or other technical language) that appear in the theory and that need to be interpreted: for example, words for entities or properties ('bacteria', 'DNA strand', 'cell mobility', etc.), expressions on one side of an equation, mathematical symbols for force, mass, charge, fields, etc.[20]

Standard texts on semantics (cf. footnote 19) distinguish between three kinds of linguistic expressions, which Carnap calls 'designators': namely, sentences, individual expressions (such as names, descriptions or words for properties[21]) and predicates (such as

[19] For more about intensional semantics, see Carnap (1947: pp. 177-182) and Lewis (1980). For an excellent introduction to intensional semantics in connection with sentence structure, see Heim and Kratzer (1988: Chapters 2 and 12). For an overview of theories of meaning, see Speaks (2019).

[20] Recall that my interpretation maps i from Section 1.2.3 maps theories and models directly to the world, and that I reject (in point (b)) the idea that all the relevant aspects of a theory should, or even can, be fruitfully formalised. Thus I also reject the idea that absolutely everything about a scientific theory can be fruitfully reduced to a scientific 'language'. Therefore, my treatment in this Section (but not the next) is slightly limited by the restriction to language; this limitation is of course imposed by the need to present classical semantics in a standard form, but it will not affect my argument. See also my comments on 'the language that we use to talk about the world', later in this Section.

[21] I am here giving a uniform treatment of names, descriptions, and words for *properties* (i.e. I regard all of them as individual expressions) because scientific theories tend to treat properties (for example, 'mass', 'electric charge') as abstract entities. Thus, the term 'mass' can appear on the left-hand side of an equation, e.g. $M = 50$ kg, as if it were an object. In official expositions of semantics, properties are

words for relations and verb phrases). Different kinds of extensions and intensions are assigned to each of these kinds of linguistic expressions. The case just discussed—the referents of 'terms', or individual expressions—are indeed the most interesting cases for scientific realism: since we are interested in whether terms such as 'aether', 'caloric', etc. refer. But let me briefly introduce the semantics of the two other kinds of linguistic expressions—

The extension of a *sentence* (for example, a statement of a law, or an equation) is its truth-value, while its intension is the corresponding proposition (alternatively, a map from the set of worlds described by the theory to truth-values). The extension of a *predicator* (for example, the verb phrases 'never decreases over time' or 'is equal to zero') is a *characteristic function* from a set of individuals to truth values (alternatively, it is the set of individuals itself), while its intension is a function from the set of worlds to its characteristic function (alternatively, the corresponding property). The referents of sentences will be relevant when we discuss truth, in Section 2.3. For the moment, we focus on individual expressions.

The extensions of terms thus denote "local", this-worldly items that are contingently related to the term used—contingent in the sense that the term gives that extension only in a specific possible world described by the theory, depending on its circumstantial details. Intensions, on the other hand, denote "global" items, and thus apply to all possible worlds.[22]

Our main concern will be with terms that, like 'Morning star' and 'Evening star', have different intensions but the same extension. Furthermore, we will be interested in cases in which the equality is more general: it holds not only in a single world (i.e. instantiation, solution, or model of the theory), but in various possible worlds. Thus it will be useful to introduce the following notion:

Extensional equivalence:[23] two terms are extensionally equivalent, with respect to a set of possible worlds described by the theory, iff they have the same extension in each of those worlds (even though their intensions may differ). Two terms that are extensionally equivalent are often said to be 'coextensive' or to 'co-refer'.[24]

How does one know that the extensions are the same in a given world? How does one know that 'Morning star' and 'Evening star' have the same referent, namely Venus? This is, in astronomy as in other sciences, partly a theoretical and partly an empirical question: one works out deductively the empirical consequences of the use of the term in the theory, applied to the particular empirical circumstances, including the experimental and observational accuracy, and sometimes also the context of utterance. The identification

treated in verb phrases ('is heavy'). It is a grammatical matter of no consequence whether a property is represented by an individual expression, i.e. a singular term, or as a predicate (the two are each others' duals): and my usage has the advantage just mentioned. Throughout this thesis, we are interested in individual expressions and sentences not verb phrases.

[22] Thus Lewis (1980: pp. 22-27) defines intensions as maps from linguistic terms (or sequences thereof) to extensions.

[23] See also Carnap (1947: pp. 14, 18). An example: Norton (2012: p. 211), without using the word 'extension', discusses an example: '[t]he "caloric" of caloric theory refers to the same thing as the "heat" of thermodynamics, but in the confines of situations in which there is no interchange of heat and work.'

[24] 'Extensional equivalence' is, in philosophy, usually used for the actual world only.

of ‘Evening star’ obviously depends on various details: not only does Venus need to then be on one side of the Sun, trailing it in its course; also, the sky needs to be sufficiently clear and dark. More generally, the planet Venus is open to a more detailed investigation through spectroscopic observations, radio echo, and space probes.

I will next argue that extensional equivalence can be expressed using the *language that we use to talk about the world,* but let me first say a few words about this language. We describe the domain of application partly linguistically (e.g. ‘the nucleus of an eukaryotic cell is surrounded by a nuclear envelope that is perforated by nuclear pores’) and express identity in terms of this language. This language is an aid (e.g. words used to mention objects) to describe the intended physical things and properties, and it is in general different from the technical, formal, or mathematical language of the theory. It is ‘the language of experimental physicists (and lay people)’ used to describe the world, and it may contain *extra-linguistic elements* such as diagrams, images, ostension, etc.[25]

Kuhn and Feyerabend objected to the comparison of the meanings of the terms of different scientific theories. It does not help to say, the argument goes, that this comparison is empirical, because experimentation and observation are ‘theory-laden’. Thus the Kuhn-Feyerabend critiques of meaning might lead one to think that, if the language that we use to talk about them is theory-laden, the criterion of individuation of entities in the domain of application is problematic, and so we cannot compare theories to assess their extensional equivalence. Such a comparison would require a theory-free, neutral language—which we do not have.

But I will argue that incommensurability of the Kuhn-Feyerabend type applies only to intensions (and only partially so) and not extensions. Specifically, the theory-ladenness of empirical data is not an obstacle to assessing extensional equivalence. This is because, as I will discuss in the next Chapter, the theories that we compare in the debate over scientific realism are related by various formal (i.e. mathematical, abstract) and-or predictive, relations of linkage or correspondence: such as the numerical and formal agreement between theories, in various limits and specific cases. This formal and-or predictive linkage can be taken to be conceptually independent of the specific domain of application. But since the domain of application reflects the theory (it is ‘theory-laden’), linkage can be used as a guide to *compare* extensions between theories, so that there is no incommensurability. Furthermore, we will see in Section 2.2 how the notion of *correspondence* can be used to also compare conceptual and material aspects of theories.

Therefore, for theories like the ones we are interested in here, which are related by linkage relations, one can combine three elements to compare the extensions between them: (1) the linkage relation between the *theories,* (2) the relation between each *theory and its domain of application,* (3) the *empirical data* about the domain(s).

Thus my account will stress the need to combine theoretical and empirical (including

[25] Galison uses the metaphor of the ‘trading zone’ to understand how scientists from different traditions engage with each other, developing pidgins or creole languages: ‘people have and exploit an ability to restrict and alter meanings in such a way as to create local senses of terms that speakers of both “parent” languages recognize as intermediate between the two. The resulting pidgin or creole is neither absolutely dependent on nor absolutely independent of global meanings’ (1997: p. 47). For examples of what I have called ‘the language that scientists use to describe the world’, see De Haro (2019: Section 5).

material) aspects of correspondence in order to establish extensional equivalence.[26]

In particular, a putative "theory-free" language is neither needed nor desired to compare the extensions.[27] Also, incommensurability is usually only *partial,* and the domains of application are often partly overlapping and partly distinct.

Galison (1997: pp. 782, 804-805) has argued forcefully (and, I think, convincingly) against the need for a theory-neutral language to secure continuity and communication between scientists working in different traditions or paradigms. On his account, it is the 'trading zone' at disciplinary boundaries (with its pidgins and creole languages) that secures successful communication (p. 838):

> Superficially, the handing of charts, tubes, and circuit boards back and forth across the various interexperimental cultural divides might look like a case of worlds crossing without meeting. This description, however, would do violence to the expressed experience of the participants. They are not without resources to communicate, but the communication takes place piecemeal, not in a global translation of cultures, and not through the establishment of a universal language based on sense-data... laboratory and theoretical work are not about translation, they are about coordination between action and belief.[28]

A consequence of the above is that determining the extension of a theory is never a "purely factual" matter. If there was a "theory-free language" that allowed us to talk about extensions, then we might talk about the facts with no reference to theoretical concepts at all. But we have just seen that this is incorrect, since determining extensions and comparing them is partly a theoretical matter. It surely employs the concepts of the theory, and it even involves the consideration of counterfactual situations, which e.g. allow one to consider what would happen—what difference would it make—if one changed some aspects of the system considered, or some of the circumstantial physical conditions.

This is one reason why the extensional scientific realism that I will develop in Chapter 2 is a form of *realism*: even though the extensions are themselves factual, these facts are not brute or taken in isolation, but in their relation to other facts and concepts: determining an extension is both an empirical and a conceptual matter.

Extensional equivalence will give, in the next Chapter, a notion of referential (i.e. extensional not intensional) continuity between theories that we need in order to compare the truth of theories.

[26] The aspects of correspondence that establish extensional equivalence are in broad agreement with the three elements that, according to Galison (1997: pp. 799-800), secure the possibility of interaction and communication between different scientists: namely, theory, experiment, and instrumentation: 'we do not expect to see the abrupt changes of theory, experimentation, and instrumentation occur simultaneously... When a radically new theory is introduced, we would expect experimenters to use their best-established instruments, not their unproven ones.'

[27] It is always good if one's "realist" point is also made by an anti-realist, even though otherwise our arguments differ substantially: 'Its central flaw [of the pessimistic conclusion by Quine, Kuhn, and Feyerabend, to the effect that theory succession is not quite a rational process] lies in its presumption that rational choice can be made between theories only if those theories can be translated into one another's language or into a third "theory-neutral" language' (Laudan (1977: p. 142)). For a summary of the argument about theory-ladenness, see (ibid: pp. 141-142).

[28] See also Galison (1997: p. 790), where he argues that experimentation can give continuity even in the presence of Gestalt switches.

1.3.2 Extensions are determined by intensions, circumstances, and context

One main goal of the discussion of the reference of terms in debates over scientific realism is to give a reference-based criterion of continuity between scientific theories of different historical epochs.

In this Section, I will apply the general intensional semantics articulated in Section 1.3.1 to scientific theories. This Section focusses on the general elements that determine extensions; Section 2.2 will give more detail and examples, and connects the discussion to the literature about scientific realism.

In classical semantics, extensions are determined by intensions and, in addition, context of utterance, and circumstances of evaluation. The context and the circumstances are both made up of indices. Roughly speaking, the context of evaluation has to do with the *speaker* (or where and when the expression is uttered) while the circumstances have to do with the particular state that is relevant for the expression, given the context. Thus roughly, the distinction is between 'who speaks and where and when' vs. 'what is being said and about what'. Rather than going further into the semantic details here,[29] let me just say how I will use these expressions for scientific theories: namely, the relevant contrast will be between 'the scientists and their activities' vs. 'the theory and its set of models'.

Thus I will use the phrase *circumstantial physical conditions* for the choice of state of the world (see Section 1.3.1), together with other physical conditions relevant in determining an extension, and which are modelled by, or have a correlate in, the theory or model. These are, for example, boundary conditions and other conditions about the empirical situation considered: including measurement or observational accuracies, the approximations made, in so far as these correspond to physical properties of the system considered (for example, low speed, or large orbital angular momentum, etc.). They have a formal, mathematical, or technical counterpart in the theory, which I will call the theory's *level of abstraction*.[30] The level of abstraction depends on the properties of the states and quantities that are relevant for a specific system, together with the way in which other practical aspects are implemented in a theory or model. Thus the level of abstraction is the formal correlate that reproduces (and thus determines) the circumstantial physical conditions under which the theory is applicable.[31] (I will use the phrase 'the level of abstraction determines the circumstantial physical conditions' in this sense.)

In what follows, I will use the phrase *domain of application* to refer to a possible world (in the usage introduced in Section 1.3.1) with specific circumstantial physical conditions. Since the circumstantial physical conditions include approximations (see below),

[29] For an introduction and examples, see Speaks (2019: Sections 2.1.4 and 2.3). See also Kaplan (1977: pp. 494-495), Lewis (1980) and Perry (1979).

[30] I take term 'level of abstraction' from Floridi (2011: Chapter 3). Levels of abstraction have an objective character, in so far as they determine circumstantial physical conditions which belong to a physical system and to which the theory applies.

[31] Levels of abstraction are in the realm of language, while the circumstantial physical conditions are in the world. (Both are usually expressed linguistically).

this means that a domain of application of a theory, unlike a possible world with a specified state, is usually not an exact solution of the theory's dynamical equations, but only an approximate solution.[32]

Let me here summarise, without aiming for completeness, some of the content of the various *levels of abstraction* (and the *circumstantial physical conditions* they determine) involved in determining the extension from the intension:

- The specification of a model, viz. the relevant kinematic and dynamical situation (including choices of initial and boundary conditions, velocities, energies, distances, assumptions about the population, environmental conditions, chemical composition, etc.).
- The values of the theory's and the model's free parameters (such as mass, coupling strength, chemical concentration, etc.).
- Extra-theoretical facts, in so far as these are modelled by, or incorporated into, the theory or model: including the allowed measurement or observational errors; other influences, such as noise, not described by the theory; aspects of the experimental—including material—conditions, weather conditions, etc.
- Approximations and idealisations (both theoretical and empirical) carried out in order to accommodate all of the above, i.e. to make the theory empirically adequate under the stated conditions and accuracies.[33]

The phrase *context (of application)* will denote all those aspects that are compatible with, but are not fully modelled by, the theory or the relevant models, and that bear on how reference is established in practice. This includes scientific methods and practices, material realisation of models and experimental procedures, ostention and intentions of speakers, influences (or lack of them) that are not described by the theory, etc., in so far as they are not encoded in or modelled by the theory or the relevant models. The main context of application that I will discuss in this thesis is the material realisation of an item (e.g. a chemical substance, a particle) in an experiment, which will secure partial continuity.

[32] In the discussions of correspondence between two theories, T and T' (or reduction of one to the other), the literature often uses the notation 'T^*' to indicate an intermediate theory, T^*, that relates T' and T, since they are not *directly* in correspondence (see e.g. Lewis (1970: pp. 441, 445)). I will not use this notation, because it assumes that T' corresponds to T, via T^*, *globally,* i.e. at the level of the theory. But this is of course often not the case, since one often reduces *models* (rather than theories) to models, and each model has its own circumstantial physical conditions, which may be inconsistent with those of some other model. Thus I will compare *domains of application,* which are worlds, each with its own circumstantial physical conditions. Because the circumstantial physical conditions for each of the worlds are already included in the domain of application, I do not need to consider the intermediary T^*; as I said before, such a theory need not exist.

[33] For more on this, see Chapter 3.

Scientific practice has been the focus of several prominent accounts of science, especially the "Stanford school",[34] which range from anti-realist to realist views.[35] The focus on scientific practice has sometimes been accompanied by a repudiation, or at least a critique, of theory. However, as I hope to show in the next Chapter, pragmatics need not militate against the use of theory. I will argue that the critics are right in insisting that theory alone is not enough. But this does not militate against scientific realism: for, as I hope to show, there is a sensible scientific realist account that makes use of both theory and practice.

Given a specific level of abstraction, and often also a context of application, the extensions of two theories with very different intensions can be identical. For a given level of abstraction that determines a specific set of physical conditions, the two theories refer to the same physical system. In the next Chapter, I will discuss in more detail various ways in which sameness is established.

[34] See, for example, Galison (1997), who does not easily qualify as either a realist or an anti-realist (he qualifies his own position as a 'historicized neo-Kantianism'). As can be seen from my earlier quotations of Galison, my account could be seen as adding a semantic/truth dimension to the aspects of theory, experimentation, and instrumentation that he focusses on (as a result of which our accounts of course differ widely).

[35] Besides the Stanford school, important realist accounts are Cartwright (1983), Hacking (1983), Massimi (2016, 2018), Chang (2012, 2018), and Radder (1991, 2012 [1984]). The accounts of Massimi, Chang, Radder, and mine are, in several aspects, kindred accounts: however, there is much in which they differ. There are, for example, some structural similarities with Massimi's account (e.g. between her 'context of use and context of assessment' (2016: pp. 356-357) and my 'circumstantial physical conditions and context of application'), but there are many differences in content—more than I can discuss here. Also, my account agrees with Chang (2018: pp. 176, 179) that theories can contradict one another (in some ways) and still be true (in other ways), but does not need to endorse Chang's pragmatism, or the metaphysics that underpins his pluralism (2018: pp. 181-184). Radder's (2012 [1984]) 'referential realism' is proposed as a scientific realist account that focusses on scientific practice and is, in some aspects, close to my account. A detailed comparison cannot be given here (for some details, see Section 2.2.2). One important difference is that Radder believes that there is no conceptual correspondence between at least some of the main theories at issue in the debate over scientific realism, like for example classical and quantum mechanics (however, see footnote 18): 'An important result of analyses of the actual historical development of the natural sciences is the insight that there have been big and radical conceptual discontinuities' (p. 83). (Radder (p. 89) of course does not deny that there are also continuities).

Chapter 2

An Extensional Scientific Realism

Apparently we do not have any way of knowing for sure... that science is true, or probable, or that it is getting closer to the truth (Larry Laudan, 1977).

All of us would like realism to be true; we would like to think that science works because it has got a grip on how things really are (Larry Laudan, 1981).

This Chapter aims to point to a relatively simple answer to the pessimistic meta-induction argument against scientific realism: an answer that is based on an application of intensional semantics. While this Chapter focusses on answering the pessimistic meta-induction argument, the application offers what I take to be a credible scientific realist position. The position builds on previous accounts by Psillos (1999), Chakravartty (2007), Radder (1984), Hacking (1983), and especially Kitcher (1993): but it has, so far as I know, not been articulated before.[1]

The position is a relatively straightforward application to scientific theories of the distinction between extensions and intensions, and it can be stated in simple terms: *We are justified in being scientific realists about extensions but not about intensions.* This is inspired by the logical empiricists' reply to anti-reductionist charges (due to Kuhn and Feyerabend) that meanings are incommensurable. The logical empiricists' reply was to admit that meanings (intensions) are *partly* incommensurable: but to argue that extensions are *continuous* across theory change.[2]

While I am not wedded to the cogency of the empiricists' reply to the anti-*reductionist* challenge, I will argue that one can base on it a good reply to the anti-*realist* challenge. Thus my aim here is neither to defend reductionism, nor to revive the logical empiricists' doctrines—since I am interested in developing a *realist* account of science. Rather, I propose that, with appropriate modifications, Nagel's basic reply to Feyerabend—that, despite the discontinuity of the intensions, the extensions are continuous—is true, and deserves development as a reply to anti-realism. Thus my realism is a modest one: I accept the truth of a *partial* incommensurability thesis for intensions, while I argue for the *continuity* of extensions.

[1] For a recent review of the debate over scientific realism, see Psillos (2018: pp. 22-32).

[2] See Scheffler (1967: pp. 60-64) and Nagel (1979: pp. 914-915). For the relation between realism and logical empiricism (in terms of the 'emergent realist tendencies in logical empiricism'), see Neuber (2018).

The cited accounts of scientific realism by Kitcher, Psillos, and Chakravartty are sophisticated and have, I believe, many elements of truth in them that mesh with, and are complementary to, my own considerations. Thus Kitcher (1993) and Psillos (1999) are right to point out that a scientific theory is not a single monolithic entity that can be an object of belief: for it consists of different kinds of statements and laws, models of various sorts, experimental practices, etc., which should receive various degrees of confidence.

More controversially, Psillos and Kitcher have argued that one should differentiate between, roughly, different theory "parts". For example, Psillos proposes that one should distinguish between core and idle elements of theories: so that the theoretical elements which contributed to the success of past theories turn out to be those core elements which were retained in subsequent theories, while those that 'were not essential for the success of the theory were treated with suspicion' (1999: p. 107). Stanford (2006: pp. 153) has pointed out (rightly, I think) that past scientists held many beliefs, such as the belief in the aether, that they thought were core parts of their theories, but which nevertheless were not retained by later theories. Thus a more conceptual distinction is needed: one that applies more homogeneously to a theory.

The distinction between extensions and intensions allows us to say that, at any given moment in time, and in so far as a scientific theory is empirically successful and the scientists have good reasons to use the various terms they use, the terms of a theory refer extensionally, and *homogeneously, for all the terms of the theory* (where, by 'homogeneous', I will mean that in general one does not need to distinguish between different occasions of utterance, and also not between different theory parts). And to respond to the pessimistic meta-induction argument, which is concerned with the problem of referential correspondence of different theories, we will say that there is an extensional correspondence between well-confirmed theories that succeed one another, even though some of their terms are intensionally inequivalent: and that this extensional correspondence is sufficient to secure the conceptual continuity that is disputed by the pessimistic meta-induction argument.[3]

The resulting scientific realist proposal is a modest one: as realists, we only need to commit ourselves to extensions, and not to intensions. This will lead in to a notion of approximate and contingent truth.

2.1 Scientific Realism and its Critiques

My motto in this Section is: listen to your enemy. Or, in more equable terms: take the criticisms of scientific realism seriously, and make them a starting point for scientific realism. In this Section, I will first state the pessimistic meta-induction argument. Then

[3] Stein (1989: p. 57) has argued for the existence of terms, like 'atom', that are retained, but change their meaning, in later theories. This is of course familiar from Feyerabend's critique of meaning, discussed in Section 1.3.1, which focussed on the different meanings of 'mass', in special relativity and in classical mechanics. And that the word 'planet' changed meaning during the scientific revolution is of course an example of incommensurability in Kuhn (1962: p. 115). These kinds of cases are resolved using the same tools of reference that I develop in this Chapter, and I will treat the example of 'mass' in Section 2.2.3. I agree with Stein that the discussions of reference have not always been conducted clearly in the literature: but, as I will argue, I disagree with Stein's (1989: p. 57) dismissal of the notion.

I will agree with some aspects of these critiques, and indicate how they can be used to develop a version of scientific realism that is free of these problems.

I will defend a version of scientific realism that is committed to scientific theories that give an empirically adequate description of some domain of application. Roughly speaking, scientific realism can be characterised as follows:

Scientific realism is the belief in the (approximate, or probable, or partial) truth of scientific theories that are empirically well-confirmed in a (significant) domain of application.

This general characterisation of scientific realism is still imprecise, since the various notions need to be unpacked, and a choice needs to be made (I will settle, in Section 2.3, for 'approximate truth'). It is also potentially vulnerable to the *pessimistic meta-induction argument,* which starts by noting that past scientific theories that were well-confirmed in a significant domain of application have turned out to be false. And since the past well-confirmed, now discarded, theories were false, *belief in the truth* of today's well-confirmed theories is unjustified.

Before I go on to describe the pessimistic meta-induction argument, let me motivate my general characterisation of scientific realism a bit more. Scientific realism is often characterised, as I did above, in terms of assertions and beliefs about the *epistemic status of scientific theories,* i.e. in terms of approximate truth (see e.g. Chakravartty (2017: Section 1.1)). However, it is sometimes alternatively characterised in terms of the epistemic *aims* of science: famously, by van Fraassen (1980: p. 8)—who, in addition, introduces a philosophical word, viz. 'scientific gnostic', for the person who believes the science she accepts to be true (1994: p. 182).

Of course, no one wants to call themselves a 'gnostic' (and the word, as defined by van Fraassen, has to do only with belief, and not with justification): but names and puns apart, there are several reasons to resist the replacement of talk about beliefs by talk about aims. Mine are as follows: First, 'aim' is too weak a term, and a definition of scientific realism in terms only of the aims of science, unless strengthened by some other condition, is liable to requiring the realist to be a realist about activities that have the same aims as science, but are not science. Second, the aims of science are themselves a subject of philosophical debate (see Section 1.1), and it does not seem desirable to make the definition of scientific realism directly dependent on the outcome of that debate. Third—and this is my main reason—scientific realism is *philosophical realism about something specific:* namely, scientific theories, and in particular the descriptions that they give of the world. To be a realist about objects, in the general philosophical sense, is to believe in the mind-independent *existence* of objects: it is not to have any specific beliefs about the *aims* of those objects (if objects have aims). And to be a realist about ideas is to say that those ideas *exist,* independently of the mind: it is not to have any beliefs about the *aims* of ideas (if ideas have aims). To be a realist about any subject-matter, likewise, is to believe that the part of the world, that the subject-matter describes, is as the subject-matter says it is, independently of the mind: to believe that the subject-matter's descriptions are *true.* This can be stated without reference to the *aims* of the subject-matter. Thus it would be a mistake to turn the definition of scientific *realism* into a definition of the *nature* of science. Scientific realism is an important foundational question in the philosophy of science, but it is not identical to the question 'What is science?'

In his vigorous attack on scientific realism, Laudan (1981) *severs the link between the 'success' of theories* (i.e. the degree to which theories are confirmed) *and their reference,* i.e. the entities that the terms of theories refer to. Thus he criticizes the idea that 'the world probably contains entities very like those postulated by our most successful theories' (p. 22), which one might have expected was true, if well-confirmed theories are true or approximately true.

Indeed, he famously lists an impressive record of theories that were once successful but whose central terms for unobservable entities we now regard as non-referring, including (p. 33): the crystalline spheres of ancient and medieval cosmology; the phlogiston theory of chemistry; the caloric theory of heat; the electromagnetic aether; etc.

Laudan's charge that the old theories 'fail to refer' or, alternatively, that the referrents of old theories and their successors are very different, goes back of course to Kuhn and Feyerabend's ideas about incommensurability. Feyerabend (1963: pp. 929-931, 939-941) emphasises that *meanings* change across theory boundaries, because the meaning of every term depends upon the theoretical environment in which it occurs.

Some of the responses to the pessimistic meta-induction argument include an *account of reference.* The rough idea here is that a well-confirmed theory, whose terms for unobservable entities refer to entities existing in the world, is also approximately true.

For example, in some accounts of reference, terms refer to particular objects in virtue of a *causal relationship* between the object and the speaker who introduced the term in the first place. Thus Kitcher (1993: Chapter 4) argues that the speaker's *intentions* must be considered in assigning reference: 'thus Priestley's tokens of 'dephlogisticated air' are fixed by his intention to refer to air with the substance emitted in combustion removed from it' (p. 149).

This causality principle is also used by *causal theories of reference,* where a term for an unobservable entity refers to whatever entities cause the phenomena that led theorists to introduce the term in the first place. So, 'when Franklin and Ampère used the term 'electricity', it referred to the *actual* electricity causing the phenomena (like sparks, lightning bolts, currents, and electromagnets) that led them to posit the existence of an unobservable physical magnitude responsible for these phenomena' (Stanford (2006: p. 148)). In the same way, the 'aether' of the early electromagnetic theories refers to the electromagnetic field itself, because it plays the same sorts of causal roles that the electromagnetic field is supposed to play in carrying electromagnetic energy.

Stanford seems to partly agree with some of these accounts of reference, but criticizes their use as arguments for scientific realism, on the grounds that they miss the point of the meta-induction argument. For if even these central terms for unobservable entities refer, the fact that they are embedded in theories that repeatedly turn out to be 'radically misguided', implies that it would be foolish to believe such theories—they cannot be approximately true, despite the fact that their central terms refer.

There is, in most of the literature on scientific realism of which I am aware, what I think is a major weakness in the analysis of meaning. The literature mostly appears to assume a *single kind of meaning,*[4] i.e. a single way in which words are meant when they are used by scientists.

The literature includes both convincing arguments that some disputed term does refer (for example, the 'aether' in the old electromagnetic theory), as well as equally reasonable arguments that the same term does *not* refer: see Laudan's (1981: p. 33) list. This suggests that the pessimistic meta-induction problem is largely linguistic, and that there is a need for a more nuanced conception of what we mean by 'reference'.

The distinction between *two kinds of meanings* of words when used by scientists, namely intensions and extensions, is standard in referential semantics and well-known in philosophy of science: but it seems nevertheless to have been largely neglected in this discussion:[5] while it is prominent in the early literature on *reduction*.[6] For example, Nagel (1979: pp. 914-915) used that distinction to defend his own account of reduction against Feyerabend.

The distinction between intensions and extensions will be the starting point of my analysis.[7] I will claim that the oft-repeated slogan 'the term X fails to refer' should be replaced by 'the term X has a different *intension* in the theories Y and Z'. But of course this still leaves the possibility that X might have the same *extension* in the theories Y and Z, thus referring after all!

The extensional scientific realism that I develop here is semantic and epistemic, but not (strongly) metaphysical. That is, I aim to argue that 'the realist is justified in believing in the entities that our best scientific theories postulate (alternatively, in believing their ontologies), where the meaning of the terms is construed extensionally'. But I have no

[4] Although Kitcher (1993) distinguishes how terms refer on different *occasions of utterance,* there is still a *single kind of meaning* on those occasions in which terms do refer. Likewise, on Psillos's (1999) account, there is a single kind of meaning for those parts of a theory that do refer.

[5] Kitcher (1993: pp. 75-76) does, in one place, mention the distinction between Frege's terms 'sense' and 'reference', but then seems to dismiss it, because Frege believes that 'sense determines reference' (i.e. two expressions with the same sense must share the same referent), while Frege's notions cannot satisfy this. A similar criticism is in Psillos (1989: p. 272). But this is not a problem for me, for two reasons. First, I do not use the Fregean contrast between sense and reference, but rather the Carnapian-Lewisian contrast between intension and extension, which is now standard in semantics. [See for example Heim and Kratzer (1988: Chapters 2 and 12) and Speaks (2019). For the comparison between the pairs sense-reference vs. intension-extension, see Carnap (1947: pp. 122-133)]. Second, the standard accounts of intensions and extensions do not have this problem, since it is *not* the case that the intension alone determines the extension (in particular, it is not true that two expressions with the same intension must share the same extension): context of utterance and circumstances of evaluation are often also required (Kaplan (1977: p. 494); which Frege of course never denied: see Frege (1956) [1918]).

[6] The philosopher's traditional account of reduction is Nagel (1961: pp. 351-363; 1979), Hempel (1966: §8), Schaffner (2012). On this account, reduction is, essentially, deduction of one theory from another, almost always using additional definitions or bridge-principles linking the two theories' vocabularies. Recent proposed rebuttals of the objections to the traditional account are: Dizadji-Bahmani et al. (2010: pp. 403-410) and Butterfield (2011a: §3).

[7] The differences between the two pairs of expressions, 'sense and reference' vs. 'intension and extension', have to do mainly with opaque contexts and with some of Frege's more specific doctrines (see Carnap (1947: pp. 122-133)). Apart from opaque contexts, the two pairs of expressions are used in the same way, and give the same result, as Carnap (1947: pp. 125-126) explicitly shows. Where they differ, it is the contrast 'intension vs. extension' that I endorse.

intention to specify what those entities are, from the outset, and independently of the specifics of the relevant scientific theories.

Thus this discussion requires two steps, which will be carried out in the next two Sections:

(1) Explain the continuity of the reference of terms for unobservable entities between theories (Section 2.2).

(2) Explicate approximate truth (Section 2.3).

2.2 Extensional Equivalence as Given by Correspondence

This Section addresses how the reference of the terms of scientific theories is fixed. The way reference is fixed for scientific theories is not essentially different from how this is done in ordinary language, although there are of course aspects that are highly specific to science. Thus reference is fixed through a specification of the intensions, circumstances of evaluation, and context of utterance (see Section 1.3.1).

More specifically, this Section will list various factors that are used in practice to determine the extensional equivalence—usually between terms—of two theories that are in a relation of correspondence. 'Correspondence' is a broad notion that involves many aspects formal, heuristic, conceptual, material, empirical, etc. The list does not aim to be exhaustive, but rather to make the notion of extensional equivalence, introduced in Section 1.3.1, more vivid and applicable by giving its main examples and characteristics.

2.2.1 On the notion of 'correspondence'

Bohr used the word 'correspondence' to denote various aspects of the relation between classical and quantum theories.[8] Correspondence, as the broad idea that the successor theory should build on, or should somehow reproduce, the theory that it replaces, of course goes back much earlier. It is surely implicit in Maxwell's kinetic theory of gases no less than in his unification of electric and magnetic phenomena, and in Newton's derivation of Kepler's and Galileo's laws as special cases of his own theory of gravitation.[9]

In the philosophical literature, the notion is present in discussions of reduction, and various weakenings, of it: Nagel's (1961, 1979) heterogeneous reduction admits that strict deduction of one theory to another fails, but deduction can nevertheless be achieved by postulating additional **bridge laws.** It is a *correspondence* account of reduction, where the bridge laws are correspondence rules as empirical hypotheses relating things mentioned in both the reduced and the reducing theories. These empirical hypotheses determine the

[8] For a treatment of Bohr's correspondence principle, see Rynasiewicz (2015); for an alternative treatment, see e.g. Bokulich (2008). See also van Dongen et al. (2020: Section 5.1). For the correspondence between special relativity and classical mechanics, see Brown (1993).

[9] Post (1971) begins with quotes from Whewell, Duhem, Born, Rohrlich, Maxwell, and Wigner that effectively support the idea of correspondence between theories.

extensions of the reduced theory's predicates mentioned in the correspondence rules.[10] (Regardless of Nagel's empiricism, my own view is that determining extensions is not only an empirical, but also a conceptual matter: as I argued in the previous Section, and am arguing again in more detail in this Section, and will illustrate in the example of Section 2.2.5.)

A further step is taken by Fodor (1974), who denies the complete reduction of the special sciences to physics, on the basis that bridge laws are not identity statements between types, but rather contingent identities between tokens: what he calls **token identity** (especially: of events, rather than properties).[11] I will use the idea of token identity here, divorced of its anti-reductionist connotations.

Token identity can be recognised through various empirical and conceptual procedures, and not only through the strict identity of the bits of matter stuff of e.g. a sample of a substance, i.e. an individual lump. For example, recognising that, in a specific case, mean kinetic energy and temperature are contingently identical is aided by judging the roles that temperature and kinetic energy play in the two theories, including their **causal roles** (the heating up of a gas at constant pressure *causes its expansion:* in one case because of the Boyle-Charles ideal gas law, and in the other because of the free motion of the particles. More on 'causal roles' below[12]).

Post (1971) construed 'correspondence' as a heuristic guiding principle in science. The (generalised) correspondence principle is 'the requirement that any acceptable new theory should account for its predecessor by 'degenerating' into that theory under those conditions under which [the predecessor theory] has been well-confirmed by tests' (p. 16).[13] Hartmann (2002) and Radder (1991) perceptively analyse and criticise the subsequent literature on the generalised correspondence principle (which it is not my aim to do here). Radder (1991), in his investigation of the relation between classical and quantum mechanics, distinguishes three main kinds of correspondence: numerical, formal, and conceptual correspondence (the last is also called 'conceptual continuity'). In **numerical correspon-**

[10] Nagel (1979: pp. 914-915) distinguishes two types of correspondence rules: those of the first type state conditions, often in terms of a micro-theory, for the occurrence of traits characterising various things, often macroscopic ones; such bridge laws are empirical hypotheses concerning the extensions of the *predicates* mentioned in the correspondence rules. The bridge laws of the second type relates the *individuals* or entities designated by different predicates. As I mentioned in the preamble of Chapter 2, this was a response to Feyerabend's (1963: pp. 929-931, 939-941) accusation that Nagel did not respect meaning variance: but Nagel (1961: p. 357) had already said that, for example, the meaning of 'temperature' in thermodynamics and in classical mechanics is different, but that nevertheless a postulate can be introduced that relates them ('this postulate cannot be warranted by simply explicating the meanings of the expressions contained in it'). In his 1979 reply, Nagel stated explicitly that 'temperature' has a different intension in the two theories, but the same extension.

[11] Sober (1999: pp. 552-554) argues that multiple realisability does not preclude reductionism, and criticises one aspect of Fodor's (1974) anti-reductionist argument: namely, his contention that laws cannot be disjunctive. Regardless of whether or not Sober's critique invalidates Fodor's argument: the aspect of Fodor (1974) that is interesting for scientific realism—namely, his consideration of token rather than type relations between theories—is immune to Sober's critique, and is close to the view of extensional continuity that I am defending.

[12] See also Lewis (1972: pp. 249-252).

[13] For an overview and critical discussion of both Post's work and a collection of essays in his honour, see Hartmann (2002).

dence, the two theories agree (perhaps after appropriate approximations are done) on the numerical values of a distinctive set of salient quantities. In **formal correspondence,** the (mathematical or technical) form of the laws of one theory can be obtained from those of the other (Radder (1991: p. 208)). Radder's own conclusion is that between classical and quantum mechanics there is numerical and formal correspondence, but no conceptual correspondence (in the sense of Radder's 'conceptual continuity'). My own view is that, although there is no conceptual *identity* (i.e. intensional equivalence of concepts) there is **conceptual correspondence** (i.e. extensional equivalence, and in addition similarity of intensions), as I will explain in the next Section. Although numerical, formal, and conceptual correspondence between theories on a given extension are a matter of equivalence of models, the *circumstantial physical conditions* listed in the previous Section usually play an important role in obtaining this correspondence.

Post's 'conditions under which' the new theory reproduces the predecessor theory are what I call the 'levels of abstraction' (see Section 1.3.2) at which the theory or model are treated, and which determines the circumstantial physical conditions of the domain. Thus, for example, the approximations made to get numerical correspondence between quantum mechanics and classical mechanics are part of the level of abstraction required to describe the domain of classical physics.

2.2.2 Conceptual correspondence in quantum mechanics

In this Section, I illustrate the notion of conceptual correspondence using the example of quantum mechanics.

One might wonder whether there is *conceptual* correspondence at all between quantum and classical mechanics. Indeed, one could easily take the conceptual *dis*continuity between classical and quantum mechanics to be obvious: namely, quantum mechanics introduces a number of concepts that are not identical with any classical concepts (for example, the quantum state, observables as operators, etc.).

But I think one can best say that there are conceptual *differences* between classical and quantum mechanics: this is indeed obvious and undeniable, given that their intensions are not the same, and given my acceptance of the partial incommensurability of the intensions. And yet I maintain that there is also a conceptual *correspondence* between the two theories, which gives *partial continuity.*

For example, the concept of 'position', in the quantum theory, corresponds to the concept of 'position', in the classical theory, in the following sense. They play similar roles in both theories: namely, they both give information about the spatial location of a system (without, in the quantum theory, defining the position with infinite precision). They play similar roles also in terms of the **relations** they bear with other quantities, for example with the canonically conjugated 'momentum', and of how they are measured.

Indeed, there is a long tradition of studies and theories along the lines of such relations. Two standard examples are: (1) The correspondence between classical and quantum mechanics, whereby classical Poisson backets go over in commutators[14] (hinted at at the

[14] The idea of replacing Poisson brackets by commutators goes back to Dirac (2001 [1964]: pp. 35-43, 77); see also Messiah (1969: pp. 318, 337). The idea of a 'quantisation map' was first proposed by

end of the previous paragraph). (2) Ehrenfest's theorem, whereby the expectation value of an operator satisfies a close analogue of the classical equation of motion, and reproduces the latter in the classical limit (for example, the expectation value of the position operator satisfies an analogue of Newton's equation).[15]

In both cases, the classical result is usually reproduced in the limit $\hbar \to 0$ and $N \to \infty$, where $\hbar$ is Planck's constant and N is, roughly, the size of the system, e.g. the number of particles or components.[16] Once the limit is taken, we indeed have formal identity of the quantum and classical results. There is also conceptual correspondence, because the expectation value of the position operator plays a similar role to the position of the particle in the classical theory: as we have seen, it obeys Newton's second law; and, after the limit is taken, the quantum and classical quantities can be *identified.*[17]

Under the consideration of **counterfactual situations** in other possible worlds, this similarity persists, for the class of circumstantial physical conditions under which classical mechanics is valid. In fact, without the existence of such a conceptual correspondence, as a *similarity or resemblance between distinct but related concepts,* it would seem to be impossible to identify quantum and classical mechanics as both dealing with the same world of phenomena. I believe that this is something that an insistence on incommensurability, without further semantic distinctions, misses.

Like I said above, conceptual correspondence is not synonymous with conceptual identity: conceptual identity is intensional, and therefore it is usually not available and not needed. To be scientific realists about extensions, we only need that there is a *similarity* between the old and the new concepts, which play the *same local roles.* Partly through the local roles of similar concepts, we are able to identify extensions.[18]

Heisenberg, who reinterpreted the mathematical symbols of the classical equations of motion in terms of matrices. It was then used by Schrödinger in understanding his own equation: for a lucid discussion, see Landsman (2007: pp. 427-428). For the mathematical implementation of Heisenberg's idea (of a systematic map between classical and quantum objects) by von Neumann, see (ibid, p. 430). For more modern and mathematically rigorous approaches to this problem, especially 'deformation quantisation', see Landsman (ibid, pp. 446-464).

[15] See Messiah (1969: pp. 316-319), Landsman (2007: pp. 475-476).

[16] The limit $\hbar \to 0$ should be interpreted as a limit of small $\hbar$ *compared to* an energy scale usually set by the quantum mechanical potential, the mass of the particle, etc. See Messiah (1969: p. 214). For a thorough discussion, see Landsman (2007: pp. 471-515). This is not to claim outright a solution to the measurement problem—but it is surely relevant to it. See, for example, the recent programme in Landsman (2013: pp. 378-383), who treats classical and quantum mechanics using a common C*-algebra formalism. He then shows convergence of judiciously chosen quantum states to classical states (and the corresponding expectation values of quantities converge to the classical values). Another approach to the emergence of quasi-classical behaviour is in terms of the dynamics of decoherence, see Joos and Zeh (1985), Joos et al. (2003) and Zurek (2002).

[17] Note that this correspondence goes in the opposite direction to the familiar correspondence that assigns quantum operators to classical quantities, because it attempts to give a rule for the quantum theory starting from the classical theory, rather than the other way around: see Rynasiewicz (2005). Messiah (1969: p. 70) is a *locus classicus.*

[18] Thus I here disagree with Radder (1991, 2012 [1984]), who says that there is no conceptual correspondence between quantum mechanics and classical mechanics, and even in general: 'In general the correspondence is not of a conceptual but rather of a formal and numerical nature' (p. 214). 'In surveying the whole episode in the history of 20th century physics we may conclude that, ultimately, the success of the correspondence principle appears not to rest upon a conceptual correspondence but rather

Historically, this similarity of concepts played an important heuristic role: for Bohr's numerical and formal correspondence, in so far as they were successful, would have never succeeded without the conceptual correspondence, which was crucial in guiding his ideas, and in establishing how classical and quantum mechanics were supposed to agree in specific situations.[19] (In Section 2.3.4, I will argue that the similarity between T's and T''s intensions means that T captures a significant part of T''s intension.)

The correspondence between quantum and classical mechanics usually requires, as I mentioned above, setting some quantity to a large value, compared to Planck's constant $\hbar$, i.e. doing an **approximation** or taking a **limit.** For the hydrogen atom, Bohr's correspondence principle entailed taking the limit of orbits with large quantum numbers (in units of $\hbar$); in that limit, the classical orbital frequency of the valence electron coincides numerically with the frequency of the emitted radiation, and the electron can be treated as a charged oscillating classical source. Again, there is no intensional equivalence, but there is an extensional equivalence of the quantum and the classical system, in the limit considered. That is, the electron of the Bohr model of the atom is, *pace* Kuhn and Feyerabend, *extensionally equivalent* with the electron of the classical theory of radiation: the physical systems, and the possible experimental situations, to which these models can be applied are the same, even though intensionally they are of course very different.

My argument for the extensional equivalence of the electron of the classical and of the quantum theory, for the hydrogen atom under specified conditions, is similar to Lewis' (1972: pp. 249-252) argument for the identification of what he calls 'occupants of the same causal roles', i.e. if x and y occupy the same causal role (in two different theories, viz. T and T', respectively), and provided the occupants are uniquely realised, then also $x = y$ (where the equality sign indicates the extensional not intensional equivalence: see Carnap (1947: p. 14)).[20]

For example, the electron fulfils equivalent *causal roles* in both cases, in that it produces the radiation, and it is ascribed **properties** that are conceptually distinct, but extensionally equivalent[21] (for example, the orbital frequency in one case, and the frequency of the radiation that it emits in the other).[22]

upon a combination of numerical and formal correspondence' (p. 208). But Radder appears to have changed his mind, for in the *Postcript* to *The Material Realization of Science,* written 28 years after the first edition (2012 [1984]: pp. 161-189), he admits the following inadequacy of his earlier treatment of numerical and formal correspondence: 'I should have been more careful myself in using the phrase of formal-mathematical *continuity*, by making explicit that this type of continuity implies intertheoretical correspondence but not necessarily preservation' (Radder (2012 [1984]: p. 182)). Thus in his *Postcript,* Radder suggests a uniform treatment of numerical, formal, and conceptual correspondence, which I also advocate.

[19] Bohr used a combination of conceptual, formal, and numerical reasonings, and drew an analogy between the hydrogen atom and the Kepler problem that was crucial for the correspondence between classical and quantum mechanics (see Rynasiewicz (2005)). See also Saunders (1993).

[20] Lewis' argument is formulated in the context of a discussion of reduction of mental states to neural states: but the argument does not, in itself, require reduction.

[21] For a list of "stable properties of the electron that are retained" by all theories after Thomson's discovery of the electron, see Bain and Norton (2001: pp. 453-455). I would like to restate Bain and Norton's claim by saying that the properties they list are intensionally different from each other, but extensionally equivalent in the sorts of experimental situations that they discuss.

[22] The definition of extensional equivalence of properties is as in Section 1.3.1. See also Carnap

The approximations or limits mentioned above are part of the level of abstraction, which determines the circumstantial physical conditions of the domain.

2.2.3 Material and predictive correspondence

In this Section, I discuss material correspondence and its relation with predictive correspondence. Material correspondence will be my main example of how the *context of application* (see Section 1.3.2), here given by the concrete experimental set-up, contributes to determining reference. I will do this using another example.

Consider the case of fullerenes, which are (possibly large) molecules consisting of carbon atoms connected by single and double bonds. They can form closed meshes, with rings of five to seven atoms. A famous example is the molecule buckminsterfullerene, or C_{60}, a type of fullerene with the colloquial name "buckyball", because its structure of twenty hexagons and twelve pentagons resembles a football. Using spectroscopy on a carbon gas obtained from e.g. a condensate of vaporised carbon, discrete infrared bands of fullerenes can be distinguished. But fullerenes can be extracted from soot in other ways, for example using organic solvents. Also, their (slow) reactions with other chemical compounds can be studied.[23]

Is there an extensional equivalence or *in*equivalence between (i) the fullerene of the spectroscopist, who thinks of the substance as a quantum molecule with a discrete spectrum, (ii) the fullerene of the organic chemist, who applies solvents to the soot in a large diesel truck's fumes to extract the fullerene molecules, and (iii) the fullerene of the chemist who is interested in how fullerene interacts with for example hydrogen or nitrous oxide gases? I claim that there is here an extensional *token identity,* of the type that Fodor (1974) discusses, that brings out the importance of **identity of bits of matter stuff,** e.g. a token sample of fullerene, in specific laboratory situations.[24] (There *may,* in addition, also be a type identity with a narrow type: but in this Section we are mainly interested in the role of token identity.) To see the token identity, notice that the fullerene molecules in the sample that are studied by the quantum spectroscopist are (can be) the *same* fullerene molecules in the sample studied by the chemist.[25] The identity of the tokens is easily established: simply bring a sample of fullerene, extracted by the organic chemists, to be analysed in the spectroscopy lab. (The claim here is that the extensional identity is open to explicit investigation using token identity: but it is of course not necessary

(1947: p. 14), who defines both extensional and intensional equivalence of properties (as special cases of what he calls 'designators', see Section 1.3.1) in terms of the truth of the corresponding extensional and intensional sentences.

[23] See Kroto et al. (1985, 1996).

[24] It should be clear that there is an important element of indexicality ('*this* is the sample of fullerene that we extracted from the soot') in establishing the identity of the bits of matter stuff. Hence the claim, from the previous Section, that—even though scientific theories aim to abstract from context—*context* can be important when it comes to the experimental practices that enter into establishing extensional equivalence.

[25] I am, in effect, saying something similar to Hacking (1983: p. 271), when he argues that people came to believe in the existence of electrons when electrons could be manipulated. For me, manipulation is but *one* of the factors that contribute to establishing reference. For a critical discussion of whether manipulability is sufficient to establish the reality of an entity, see Egg (2018: pp. 123-125).

to have token identity every time. A claim of equivalence also needs to be reproducible by different samples of the same type, and so only the weaker identity is needed.) Thus there is an identity of tokens, even though the intensions are different, i.e. the quantum mechanical descriptions and the classical (chemical, and material-science) descriptions are very different.[26]

But notice that it is not just the identity of the bits of matter stuff that is required for full extensional equivalence: for if the chemical and the quantum mechanical descriptions did ascribe different, indeed empirically conflicting, properties to the fullerene sample, we would not think that there is equivalence. Thus *predictive correspondence* is required, i.e. the two theories should agree where their domains overlap, to the accuracy with which the domain is given by the level of abstraction. The predictive correspondence entails, in this case, numerical equivalence to the given accuracy, as well as formal correspondence. In some cases, this may be difficult to check in practice, but the numerous experiments confirmed that the atomic structure of buckminsterfullerene and other fullerenes lead to the chemical properties that the theory predicts.

Consider Feyerabend's (1963: pp. 928-932) treatment of special relativistic versus classical mass. Feyerabend admits that, in the limit of small velocities compared to the speed of light, the two masses coincide *numerically.* But he claims that their meanings disagree, because the laws in which they figure are formally and conceptually different (for example, according to Feyerabend, mass is conserved in classical mechanics, but not in special relativity).[27] I agree with Feyerabend that the intensions of the property 'mass' are different: but once again, the extensions are the same, given the circumstantial physical condition of small velocities at this level of abstraction.[28]

This is because of a combination of material, formal, and conceptual correspondence.

[26] Notice that the argument also does not rely on Fodor's (1974) rejection of type identity. In the present case, *even if* one granted that all organic fullerene samples are also spectroscopic fullerene samples and vice-versa, the intensions would still differ, but there would still be equivalence of extension. Thus it is equivalence of extension, rather than mere token identity, that matters.

[27] In special relativity, the physical quantity of interest is the relativistic energy (related to what is sometimes called "relativistic mass" by a constant), and there is a conservation theorem for this quantity, which takes the form of the 'energy-work theorem', just as in classical mechanics. This conservation law reproduces the usual conservation law in the limit of small velocities. Thus there is not only a formal, but also a conceptual correspondence between special relativity and classical mechanics.

[28] Galison's (1997: p. 812) historical analysis of the measurement of the "relativistic mass" of the electron (or, rather, of what was then called the 'transverse mass'), and of Einstein, Lorentz, Poincaré, and Abraham's engagement with it, shows quite convincingly that all of these theorists are able to engage with the same experiment, communicate with each other about it, and even change their minds in the face of it. For example, Lorentz dramatically declared: 'my hypothesis [of explaining mass by] the flattening of electrons is in contradiction with Kaufmann's results, and I must abandon it. I am, therefore, at the end of my Latin' (p. 812). And Poincaré made similar concessions. Galison goes on: 'despite the "global" differences in the way "mass" classifies phenomena in the Lorentzian, Abrahamian, and Einsteinian theories, there remains a localized zone of activity in which a restricted set of actions and beliefs is deployed. In Kaufmann's and Bucherer's laboratories, in the arena of photographic plates, copper tubes, and electric fields, and in the capacity of hot wires to emit electrons, experimenters and theorists worked out an effective, though limited, coordination between beliefs and actions. What they worked out is, emphatically, not a protocol language—there is far too much theory woven into the joint experimental-theoretical action for that. Second, there is nothing universal in the establishment of jointly accepted procedures and arguments' (p. 813). See my discussion of 'theory-ladenness', at the end of Section 1.3.1.

First, the mass is measured using the **same procedure** in both theories: for example, the mass of the electron was measured historically through a combination of cathode tube and oil drop experiments (see e.g. Bain and Norton (2001: p. 453)). Thus a *single sequence or combination of experiments* suffices to determine the mass in both theories,[29] because of the material continuity between them, without which mere numerical agreement would be insufficient—which again highlights the importance of the context of application in establishing extensional equivalence.[30] In other words, without identity of the bits of matter stuff one would not know which numbers one should measure.

Second, in the limit in which the two theories give the same value for the mass, the mass stands in the *same relations* to other quantities in both theories. For example, in the case of the measurement of the charge of the electron: the electric field, the deflection of the trajectory of the electron, and the electric charge. This conceptual and formal correspondence secures that the material continuity relates the same quantities measured: since the theory enters in combining two experiments to "measure" the mass (see footnote 29). In other words, if the formal and conceptual equivalence of the two masses were false, then one should, despite the numerical agreement, have *denied* that a single sequence of measurements suffices to measure the quantities in the two theories. Thus *material, formal, and conceptual correspondence* work together to secure that mass in special relativity and in classical mechanics are extensionally equivalent.

2.2.4 Overview of the elements of correspondence

I have gone through a number of examples, highlighting features that are jointly surely sufficient to establish extensional equivalence, although they are neither jointly necessary nor independent—establishing extensional equivalence requires judgment.[31] Generally, extensional equivalence is a kind of interpretative correspondence, and so it involves both *theoretical* and *empirical* aspects, as I already argued in Section 1.3.1 (and will argue again in Section 2.2.5). A completely theory-free language is neither necessary nor desired, as we have seen: for the conceptual correspondences depend on the theory-ladenness of the domain of application, and are actually *used* in establishing formal and numerical corre-

[29] This episode also highlights the combination of theory and experiment involved in the mass "measurement", however this is not the point that I wish to stress here.

[30] This way of establishing the token identity of bits of matter stuff echoes Galison's (1997: p. 804) emphasis of the importance of laboratory practice and material culture: 'different global meanings can nonetheless come to (even very complex) coordination in specific contexts. Such a partial sharing of meanings became salient at many points in the history of laboratory practice and the material culture of physics'. See also Radder's (2012) [1984] idea of material realisation.

[31] That 'extensional equivalence requires judgment' (i.e. it does not seem possible, using this argument, to identify a minimal set of partial aspects of theory or experiment that are both necessary and jointly sufficient to establish extensional equivalence and hence continuity), echoes Galison's (1997: p. 799) well-known diagram for the periodization of science. He contrasts it with the positivist (p. 785) periodization (where, despite the theoretical discontinuity, there is continuity in observation and a strict accumulation of experimental results) and the 'anti-positivist' periodization, where large changes in theory also produce changes in observation. Galison's own diagram is similar of the view I am presenting here, in that in his account the continuity is sometimes secured by (a combination of) instrumentation, theory, and experiment: in my account, continuity is secured by a combination of the correspondences that I list in this Section. Otherwise, our accounts of course differ widely in their aims.

spondence. Also, extensional equivalence is *factual,* but not restricted to the *actual* world: it involves the consideration of *counterfactual situations* in other possible worlds.

I will now give an overview of the various elements that enter into assessing extensional equivalence, which we have discussed. As we have seen in our three main examples—the correspondence between quantum and classical mechanics, the case of fullerenes, and the extensional equivalence of mass between special relativity and classical mechanics—*predictive, conceptual, and material correspondence* are inter-dependent. All three of them are usually required to establish extensional equivalence.

In general, extensional equivalence requires *bridge laws,* which say how the terms in one theory are to be related to those in the other. In all the cases we have seen, there are *numerical and formal correspondences* between theories. *Numerical identity* holds in an approximation that is appropriate to the particular extension, and is part of the empirical adequacy of the theories. Underlying these identities is a *conceptual correspondence,* namely a conceptual similarity between the concepts of the theories, so that the concepts have the same local roles in the given extension (for some concepts, one may have a stronger conceptual *identity* on a given extension). For example, two entities can have different *properties,* but these properties have the same instantiations on the relevant extension. Also the *causal roles* of entities, and more generally their *relations* to other entities, can be equivalent on an extension. Finally, conceptual correspondence includes agreement in other *possible, counterfactual situations.*

A further important aspect of the empirical side of extensional equivalence is the contingent *token identity* between entities (be they objects, properties, etc.) on a given extension, in the sense that they pick out the same items. This involves the consideration of material realisation, experimental and observational set-ups and procedures, methods of measurement, etc.

Note that the identity of the extensions of the terms of different theories does not only hold for obviously observational terms; it also holds for theoretical terms, such as 'state' and 'aether', that make an empirical difference.

Since anyone claiming to have a solution to the pessimistic meta-induction argument must show that—at least some of—Laudan's 'evidently non-referring terms' *do* refer, I will take up this challenge with one more example: namely, the electromagnetic aether.

2.2.5 How "the electromagnetic aether" refers

In Section 2.1, I already mentioned the idea that the aether of the old electromagnetic theory might refer to the same entity as that of the new electromagnetic theory—namely, the electromagnetic field itself. In this Section, I develop this suggestion.

From a contemporary perspective, and depending on the properties that one ascribes to the aether, Maxwell's theory either does not require an aether, or it is inconsistent with it. If there *was* an aether that was a fixed background medium on which to propagate, then there would be motion relative to it, and so light would not propagate isotropically in space independently of the motion of the source. But such effects are not measured.[32]

Careful reading of Maxwell's text shows, however, that the reasons for which he maintained the existence of the aether were sensible. I will argue that the aether's role is fulfilled by other quantities in the modern theory: namely, the electromagnetic field.

More specifically, I wish to suggest that Maxwell's aether can be identified extensionally with the electromagnetic field itself, *together with an appropriate set of Galilean frames of reference,* where by 'an appropriate set of Galilean frames of reference' I mean a set of inertial frames of reference that move with low speed (much smaller than the speed of light) relative to earth and to each other.[33] The electric and magnetic fields are indeed invariant under Galilean transformations within reasonable accuracy for the Galilean transformations between these frames, so long as their speeds are much lower than the speed of light, and so it makes sense to identify the aether with the electromagnetic field itself, together with the set of Galilean frames of reference.[34]

This is not to say that the two concepts are the same: for they are not. It means that the two terms are *extensionally equivalent* (see Section 1.3.1). If true, this gives us an extensional equivalence between the new electromagnetic theory and Maxwell's old theory.

The notions 'aether' and 'electromagnetic field plus the theory's set of approximate/select Galilean frames' are of course distinct in principle. For one, if we consider frames moving at speeds close to the speed of light, then the electromagnetic field transforms non-trivially, and then its properties may be different from whatever properties one's aether model ascribed to the aether. This means that there are counterfactual situations where the two notions ('aether', and 'electromagnetic field plus set of Galilean frames') disagree: their intensions are clearly distinct, because one can think of possible worlds in which they differ. Nevertheless, the two concepts play the same roles in the situations that Maxwell was interested in (see the previous paragraph), and *there,* on that specific domain, no distinction can be made. Thus for those situations, the two intensions give the same extension. Just like the Morning star and the Evening star are intensionally clearly distinct but happen to coincide extensionally, so the aether and the electromagnetic field (for a choice of a Galilean frame) are intensionally distinct, but extensionally equivalent.

In the final paragraph of his *Treatise,* Maxwell argues that the transmission of energy in electromagnetic interactions seems to require a medium—some kind of medium or

[32] In the Lorentz-Poincaré aether theory, clocks and rods contract and expand relative to the aether, such that all the effects of the motion relative to the aether, to a given order in the approximation, are cancelled out. The resulting theory is formally equivalent to Einstein's special relativity, in the relevant approximation, but interpretatively distinct. So—although it may not be very well-motivated—it is possible to hold on to the hypothesis of the aether even from a contemporary perspective. For a pedagogic discussion, see Sklar (1974: pp. 249-251). For details about the 19th-century aether theories, see Schaffner (1972) and Whittaker (1951). See also Myrvold (2019: pp. 5-14).

[33] Maxwell considers Galilean transformations in his *Treatise* (1954: pp. 241-243), and shows that 'In all phenomena, therefore, relating to closed circuits and the currents in them, it is indifferent whether the axes to which we refer the system be at rest or in motion.'

[34] The important aspect here is the level of abstraction at which the fields can be taken to be invariant, i.e. a regime of speeds that are small, and an approximation scheme where they are taken to be negligible compared to the speed of light. Extrapolating to small but non-negligible speeds sets one outside of the domain of application where the result is valid.

substance:

> If something is transmitted from one particle to another at a distance, what is its condition after it has left the one particle and before it has reached the other? If this something is the potential energy of the two particles, as in Neumann's theory, how are we to conceive this energy as existing in a point of space, coinciding neither with the one particle nor with the other? In fact, whenever energy is transmitted from one body to another in time, there must be a medium or substance in which the energy exists after it leaves one body and before it reaches the other, for energy, as Torricelli remarked, 'is a quintessence of so subtile a nature that it cannot be contained in any vessel except the inmost substance of material things.' Hence *all these theories lead to the conception of a medium in which the propagation takes place,* and if we admit this medium as an hypothesis, I think it ought to occupy a prominent place in our investigations, and that we ought to endeavour to construct a mental representation of all the details of its action, and this has been my constant aim in this treatise (Maxwell (1954: p. 866), my italics).

This is an excellent argument for the need of a medium to transmit the electromagnetic energy between two bodies that are separated in space: the energy does not simply *disappear* from the sender to suddenly pop up at the receiver at a different point in space. Rather, it is (in the modern understanding) carried through space by the electromagnetic field: and indeed a conservation law relates the change of the energy of a system in time to the local energy flow through space.

Thus the medium that Maxwell is talking about is extensionally equivalent to the electromagnetic field of the modern theory, for a fixed reference frame which is determined by the level of abstraction:[35] because the electromagnetic field *has a value at each point in space,* it can carry energy between two bodies at different points in space, in the form of electromagnetic energy.

Also, notice Maxwell's caution about the exact properties of the aether:[36] he only says that 'there must be a medium or substance', but he does not commit to any of the extant models of the aether. Thus Maxwell's nearly functional definition of the aether in his *Treatise* is perfectly compatible with the aether being (extensionally) the electromagnetic field itself plus a set of Galilean frames.

Disambiguating meaning into intensions and extensions, as classical semantics teaches us to do, shows that Putnam's (1978: pp. 20-22) and Laudan's (1981: p. 25) verdict, namely 'the aether is a non-referring term', is only half the truth. It is a coarse verdict that gets the intensions right, but the extensions wrong.

What if quantum field theory, or a future unifying theory, tell us that the electromagnetic field is something entirely different—not a continuous medium, but for example something grainy? No matter, say I: since, on the level of abstraction of classical electrodynamics, Maxwell's description is still *extensionally* true. This explains, more generally, why the scientific realist does not need to know the future of science.

[35] Maxwell's argument that a medium must exist fixes the extension not the intension.

[36] This caution is also apparent from his rare use of the word 'aether' throughout his *Treatise*: he normally uses 'medium', except in the last, more speculative, chapter from which I just quoted. Relatedly, the way in which he talks about the 'medium', in other places of the *Treatise,* is similar to how one would nowadays talk about a field, under similar circumstantial physical conditions.

Hardin and Rosenberg (1982: pp. 611-614) have also argued that—in the theories of Fresnel, MacCullagh and Lorentz, in addition to Maxwell—the 'aether' refers to the electromagnetic field. On the other hand, they take Cauchy's 'aether' to not refer. For he postulated 'types of elastic solid which do not exist, and for [whose] assumed properties no dynamical justification is offered' (Whittaker (1951: p. 137)).

While this judgment—that 'aether', in the former set of theories, refers to a single entity[37]—indeed seems correct, our views differ in two important points, of which the first leads in to the second.

First, I have not argued, as Hardin and Rosenberg do, that 'the aether refers to the electromagnetic field': but, rather, 'the electromagnetic field, together with an appropriate set of Galilean frames' (where 'appropriate' is further specified, above, in terms of low speeds). This key addition specifies the level of abstraction.

Second, the justification that Hardin and Rosenberg give for why 'aether', in these theories, refers to the same entity, is very different from mine. For them, only causal roles seem relevant: there is no consideration of a level of abstraction or a similar notion, and what determines reference is left very implicit. But an abstract consideration of "causal roles", with no attention for the elements that determine reference in daily life and in science (see Section 1.3.2), i.e. circumstantial physical conditions and context of application (Sections 2.2.1-2.2.4), is insufficient.

Thus Laudan's (1984: p. 160) criticism of Hardin and Rosenberg on this point is on the mark. Their notion of reference seems liable to being *too tolerant.* By contrast, a notion of reference that takes into account the circumstantial physical conditions and the context of application is as tolerant as it should be: because the situations under which the terms refer are limited to those that are relevant to the theory's domain of application.

As Hardin and Rosenberg (1982: pp. 610) convincingly argue, some of the examples in Laudan's list are not serious candidates for realist consideration, because the corresponding theories were not sufficiently well-confirmed at the time when they were formulated. One is not justified in being a realist about just any hypothesis or speculative theory that scientists propose; empirical adequacy and other aspects of confirmation, in a given domain of application, also should be met.[38]

[37] Whittaker's taking the lack of 'dynamical justification' as the reason why Cauchy's 'aether' lacks reference fits with my own demand for good reasons before we can take the terms of a theory to refer. That is, belief in the reference of a theory's terms, like theory acceptance, requires epistemic justification.

[38] I do *not* here advocate the idea of 'mature science' with a 'take-off point' (Hardin and Rosenberg (1982: p. 609)), since these expressions suggest that, once a science is mature, it may be justified to believe all of the theories and hypotheses that it produces. And that would of course be naïve, since mistakes happen at every turn. Rather, my scientific realism is *cautious:* belief is justified only under the discussed epistemic conditions. Furthermore, this justification is, as in every kind of knowledge, not absolute, but a matter of degree. Some theories are better confirmed than others, and so belief in the entities that they postulate is better justified. Also, one should beware of possible cases of epistemic luck, which happen in science as in everyday life. However, we need not solve the general epistemological problems while we are doing philosophy of science.

2.2.6 A Pyrrhic victory?

In this Section, I address what seems to me is perhaps one of the most difficult criticisms of the recent scientific realist views, discussed in the preamble of this Chapter: namely, Stanford's (2006: pp. 144-180) charge that *selective confirmation versions of scientific realism* require either knowing facts about the beliefs and intentions of past scientists that we do not have access to, or require us to say that past scientists have repeatedly misidentified the parts, features, or aspects of their theories that genuinely contributed to their success (p. 168). At its worst, the charge is that some defenders of selective confirmation incur in a selective reading of the historical record (p. 174). That is, selective realism seems liable to the accusation of doing Whig history. Stanford calls this scientific realism's 'Pyrrhic victory'.

So the question for me is: is my extensional scientific realism vulnerable to the same accusation?

The answer is a clear 'No', as I will now substantiate. First, the common feature of selective confirmation theories that puts them potentially in conflict with the historical record is the fact that they distinguish between different 'parts, features, or aspects' of theories, and then ascribe different "weights" to these parts. For example, some parts may be more causally involved than others in producing the relevant phenomena; or some parts may incite stronger belief than others. Those parts that are identified as more important, significant, or sound are then, *ex hypothesi,* expected to recur as parts of *later theories*: thus the historical record should preserve them, and in this way continuity is secured. Although this commonsensical view about scientific theories may at first sight strike one as sensible, one can see how a philosophical theory based on it can easily drift into a historical claim about how theories succeed one another.

In addition to the *historical* claim, this form of realism seems to also require substantive *scientific* claims about the relative merits of different theory parts. Those claims are no doubt made partly through a careful study of the historical record, and partly using the vantage point of *current theory.* In which case, one is liable to Stanford's critique.

But extensional scientific realism is *not a form of selective confirmation.* For it does not distinguish between different theory "parts" of the kind just discussed. Rather, it is a *linguistic and philosophical theory* about how to interpret the statements of scientific theories. It treats the different sentences, individual expressions (terms) and predicators of a theory *uniformly,* i.e. it does not give more weight to one type of sentence than another, one term than another, or one example of a predicator than another. It also does not in general distinguish occasions of utterance, i.e. the terms are used homogeneously. This is why we never needed to decide which terms refer and which ones do not: for all we know (assuming empirically adequate theories and sound scientific arguments: cf. the previous Section), *all the terms that scientists claim refer, do refer.*

This is worth stressing: that whenever scientists have good reasons to introduce entities into a theory, and in so far as the resulting theory is empirically supported, extensional, though not intensional, belief in those entities is justified.

Some amount of hindsight if of course inevitable in *any* discussion of the pessimistic meta-induction argument, since the point at issue is precisely the truth of *past* theories. For example, delimiting, in precise terms, the domain where a theory is empirically adequate is something that can, partly, be done only with hindsight: i.e. by finding domains where the theory fails to be empirically adequate, or by contrast with later theories. Nevertheless, when Newton says that there is a force of gravity whose properties are such and such, he secures reference in a domain of application whose precise boundaries have not yet been precisely established. It is of course further experimental work, and partly also the contrast with later theories, which further determines the boundaries of the domain in which belief in this referrent is justified. But Newton's original posit already secures reference. Also, standards of empirical adequacy are subject to historical change.

But this is a benign form of hindsight, that is part of normal theory construction and does not require Whig history. For scientists normally will not make claims (other than as their expressed *expectations* or educated guesses) about the empirical adequacy of their theories *beyond* the domain where the theory is well-tested: namely, a *claim* that a theory is valid beyond the domain where it has been well-tested would be—to use a Popperian word—dogmatic, and lacking in argumentative power. Such claims are usually not made as part of the work of scientists.

Thus we do not need to change the meanings of the words of scientists, since they are limited to particular extensions. Rather, extensional scientific realism distinguishes two *kinds* of meaning, viz. one extensional and one intensional. It also explains why we are justified in believing in the reference of each term as regards its extension but not its intension. Thus the theory makes no *specific* historical or scientific claims about the pronouncements of scientists, over and above the ordinary analyses in history and philosophy of science.[39] Rather, it *adds* an extra philosophical layer to the interpretation of the statements made by scientists. It says *in which manner the statements ought to be interpreted and believed,* i.e. how they refer to the world: without changing those statements themselves. This is one of the things that philosophy is supposed to do, and regularly does.

Thus my proposal is analogous to what van Fraassen (1980) does when he tells us that both realists and empiricists should listen to what scientists say (i.e. both can agree about 'the picture of the world drawn by the theory' (1980: pp. 14, 43, 57)), but the realist and the empiricist have different degrees of belief in the entities that scientists postulate.

Likewise, extensional scientific realism tells us to *take literally what the scientist says, but to only believe it as an extensional not an intensional kind of meaning.*

2.3 Approximate Truth

The first part of the reply to the pessimistic meta-induction argument was to explain the referential continuity of theories, so as to remove the initial perplexity about some terms' purported 'lack of reference' that appears to ensue. In the previous Section, I did this through the notions of extensional equivalence and correspondence.

[39] The pessimistic meta-induction argument *itself,* as presented by Laudan (1980: p. 33), contains, in addition to a prediction about the future, a claim about the reference of certain terms in past theories. On the other hand, no one working on the pessimistic meta-induction argument can of course escape interpreting the historical record and the utterances of scientists at one point or another, and doing this partly with an eye on current theory. But this is part of normal philosophy and history of science.

The second part of the reply is to show how this continuity between theories is *progressive,* in the sense that later theories generally approach the truth more than previous, discarded, theories which they succeed. Thus, in Section 2.3.1, I will first discuss the truth about extensions and intensions of *sentences* and then, in Section 2.3.2, the extensional truth of *theories.* In Section 2.3.3, I will define approximate truth and then answer the pessimistic meta-induction argument. Finally, Section 2.3.4 addresses whether there can be a more ambitious, intensional, notion of truth of theories.

2.3.1 Truth about extensions and about intensions of sentences

In the previous Section, I pointed out that a preoccupation with meanings, i.e. intensions, seems to have led most philosophers working on scientific realism to neglect extensions. This coarse approach to reference introduces a bias towards an intensional notion of truth.

In this Section, I develop 'truth' as an *extensional* notion, in the straightforward sense of 'truth *about* extensions': and as opposed to truth about intensions. As such, the two applications of 'truth' are close to Carnap's (1947: pp. 10-12) notions of truth and L-truth (i.e. 'logical' truth), but with adaptations relevant to scientific theories.

The points in this Section are not new: but they have, so far as I know, not been made in discussions about scientific realism, in the philosophy of science.

Kuhn and Feyerabend were of course right about the (partial) incommensurability of intensions: classical and quantum mechanics, classical mechanics and special relativity, statistical mechanics and quantum mechanics, etc. are partly incommensurable theories in so far as they have very different, indeed incompatible, intensions. If the intensions of the sentences of quantum mechanics are true, then those of classical mechanics must be false.

However, they were wrong in taking this as a reason to disregard the question of the truth of theories *tout court.* For, once we recognise that meanings are ambiguous between extensions and intensions, we have to distinguish between contingent and necessary truths, i.e. truths that obtain on a set of (some but not all) possible worlds, and truths that obtain in all possible worlds. Thus I define the following notions for sentences:

Truth about extensions (E-truth): a sentence is E-true, relative to a possible world, iff its truth value, in that possible world allowed by the theory (with its specific circumstantial physical conditions), is 1.

Truth about intensions (I-truth): a sentence is I-true iff it is true in all possible worlds to which the theory can be applied (under all possible circumstantial physical conditions).[40]

Notice that I-truth of course entails E-truth, i.e. I-true means "merely" 'E-true for all extensions' (i.e. all worlds to which the theory applies, under all their possible circumstantial conditions).

[40] I-true sentences are in general not necessary, in the sense of 'true in all worlds including those to which the theory does *not* apply'. Namely, a sentence is I-false if there is at least one possible world where it is false.

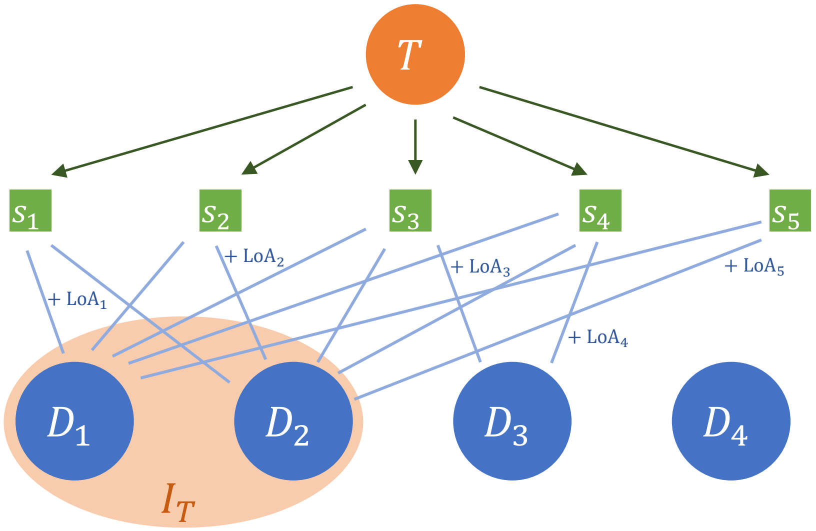


Figure 2.1: Sentences $s_1, \ldots, s_5$ implied by the theory T, in the domains $D_1, \ldots, D_4$. Each sentence, taken at a given level of abstraction, is true in some domain (blue lines). The theory's intension is $I_T = D_1 \cup D_2$, since all five sentences are true at both D_1 and D_2.

Here, I take the truth of sentences in the usual Tarskian (1935, 1944) sense of the word, as expressed by disquotation: to understand 'p is a true sentence', just disquote p.[41] Apart from the disquotation of sentences (his 'convention T', or 'equivalence T'), Tarki's theory of truth is of course also concerned with the recursive definition of the truth of sentences in terms of satisfaction relations for their components. See Tarski (1944: pp. 352-354) and Davidson (1967). In this sense, truth is a device for semantic ascent from the object language into the meta-language.[42]

Since, in order to carry out the truth valuation, a possible world and an appropriate level of abstraction are required (which determines the circumstantial physical conditions), E-truth is factual, and related to a specific physical situation, as already discussed in Section 1.3.2. Thus the resulting notion is the contingent truth of sentences that are true in an approximation scheme that is defined by the level of abstraction.[43] Thus *in stating the E-truth of a theory, it is essential to also state the level of abstraction.*

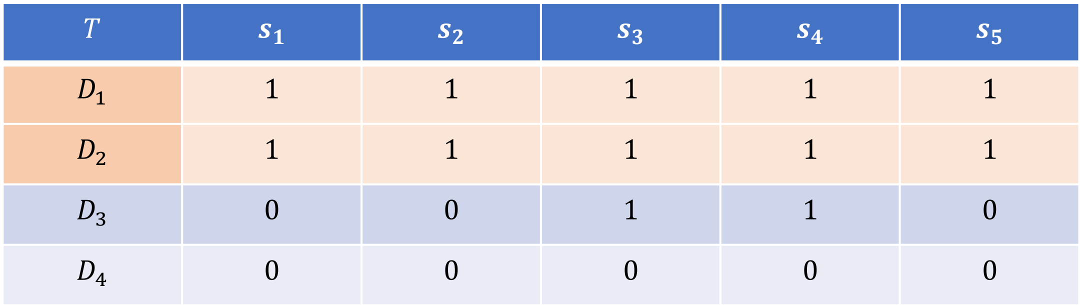
This of course means that a sentence can be I-false while being E-true, and it can agree with another sentence of the same theory that is both I- and E-true.

The above is illustrated in Figure 2.1, which shows a theory, T, with the five consequences that it implies, namely the sentences $s_1, \ldots, s_5$ (for example, these could be five laws, or statements of the values of five different quantities). $D_1, \ldots, D_4$ are four domains

[41] For a clear exposition, see for example Putnam (1978: pp. 9-17).

[42] For a defence of Tarski's referential semantics as "truth enough for the scientific realist", see Musgrave (1989: pp. 385-387) and Butterfield (1988: p. 294).

[43] Although my use of 'sentences' suggests a syntactic notion of a theory, intensions and extensions are here defined in a possible worlds semantics, and nothing in my account depends on a syntactic construal of a theory. (In fact, my account of a theory already combines elements from the syntactic and semantic traditions.) Lutz (2017: p. 347) has recently argued that the recent debate about syntactic vs. semantic construal of scientific theories does not capture any significant differences.



T	s_1	s_2	s_3	s_4	s_5
D_1	1	1	1	1	1
D_2	1	1	1	1	1
D_3	0	0	1	1	0
D_4	0	0	0	0	0

Figure 2.2: Truth table for the sentences $s_1, \ldots, s_5$ in the domains $D_1, \ldots, D_4$. The theory's intension is $D_1 \cup D_2$.

of application of the theory, i.e. four possible worlds, each taken together with its own circumstantial physical conditions (e.g. the application of the laws or calculation of the quantities for four possible states). To test the E-truth of the sentences, i.e. their truth over each of the domains, each sentence requires a level of abstraction. As is seen from the Figure, s_1 and s_2 are E-true in D_1 and D_2, etc. The truth table in Figure 2.2 contains the extensions of each of the sentences, evaluated in each of the domains.

E-truth invites a progressive and approximate notion of truth, as I will explain in the next two Sections.

2.3.2 The E-truth of theories

So much for the extensional and I-truth of *sentences*. In this Section, I extend this to *theories*.[44]

The truth of sentences allows us to analyse the truth of a theory, T, as the truth of all of its consequences or predictions. Namely, we see from the table in Figure 2.2 that all the sentences, $s_1, \ldots, s_5$, have extension 1 in the first two rows, D_1 and D_2, and 1 or 0 for D_3 and D_4. Thus we can identify the corresponding domains of application, D_1 and D_2, with T's intension: namely, the set of possible worlds and circumstantial physical conditions where all of its predictions are true, i.e. $I_T := D_1 \cup D_2$. This is the domain of application that is correctly described by the theory. Thus T is E-true in $D_1 \cup D_2$, and extensionally false elsewhere. T is also *trivially* I-true in $D_1 \cup D_2$, since the intension is the maximal set of possible worlds (each with its circumstantial physical conditions) where T is E-true. In Figure 2.1, T's intension is the shaded region that consists of D_1 and D_2. I will not use the word 'I-truth of T' for T's truth on its intension, since this is a trivial notion, but rather for T's truth in *every physically, i.e. nomologically, possible* world.

Let me give an example: let $D_1 \cup D_2$ be the region of validity of classical mechanics, restricted to large systems with low speeds and small velocities. And let D_1 be the *actual* world with these restrictions, and D_2 some other possible world. D_3 and D_4 are whatever other worlds are of interest—for example, let D_3 be the special relativistic world of high energies and small gravitational masses, and D_4 the world of cosmology, where

[44] Standard treatments of intensional semantics do not, so far as I know, give separate treatments of extensional the truth of scientific theories. I am here giving the natural one.

gravitational masses and acceleration are large.[45] In this example, classical mechanics is E-true in $D_1 \cup D_2$ but not I-true, in the sense of 'true in every physically possible world', since it does not describe well situations with large velocities or large gravitational masses, i.e. D_3 and D_4. More generally, we cannot hope that any of our current theories are true in every physically possible world.

As I noted at the beginning of Section 2.3.1, when people object to scientific realism on the grounds that "all scientific theories are false", they appear to have in mind something like a notion of I-truth, in the tradition of Kuhn and Feyerabend. They assume that, because truth can only have two truth values, theories either get everything right or they are false. But intensional semantics allows a more nuanced notion of truth: E-truth is a contingent and factual notion of truth that is relative to a given level of abstraction. Thus a theory's statements can be true for some level of abstraction, and false for some other. That we never confront the world as it is in itself need not imply that all our theories are wrong in every aspect. Thus there can be a limited domain where a theory is true, and this domain already carries with it a restriction to a specific physical situation.

2.3.3 Approximate truth and the answer to the pessimistic meta-induction

Scientific realism requires us to talk about the approximate truth of theories (cf. Section 2.1). Having defined E-truth in Section 2.3.2, and with an eye on answering the pessimistic meta-induction argument, this is now relatively straightforward to do, and is the main job of this Section.

I will here be interested in the following problem: How to compare the relative truth of theories, i.e. how to say that one theory 'approaches the truth more than another theory', or 'is closer to the truth than another theory'.[46]

This is not a question of purely philosophical interest. For it is reasonable to think that a scientist who is also a scientific realist estimates, based on good reasons, that quantum mechanics approaches the truth more than classical mechanics. And thus one would like to have a construal of the comparative notion of 'approximate truth' that looks something like what realist scientists use in their actual judgments.

This is also the main question that one needs to answer in order to reply to the pessimistic meta-induction argument. The main perplexity that allows the argument to get off the ground is that theories that, at one point in history seemed true, contain terms—usually concepts like the 'aether', 'caloric', etc.—that are not present in later, more successful, theories: and thus it is not clear how the sentences involving those terms can *ever* have been true at all—for the terms do not even seem to refer.

As I argued in Sections 2.1-2.2, such apparently non-referring terms as the 'aether' and 'caloric' *do* refer after all, because meaning can be a matter of extensions *or* intensions. This is reflected in the truth table in Figure 2.2.

[45] The example is simplified, since for example the domains of cosmology and the special relativistic world of high energies are partly overlapping. But this simplification is not important for the illustration.

[46] There is also a separate question, about how our current theories approach the truth, without further qualifications. We do not need to address this question, and I will not have much to say about it here.

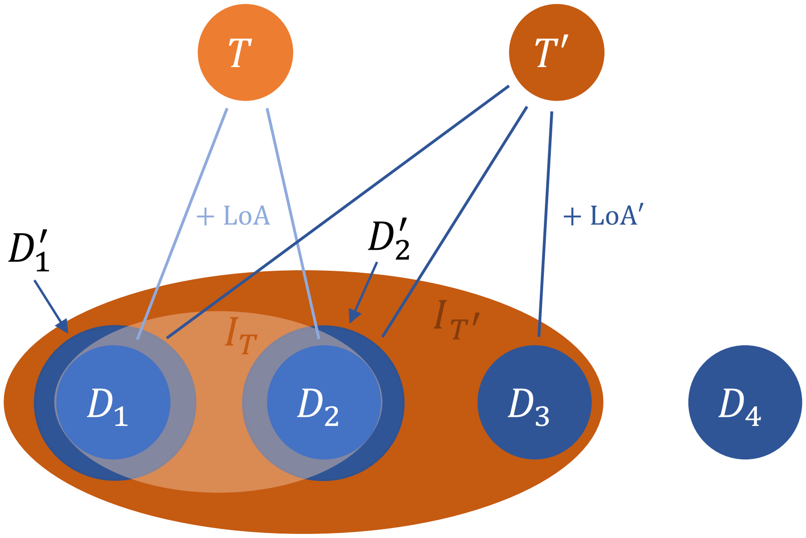


Figure 2.3: E-truths of the theories, T and T', over the various domains (light and dark blue lines). T's intension is $I_T = D_1 \cup D_2$, and T''s intension is $I_{T'} = D_1' \cup D_2' \cup D_3$. The domain in which T' is true is larger than that of T, and $D_1 \cup D_2 \subset D_1' \cup D_2'$.

Consider again, as an example to analyse the above question, Maxwell's electromagnetic theory involving the aether (see Section 2.2.5), and consider the sentence s_1 in the truth table in Figure 2.2 (for example, take this sentence to contain the word 'aether', and take D_1 and D_2 to have a fixed Galilean frame of reference specified as a contextual physical condition). The sentence is E-true in the domains D_1 and D_2, but not in the other domains. This extends, as we saw above, to the whole theory T, all of whose sentences are E-true in $D_1 \cup D_2$ only.

I will first argue that $I_T \subset I_{T'}$, as an example. Take T' to be the modern version of Maxwell's theory, which by the analysis of Section 2.2.5 (namely, that the aether is extensionally equivalent to the electromagnetic field and choice of Galilean frame) is E-true in the same domain as T, i.e. $D_1 \cup D_2$ (the corresponding domain, $D_1' \cup D_2'$, where T' is true, is strictly larger, and T's domain is contained in it, i.e. $D_1 \cup D_2 \subset D_1' \cup D_2'$: see Figure 2.3). But, in addition, T' is true in the domain D_3 where T is false: for example, take D_3 to be a domain in which the speed of light with respect to the aether can be measured, and shown to make a difference. Then T' is true there, while T is false. Since T is only E-true in D_1 and D_2, and T' is E-true in D_1', D_2', and D_3, we can say that T' approaches the truth more, or *is closer to the truth,* than T—i.e. it is true on a larger domain. Defining the intension of T', $I_{T'}$, as before, we see that the intension of T is contained in that of T', i.e. $I_T \subset I_{T'}$: T' has a larger intension than T, and so it is more widely applicable. (It is of course not necessary to assume that T''s domain strictly subsumes that of T: Kuhn losses are possible, and this leads in to emergence: see below, and Section 2.3.4).

We can now see why there is no contradiction in believing in the E-truth of, for example, *both* classical mechanics and quantum mechanics. Both theories are true at a given level of abstraction: for example, they are both true in $D_1 \cup D_2$, even though they disagree intensionally, i.e. outside of this domain. But because the extensions themselves contain circumstantial physical conditions, the truth of these theories is contextual and *contingent* truth, i.e. the truth values are all equal to one, in all the relevant worlds that satisfy the circumstantial physical conditions (which are a proper subset of all the physically possible worlds). Thus defining the extensions of scientific theories by levels of abstraction implies that *two intensionally incompatible theories can both be contingently and E-true on a given set of worlds, with the relevant circumstantial physical conditions.*

This then removes the main perplexity of the pessimistic meta-induction argument, for it answers the question: How can theories whose concepts mutually disagree be simultaneously true, and so how can the notion of 'truth' apply to theories at all? The answer is that two theories can both be E-true while only one (or none) of them is I-true: and that intensionally incompatible theories can be E-true on a common domain of application.

We have so far compared theories that satisfy a strong condition: namely, the domain of application of the new theory, T', subsumes the domain of application of the old theory, T. In such cases, there is a clear sense in which T' approaches the truth more than T.

But it is sometimes argued that there are cases that do not satisfy this strong condition: indeed, it is in principle possible that theory succession includes so-called 'Kuhn losses'. This happens if T describes some domain of application that T' does not describe: and so, while T' is more progressive than T in some aspects, it also loses some of T's descriptive power.

To address this possibility, we need a more general notion of the approximate truth of theories that can deal with these kinds of cases.

The idea is again to construe the progress of science in terms of E-truth, by comparing the domains of application where the old and the new theories are E-true.

But there is an important qualification of this project. For there are two possible answers that one can look for: namely, a qualitative or a quantitative answer. If one is looking for a quantitative or formal answer to the question, one is presumably interested in some kind of generalisation or improvement of Popper's notion of 'verisimilitude'. According to Popper, a theory is closer to the truth the more true consequences and the less false consequences it has. Indeed Popper's proposal, applied to E-truth as in this thesis, is to count the number of extensions of a theory with value 1 and subtract from it the number of extensions with value 0.[47]

The logic of Popper's proposal turns out to be problematic because the definition does not allow the comparison of false theories.[48] And, although the literature has proposed interesting solutions to this problem:[49] I do not think that an explication of Popper's proposal is needed to reply to the pessimistic meta-induction argument that working scientists are confronted with. For working scientists do not normally calculate truth measures when they judge rival theories: and, since I do not here aim to reform scientific practice, but rather to reply to a conceptual argument against this practice, I will leave

[47] If we compare the theories over the exact same domains, then we only need to count the domains where the theories are true.

[48] See Miller (1974) and Tichý (1974).

[49] See Kuipers (1982, 1987) and Schurz and Weingartner (1987).

such quantitative and formal projects aside.

This leaves a more qualitative approach as the interesting approach for our project. On this approach, one recognises that theory choice involves judgment, and that many qualitative arguments and considerations can be used that give reasonable verdicts, thus illuminating the history of science. Such qualitative arguments involve something like a quantitative reasoning: but more in the way of making a list of significant reasons for and against a general argument, i.e. without the pretence of the procedure being exact or comprehensive.

The qualitative project is the one that I think is most interesting, since it can handle real cases in the history of science and is close to how scientists actually reason. My approach is close to that of scientific realism's arch-enemy Laudan (1977), once we reinterpret his model of scientific rationality in terms of E-truth rather than mere problem-solving.

Laudan aims to develop a model of scientific rationality and progress based on the idea of problem-solving, which he sees as the main aim of science.[50] Namely, in appraising the merits of theories, the crucial question is whether they constitute adequate solutions to significant problems, i.e. substantive questions about the objects which constitute the domain of any given science. In other words, theory T' is more progressive than its predecessor, T, if it solves a larger amount of more significant problems: not only the *number* of problems counts, but also their *scientific importance.*

Laudan distinguishes three main types of empirical problems (unsolved problems, solved problems, and anomalous problems) and discusses, in qualitative terms, their relative importance. He draws attention to the importance of conceptual problems.[51] Indeed he considers these to be in general more serious than empirical anomalies (p. 64):

> The overall problem-solving effectiveness of a theory is determined by assessing the number and importance of the empirical problems which the theory solves and deducting therefrom the number and importance of the anomalies and conceptual problems which the theory generates... Progress can occur if and only if the succession of scientific theories in any domain shows an increasing degree of problem solving effectiveness... Any time we modify a theory or replace it by another theory, that change is progressive if and only if the later version is a more effective problem solver... than its predecessor. (p. 68)

Accordingly, thus reading Laudan's proposal for a qualitative assessment of the problem-solving abilities of a theory, along the realist lines that I am suggesting, is saying that: theory T' is closer to the truth than theory T iff it is E-true in a *larger, and more significant, domain.* The meaning of 'larger domain' is the one illustrated in Figure 2.3: namely, a larger number of worlds (thus the circumstantial physical conditions are weaker). And, by the 'significance' of the domain, I mean that some of these worlds and circumstantial

[50] Progressiveness need not be linear, as Laudan himself acknowledges: 'Progress can occur without an expansion of the domain of solved empirical problems, and is even conceivable when the domain of such problems contracts... theory change may conceivably be *regressive,* even when the index of solved empirical problem *increases,* specifically, if the change leads to more acute anomalies or conceptual problems' (p. 69)

[51] I wonder how Laudan's emphasis on conceptual problems, which he admits include not only logical, but also metaphysical and ethical problems, meshes with his overall anti-realist stance (perhaps a form of relativism or idealism).

physical conditions are more relevant to the theory, T, and the class of systems that it addresses, than some other worlds and circumstantial physical conditions (for example, a world can be relevant because it is very similar to the actual world).

Assessing the 'significance of the domain where the theory is true' is a matter of judgment not calculation, as it is in Laudan's case of the 'significance of the problems solved by the theory'; but it is a judgment that scientists make all the time.

For example, Newton's gravitational theory was able to describe the motions of the planets using the inverse-square law with more precision, and greater mathematical understanding, than ever before. But Newton could not point to a 'cause' for the gravitational force, as he himself acknowledges: while Aristotle could point to a teleological cause for why things fall—viz. the tendency of heavy objects to move towards their natural places, which was backed up by his overall metaphysics. (Thus there is a Kuhn loss, relative to the earlier Aristotelian physics of natural places.) Still, Newton thought that his theory of gravitation was a step forward with respect to the Aristotelian explanation: for, from the point of view of mechanics, having a quantitative description of the gravitational force is far more significant than—certaintly it is a first required step towards—a metaphysical understanding of it.[52] And I agree that Newton was correct in his judgment. Thus such judgments—namely, that the new domain of application of Newtonian theory is significant, and can compensate for the loss of part of the domain of application of the previous theory—*can* be made, and they *are* made by practicing scientists all the time. Extensional scientific realism only instructs us to interpret such judgments extensionally not intensionally.

Laudan (1977) himself admits the possibility of a realist interpretation of science, if only one can get a good notion of approximate truth:

> There is nothing in this model which rules out the possibility that, for all we know, scientific theories are true. Indeed, there is nothing I have said which would rule out a full-bodied, "realistic" interpretation of the scientific enterprise. But we do not have any way of knowing for sure (or even with confidence) that science is true... or that it is getting closer to the truth. (pp. 126-127).

We can now also see that Stanford's (2006: p. 148) criticism of scientific realist accounts of reference, mentioned in Section 2.1, does not apply to extensional scientific realism. The charge was that analyses of reference offer 'little comfort to the realist': for the referring terms are embedded in theories which regularly turn out to be 'radically misguided'.

This criticism does not apply because we have an analysis of *approximate truth,* in addition to *reference.* Scientific theories are, on an extensional scientific realist view, not misguided at all. First, there is a robust—extensional—sense in which scientific theories are true, and later theories are generally more true than previous ones: namely, E-truth. Second, there is in general no expectation, for the extensional scientific realist, that theories will turn out to be I-true—and thus no loss of comfort either, in light of pessimistic meta-induction.

[52] See Newton's *Opticks* (1721: p. 377) and his general scholium to the *Principia* (1999: p. 589). Of course, it was Galileo who had already polemicized that quantitative accuracy "matters more" than metaphysics.

2.3.4 I-truth, reduction, and emergence

We have so far considered the E-truth of theories, and argued that the realist is justified in believing that a theory is E-true on a given domain of application. This Section addresses the following question: Is the realist also justified in believing in the I-truth of a theory, i.e. the truth of the theory in all physically or nomologically possible worlds?

The scientific realist is, in general, clearly *not* so justified, at least not on the grounds of the present argument: since, as we have seen, the intensions of theories are in general different. Thus Kuhn and Feyerabend were right to point to the incommensurability of meanings—in so far as one qualifies incommensurability: as partial, and restricted to the realm of intension. The intensions of individual terms *are* sometimes incommensurable, and belief in the I-truth of two sentences containing a pair of such incommensurable terms is *not* justified if we require that intensions must be mutually compatible. Thus our scientific realism only commends belief in the contingent and extensional, not in the necessary truth of theories.[53]

Consider two theories, an older theory T and its successor, T', where the successor approaches the truth more than the older theory, in the way specified in the previous Section (i.e. $I_T \subset I_{T'}$). Let us also temporarily assume that T' subsumes T entirely, in the sense that, given various levels of abstraction, T' is extensionally equivalent with T on T's domain of application. Thus all of T's extensions agree with some of T''s extensions, but not vice-versa. Since *all* of T's extensions are given by T''s intensions (for specific circumstantial physical conditions), this means that T's intension is already captured by T''s intensions, if in addition the relevant levels of abstraction are provided. It is the case we already encountered, depicted on the left of Figure 2.4. In this case, although T' is closer to the truth than T, and so it is more widely applicable, we can say that T describes part of T''s intension—it captures a significant part of it.

The left diagram of Figure 2.4 includes cases of ontological reduction of the domain of T to that of T'. However, one should be careful in drawing from here a conclusion about the reduction between the *theories* T and T', because the domains of application, as construed in this thesis, contain circumstantial physical conditions. (The circumstantial physical conditions required for different domains can be incompatible with each other. And so, even though there is a local reduction on all of the extensions, it is not at all clear that the reduction is global, i.e. that there is a cousin of T' that is a well-defined *scientific theory,* and to which T reduces. See also the comments in footnote 32 of Chapter 1). Also, I believe that the meaning and consequences of 'reduction' of various types, in the philosophical literature, have—despite commendable attempts: see footnote 6—not yet been established sufficiently convincingly.

A more frequent scenario is the case in which T and T' have common subsets of extensions that agree (let me dub this subset $C := I_T \cap I_{T'}$), but they disagree on many others. In the right diagram of Figure 2.4, C consists of a single domain. If C is sufficiently significant (for example, if it comprises all our daily-life mechanics, including sending rockets to the moon), then the two theories' extensions overlap significantly. (And this

[53] The failure of most scientific theories to be I-true of course does not stand in the way of the fact that metaphysics and philosophy of science can and should enrich each other, as envisaged in French (2018: pp. 401-404) and Ladyman and Ross (2007: pp. 20-24).

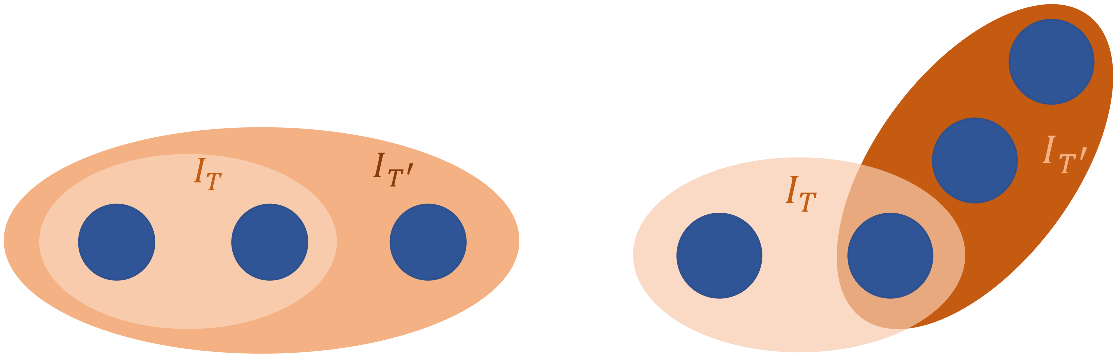


Figure 2.4: *Left*: T''s domain of application strictly subsumes that of T. This includes cases of reduction of the domains. *Right*: T' has a larger domain of application than T, but it does not subsume it. This includes cases of ontological emergence.

is just enough in many cases to ensure the continuity of science, as discussed in Section 1.3.2). This means that they capture significant parts of each others' intensions. Thus their intensions bear some resemblance to one another, since they imply a large number of identical empirical situations, say for the entire macroscopic universe, even though they disagree elsewhere.

This scenario includes cases of ontological emergence.[54] Imagine that the actual world is in the overlap between T and T''s intensions. Then we can regard T' as the more accurate, "microscopic" theory, and T as the more coarse-grained, "macroscopic" theory (viz. T' is true in three domains, while T is only true in two: and we take this third domain to describe the microscopic world). However, despite the fact that T is more coarse-grained, it has one domain that is not described by T', and so the concepts that T uses to describe the actual world will in general be different than T', since their intensions are different. This agrees with the notion of ontological emergence in Chapter 3: there is partial agreement of extensions (i.e. they agree in the macroscopic world), but difference of intensions.

A diagram such as the one on the right in Figure 2.4 should not be confused with a case of empirical under-determination. The theories T and T' overlap on a subdomain, but not an *empirical* subdomain. The subdomain in question contains both empirical and theoretical aspects; nor are the domains where the theories do not overlap to be seen as 'theoretical', since they are empirical as well. Thus the thesis that 'intensionally incompatible theories can be extensionally equivalent' has nothing to do with the empiricist idea that 'empirically equivalent theories are really the same theory and not incompatible after all' (Musgrave (1985: p. 200)).

[54] I say 'includes cases of', rather than 'is a case of', because there are additional conditions for ontological emergence: see Section 3.1.

2.4 Comparing with Structural Realism

Let me briefly spell out how extensional scientific realism differs from structural realism, and respond—albeit briefly—to some of structural realism's criticism of scientific realism.[55]

Worrall (1989) characterised structural realism as 'the best of both worlds', i.e. it is a position between empiricism and ordinary scientific realism. Continuity is seen as *more than* preservation of just the empirical content, and as *less than* preservation of the full theoretical content: namely, the structure is preserved. According to Worrall (1989: p. 117), Poincaré's structural realist position is 'the only hopeful way of *both* underwriting the 'no miracles' argument *and* accepting an accurate account of the extent of theory change'.

I will argue that structural realism is *not* the only hopeful such intermediate position: namely, extensional scientific realism is also such a position. For it agrees (with a qualification) with structural realism that scientific theories tell us about the structure of the world, but it states that scientific theories tell us rather *more* than just about structure. As I argued in detail in Section 2.2, the continuity between theories, seen as correspondence, is not only formal: indeed, it usually also involves *conceptual* and *material* aspects. Where extensional scientific realism differs from ordinary scientific realism is of course in its being *limited* to extensions, which are partly determined by the level of abstraction adopted.

Notice that a focus on structure is, by itself, not enough to secure continuity (and this is what I referred to above as a 'qualification'). Continuity is *only* secured because very different structures can *approximate* each other. Worrall notices that the same does not apply to entities: 'the notion of one theoretical entity [sic] approximating another... is extremely vague... the realist position surely becomes empty.' (Worrall (1989: p. 116)).

But this criticism applies only to a (naïve form of) ontological scientific realism. Notice that my extensional scientific realism, as a *semantic* form of realism, does not involve *entities* approximating each other. Rather, *concepts* are similar to each other on a given extension, there is continuity between *material procedures,* and the approximate truth of *theories* can be comparatively (and qualitatively) stated (see Sections 2.2 and 2.3). Thus the approximation is determined by the level of abstraction, and it is mirrored by the description of the circumstantial physical conditions (see footnote 31 in Chapter 1): we have here an approximation of *theoretical descriptions,* rather than a direct approximation of the entities themselves.

Thus I have argued that, because of the restriction to a given level of abstraction, the theories give successful conceptual, formal, and material descriptions of the world, and so are—contra Laudan—referentially successful. And yet such reference is not a trivial matter, since it is restricted to the circumstantial physical conditions and context of application given by the extension where the theory applies. And as I argued in Section 2.2.6, there is no need (in general) to interpret the words of scientists in ways that contradict their own intentions, or to distinguish different occasions of utterance,

[55] For discussions of structural realism, see for example Ladyman (2014) and Ladyman and Ross (2007). For a critical survey of the debates over structural realism, see Frigg and Votsis (2011).

i.e. reference is homogeneous. This is because *all* assertions have both extensional and intensional aspects to their meaning.

Extensional scientific realism thus *begins* with the consideration of the level of abstraction (with the circumstantial physical conditions that it determines) and, where relevant, a context of application. And since these notions apply in all of semantics—the semantics of scientific theories being no exception—there is no reason why the realist should stop at *structure*: something that, by itself, is difficult to define without the mention of concepts and material procedures.

In discussions of structural realism, 'structure' and 'intrinsic nature' are often contrasted. For example, epistemic structural realism is the position that scientific theories tell us only about the structure of the (unobservable) world, and not about its nature. It should by now be clear that extensional scientific realism, as here presented, is rather silent about the 'nature' of entities, if this phrase is meant to include the intensions of the corresponding terms (for a discussion of this point, see Frigg and Votsis (2011: p. 258)).

But the above is a false dichotomy (and indeed, such discussions are often conducted in an overly abstract way, with no appeal to the notion of levels of abstraction, nor to any kindred notion). For remaining agnostic about the nature of entities, in the above strong sense, still allows us to be realists about more than just structure. Namely, we can *consider the nature of an entity from an extensional perspective,* and in that sense be realists about it.

Ladyman and Ross (2007: pp. 85-88) criticise what they call the 'flight to reference', on two points:

(I) Causal theories of reference threaten to make reference 'a trivial matter, since as long as some phenomena prompt the introduction of a term, it will automatically successfully refer to whatever is the relevant cause' (p. 86).

(II) Restricting realism to only those parts of theories that play an essential role in the derivation of the predictions is ad-hoc and depends on hindsight.

About (I): first, 'causal roles' are only one aspect of my proposal (see Section 2.2.1 and 2.2.5).[56] And second, as I discussed in Section 2.2.5, reference is not automatically successful whenever the phenomena that prompt the introduction of a term are there: since reference depends not just on particular *phenomena,* but also requires correct circumstantial physical conditions and context of application to succeed. Belief in the reference of the terms is only justified if the phenomena in question persist over the entire domain of application.

About (II): I replied to this in Section 2.2.6—my proposal does not depend on theory

[56] Psillos (2012: pp. 220-222) convincingly argues that causal descriptivist theories solve this problem. In Section 2.2.4, I noticed that, although the elements of correspondence are not individually necessary, they are jointly surely sufficient to establish extensional equivalence (and reference). This makes my position a weak causal descriptivism, since it combines elements from the descriptivist and causal theories of reference. However, it is *weak* because it describes the conditions for reference as jointly sufficient but *not* as individually necessary. Thus it differs from Psillos (2012: p. 222) in that, in his account, causal relations *and* descriptions are both indispensable. And it differs from Lewis (1984: p. 226-228) in that (although he allows, as I also do, for a moderate amount of indeterminacy of reference and case-by-case specificity) he thinks that 'the descriptions are *largely* couched in causal terms' (p. 226, my emphasis).

parts, but rather on distinguishing extensions and intensions at a given level of abstraction.

Ladyman and Ross (2007: p. 92) give a third important argument why one ought to be structural realists:

> the success of our best current theories does not mean they have got the nature of the world right... Our solution to this problem is to give up the attempt to learn about the nature of unobservable entities from science. The metaphysical import of successful scientific theories consists in their giving correct descriptions of the structures of the world.

While I agree with the gist of the stated problem, it is not necessary to solve it in the way that they suggest. For, as I mentioned above, the 'nature of the world' contains extensional and intensional aspects. And as I have discussed, one is justified in committing to the extensions but not (by this argument) to the intensions. Thus my proposal is to replace, in the last sentence, 'structures' by 'extensions', which gives a different solution to the problem: a more cautious solution, because closer to scientific realism as usually conceived.

2.5 Scientific Realism: Conclusion

This Chapter has defended a reply to the pessimistic meta-induction argument against scientific realism. I have argued (in Section 2.1) that the analyses of realists and anti-realists alike have often been too coarse in that both consider a single kind of meaning. Likewise, they often see truth applied to scientific theories in just one way.

The reply to the pessimistic meta-induction argument consists of two main elements. The first is introducing into the debate (in Section 1.3) a standard set of notions from the philosophy of language and logic: namely, some elements of intensional semantics. Analysing the distinction between intensions and extensions of sentences and individual expressions, I have argued that one can have knowledge of the extensions—i.e. true and justified belief in what the linguistic terms express—but not (by this argument) of the intension of a theory. As in classical semantics, establishing reference to the extensions of expressions requires an analysis of their intensions, circumstantial physical conditions, and (often also) a context of application. The second element of the reply (Section 2.3) is a notion of approximate truth of theories: as contingent, E-truth.

Unlike earlier accounts, my explication of reference does not require anachronistic analyses of historical episodes (see Section 2.2.6), because it does not distinguish between different parts or elements of a theory, i.e. it is homogeneous throughout a theory, but rather adds to a scientific theory an extra philosophical dimension that is not normally considered in these debates. For, as any such account must do, extensional scientific realism is not primarily concerned with actual scientists' beliefs or intentions on various occasions, nor does it aim to reform what they say: rather, what scientists say is to be taken literally. Thus extensional scientific realism is a recipe for a specific *philosophical* interpretation of what scientists say, and a justification of the corresponding beliefs.

On an extensional realist position, the question of belief in the truth of intensions remains open. However, I have also indicated that there are similarities and extensional equivalences between intensions: so that, in theory succession, theories with smaller intensions typically capture significant parts of the ones with larger intensions that succeed them.

Extensional scientific realism recognises the important roles of both the theoretical and the empirical aspects of science, including its material aspects and its practices.

The view is also compatible with both reduction (various forms thereof) and emergence, and it does not require taking a position in the reductionism vs. anti-reductionism debate. However, the notions here developed surely help analyse these two notions, as we will see in the next Chapter.

Chapter 3

Towards a Theory of Emergence

The aim of this Chapter is to introduce and illustrate a criterion for ontological emergence. The framework is formal, where by 'formal' I just mean 'admitting of the basic notions of sets and maps'. The framework will then be illustrated by the example of the emergence of masslessness in classical relativistic mechanics.

The Chapter can be seen as a straightforward explication of the phrase 'ontological emergence'. Although my explication is related to other construals of this notion, in particular Humphreys (2016: pp. 56-93), I believe that my construal contains a number of novel aspects, and that the way I formalise it—as a single non-meshing condition between two maps—is a helpful tool for analysing cases of emergence, and for further conceptual analysis. The non-meshing condition can be understood as a difference in the intensions of two theories related by linkage (and sometimes also the extensions differ).

My immediate aim here is not strongly metaphysical, in the sense of requiring a commitment to a specific metaphysics of the world, and explicating emergence in those terms. Rather, my aim is to clarify what we mean by the phrase 'ontological emergence' (often contrasted with 'epistemic emergence') *in general*: and to give a criterion that is as straightforward as possible—a sufficient condition—for when it occurs. Thus I aim to give a minimal account of the meaning of 'ontological emergence', independently of whether we are e.g. Humeans or Aristotelians about causation—further metaphysical details then just adding to the basic picture that I will present here.

This and the remaining Chapters use the results from Chapter 1, including classical semantics, but do not presuppose notions of correspondence or approximate truth: and so, except when explicitly mentioned, they are independent of the extensional scientific realism from Chapter 2. For, as I will discuss in Section 3.2.2, the realist and e.g. the constructive empiricist can both discuss and agree about cases of ontological emergence.

Thus I will here construe 'ontology' in the straightforward sense of 'the ontology of a scientific theory', i.e. the domain of application that a theory describes, under a given interpretation. This domain of application is a part of the empirical world. Thus ontology is here not understood as a piece of language, but as a part of the world (more about this in Section 3.2).

I will also follow the new wave of emergentism in the physical sciences, which goes back to Anderson's (1972) 'More is Different'. The new emergentists have focussed on the emergence of entities and of properties[1] rather than on for example causal powers or causal properties, which are notions that require further metaphysical explication. While questions of causation are important in e.g. philosophy of mind, I follow the new emergentists in thinking that a minimal account of emergence can avoid them (see e.g. Bedau (1997: pp. 376-377), Hendry (2010: pp. 184-185)). We will see that the first steps taken by the present framework already present a number of questions and themes worth clarifying in their own right.

The writings of the new emergentists in physics are unfortunately not precise enough for us to extract from them a doctrine about emergence. The debate between emergentists and reductionists seems to have been fuelled by the alleged incompatibility between reduction and emergence.[2] Thus part of my aim is to offer a framework for emergence that allows for the coexistence of emergence and (at least one widespread type of) reduction, the possibility of which has been cogently defended for example by Butterfield (2011, 2011a).

An influential explication of emergence is by the British emergentist C. D. Broad (1925, p. 61):

> Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents, A, B and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B and C in relations of the same kind as R have certain characteristic properties; that A, B and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole $R(A; B; C)$ cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B and C in isolation or in other wholes which are not of the form $R(A; B; C)$.

While Broad's construal of 'emergence' has been influential, I also submit that his description, and others that define ontological emergence as the 'lack of deducibility', are too strong:[3] Broad (1925: p. 59) himself acknowledged that his theory of emergence fell short of finding interesting empirical illustrations: 'I cannot give a conclusive example of it, since it is a matter of controversy whether it actually applies to anything'. But we do not need to define emergence as lack of deducibility: what we need, I will argue, is to make a distinction between the *formalism* of a theory (which is the part of the theory that best allows us to use notions such as deduction) and the theory's *interpretation*, which is about the world, and need not be subject to such logical relations. This will allow a more general definition of emergence as novelty.

Another difficulty with the notion of emergence is that it is used broadly, and has various connotations. Guay and Sartenaer (2016) have recently carried out an interesting exercise in distinguishing three directions in the emergence landscape, through the contrasts: epistemological vs. ontological, weak vs. strong (i.e. emergence 'in practice' vs. 'in

[1] See for example Laughlin and Pines (2000: p. 28), Anderson (1989: p. 586).

[2] The main opponent of Anderson's emergentist doctrine was Steven Weinberg (Bedau and Humphreys (2008: pp. 345-357)). For a philosophical discussion of the debates, see Mainwood (2006: Section 2).

[3] For a critical review of the old notion of 'emergence as a failure of reduction' (where reduction is standardly defined as deduction, using "bridge-laws", *à la* Nagel), see Bedau and Humphreys (2008: pp. 10-11), Humphreys (2016: p. 37).

principle'), and synchronic vs. diachronic emergence. In this thesis, I will concentrate on ontological, synchronic emergence.[4]

Let me distinguish two main meanings of the word 'formal' that I will use. The official meaning of 'formal', as I announced at the beginning of this Introduction, is as in 'mathematics applied to philosophy': more specifically, in the sense of applying the notions of sets and maps to physical theories to articulate how they denote items in the world and theory-world relations. Thus my explication of 'emergence' is *formal* in that it applies to theories that are so formalised. What is 'formal' in this main meaning can still be interpreted, i.e. 'formal' here does *not* contrast with interpretation and ontology. The second meaning of 'formal' *does* contrast with 'interpretation': for it denotes the formalism, or the mathematical formulation, of physical theories, stripped of their physical interpretations. Since my main meaning of 'formal' is the former, I will use the phrases 'formalism of a theory' or 'formal, i.e. not interpretative' when referring to the latter.

The Chapter proceeds as follows. Section 3.1 lays out the framework for emergence: including the conception of epistemic and ontological emergence. Section 3.2 discusses three further questions that the framework prompts and compares with the literature. Section 3.3 illustrates the framework in a simple case study.

3.1 Towards a Theory of Emergence

This Section develops the main framework of this Chapter, including an explication of ontological vs. epistemic emergence that is adequate for formal sciences.

3.1.1 Correspondence and formal linkage between theories

In this Section, I give my preferred conceptions of theory, interpretation, and emergence. Talking about emergence in science of course forces us to talk about theories. And so, I will adopt the conception of theory developed in Chapter 1: in particular, the distinction between a bare theory T (regarded as a triple, $T := \langle \mathcal{S}, \mathcal{Q}, \mathcal{D} \rangle$) from Section 1.2.2, and an interpretation as the corresponding triple of partial maps, i, to a suitable domain D, from Section 1.2.3.

The general conception of interpretation from Chapter 1 is logically weak, because little is required for a structure-preserving partial map. But to discuss ontological emergence, we need interpretations that are "sufficiently good" within their domain of application. Thus, in addition to the interpretation being empirically adequate, I will impose the additional condition that every element in the codomain is described by at least some element of the theory. Though this looks like a strong condition, it is in fact innocuous. The idea is captured by the requirement that the map be *surjective*: and, as you might expect, this can always be achieved by restricting the codomain. We will return to this notion of interpretation, and develop it, in Section 3.1.3.

Discussions of emergence often work with *models,* i.e. specific solutions of the dynamical equations of the theory that are physically permitted, rather than with entire

[4] For a brief discussion of diachronic emergence, see footnote 21.

theories.[5] (The possibility of seeing emergence this way will resurface in the example of Section 3.3.) In such cases, the interpretation maps are assigned to the states and quantities of a particular solution i.e. model, rather than to the whole theory: and the domain is then naturally embedded in a single possible world.

I will also further refine the notion of interpretation through the distinction between the *intensions* and *extensions* of terms, discussed in Section 1.3.

I will construe the *domain of application, D*, introduced by the interpretation, as a set of *entities,* namely the elements of a set (fluids, particles, molecules, fields, charge properties, etc.), and relations between them (distances and correlation lengths, potentials and interaction strengths, etc.). Thus *the criterion of identity of domains is the set-theoretic criterion:* namely, the identity of the elements (and their relations).

To explain a bit more how this criterion will be applied in practice, consider that we normally use a language to talk about the world (cf. Section 1.3.1): we describe the domain of application linguistically (e.g. the theory may describe 'this red ball I am kicking'), and express identity in terms of this language. This language is an aid (e.g. words used to mention people) to describe on paper, or in sound, the intended physical things and properties, and it is in general different from the mathematical language of the bare theory. It is 'the language of experimental physicists (and lay people)' used to describe the world, and it may contain extra-linguistic elements such as diagrams, images, ostension, etc. To give some examples of how this language picks out the objects in the world: (1) 'A physicist's description of a laser beam' $\mapsto$ laser beam. (2) 'A set of numbers on a computer screen' $\mapsto$ the events in a particle detector. (3) The descriptions of experimental results (including the description of the working of the instruments) that we find in the pages of scientific journals.

Kuhn and Feyerabend said that the meanings of the terms of scientific theories cannot be compared, because theories are incommensurable. This cannot even be done empirically, because experimentation and observation are 'theory laden'. Thus the Kuhn-Feyerabend critiques of meaning might lead one to think that the criterion of individuation of entities in the domain of application is problematic, if the language that we use to talk about them is theory-laden. But we already saw, in Section 1.3.1, that—even embracing theory ladenness—the incommensurability thesis can be taken to apply (partially) to intensions, but not extensions. For various types of correspondence rules exist that allow the comparison of extensions (see Section 2.2). In this Chapter, we will discuss one of those correspondence rules, namely *formal linkage,* which—together with appropriate interpretation maps—allows us to compare different domains.

For theories that are related by linkage relations, the theory-ladenness of this language need not be an obstacle to comparing the domains of application, as my examples will also illustrate: since the linkage relates the *bare theories* whose corresponding domains of application we compare. Therefore, combining linkage between bare theories, with the domain of application's reflecting the bare theory (its being 'theory-laden'), and with the empirical data about the domain of application, we *can* compare the domains of application after all. Thus we do not need a putative "theory-free" language to compare

[5] Thus, in this Chapter, 'model' temporarily has the usual meaning in philosophy of physics. I will go back to my own conception of 'model' when we turn to dualities, in the next Chapter.

them.

Thus the practice of determining whether the elements of the domains of application of two theories are the same includes a combination of *theoretical* and *empirical* considerations, as I also discussed in detail in the previous Chapter.

3.1.2 The conception of emergence

Consider the opening words of the *Stanford Encylopedia of Philosophy*'s entry on 'Emergent Properties' (its only entry on 'emergence'):

> Emergence is a notorious philosophical term of art. A variety of theorists have appropriated it for their purposes ever since George Henry Lewes gave it a philosophical sense in his 1875 *Problems of Life and Mind.* We might roughly characterize the shared meaning thus: emergent entities (properties or substances) 'arise' out of more fundamental entities and yet are *'novel' or 'irreducible'* with respect to them... Each of the quoted terms is slippery in its own right, and their specifications yield the varied notions of emergence.[6]

On this general characterisation, emergence is a two-place relation between two sets of entities: the emergent entities, and the more fundamental ones (I will dub them the 'top' and 'bottom' entities, respectively). The top entities are 'novel' or 'irreducible' compared to the bottom ones.

'Irreducibility' here matches Broad's 'lack of deducibility' quoted in the preamble of this Chapter. As the italicized phrase in the above quote suggests, irreducibility and novelty are related: which naturally leads us to 'novelty' as the weaker, more general alternative. For irreducibility is surely indicative of some type of novelty: if the top entities are irreducible compared to the bottom ones, then surely they have something novel. But the other way around is not the case: novelty need not be expressed as irreducibility. And so, I propose that 'novelty', in this general and—admittedly—still unspecified sense, is the more general notion that one should seek to define emergence. Part of our task in developing a theory of emergence, then, is to further clarify what we mean by 'novelty'.

The above discussion is echoed by the recent philosophy of physics literature, which sees 'emergence' as a "delicate balance" between dependence, or rootedness, and independence, or autonomy (and the accounts also often compare theories rather than individual entities).[7] Roughly speaking, dependence means that there is a *linkage* between two theories (or between two items of a given theory): while *in*dependence means that there is *novelty* in one theory with respect to the other (or between the two items of the given theory). This matches the two aspects—the two-place relation and the characterisation of the top entities as 'novel'—of the above quote.

[6] O'Connor and Wong (2002): the second Italics is mine. So far as I know, the broad idea of emergence goes back at least to 1843, when it was introduced by J. S. Mill as the idea that "adding up the separate actions of the parts does not amount to the action of the living body", even though he did not use the word 'emergence'. Quoted in Landsman (2013: pp. 379-380).

[7] See e.g. Humphreys (2016: p. 26), Bedau (1997: p. 375), Bedau and Humphreys (2008: p. 1), Butterfield (2011: §1.1.1).

In this Section, I shall make these notions more precise, within the framework for theories from Chapter 1. I will take these two aspects—*linkage* and *novelty*—to align with, or to apply to, the two aspects comprising a theory: viz. the *bare theory* and its *interpretation.* Since these two aspects lie "along different conceptual axes", they can be happily reconciled: which makes it unnecessary to define novelty somehow as a 'lack of linkage', 'lack of deduction' or 'irreducibility'. In this subsection, I discuss linkage, and in the next, novelty.[8]

Thus I take *linkage* to be a *formal,* i.e. uninterpreted, relation between bare theories (or between two items of a given bare theory), hence as an inter-theoretic relation between *bare theories* (in some cases, we will compare different models, i.e. solutions, of a single bare theory). More precisely, linkage is an asymmetric relation, **link**, among the bare theories in a given family—with the family containing at least two theories. For the case of just two theories, it is a surjective and non-injective map, denoted as: $\textbf{link} : T_{\rm b} \to T_{\rm t}$. Here I adopt Butterfield's mnemonic, whereby $T_{\rm b}$ stands for 'best, bottom or basic', and $T_{\rm t}$ stands for 'tainted, top or tangible', theory. Here, $T_{\rm b}$ normally denotes a single theory, but it can be used to denote a family of theories. So, the idea is that the bare theory, $T_{\rm b}$, contains some variable(s) which provide(s) a more accurate description of a given situation than does $T_{\rm t}$.[9] (See the three cases of linkage, just below).

The idea of the linkage map, **link**, is that it exhibits the bottom theory, $T_{\rm b}$, as approximated by the top theory, $T_{\rm t}$. The map's being surjective and non-injective embodies the idea of 'coarse-graining to describe a physical situation' (but linkage is not restricted to mere coarse-graining: see below). The linkage map (and so, the broad meaning of 'linkage', as an inter-theoretic relation, used here) can be any of three kinds which do not exclude one another, but are almost always found in combination:

(i) *A limit in the mathematical sense.* Some variable (e.g. $V, c, \hbar \in \mathbb{R}_{\geq 0}$ or $N \in \mathbb{N}$) of the theory $T_{\rm b}$ is taken to some special value (e.g. $V \to \infty$, $N \to \infty$, $c \to \infty$, $\hbar \to 0$), i.e. there is a sequence in which a continuous or a discrete variable is taken to some value. $T_{\rm b}$ then refers to the sequence of theories obtained when the variable is taken to the limit, while $T_{\rm t}$ is the limit theory. In actual practice, the limit often comes with other operations, such as in (ii) immediately below—hence my warning above that these three kinds of linkage maps 'are almost always found in combination'.[10]

(ii) *Comparing different states or quantities.* Emergence often involves comparing a given physical situation, or system, or configuration of a system, to another that resembles

[8] In his excellent, and very comprehensive, book, Humphreys (2016: p. 26) distinguishes four features of emergence: dependence, novelty, autonomy, and holism. He only regards dependence and novelty as necessary for emergence, which are precisely the ones I consider here. For Humphreys' rationale for distinguishing novelty and autonomy, see (2016: pp. 32-33).

[9] There are exceptions to this motivating idea, as we will see in the example of Section 3.3 (cf. also the discussion in Section 3.2.1): where the bottom theory will be more explanatory, but the top theory will be more empirically accurate. What is essential is that the two theories are linked as here required.

[10] In the case of dimensionful parameters such as c and $\hbar$, the limit is taken such that the parameter is large or small *relative to* other dimensionful parameters in the theory (such as the speeds of particles, in the case of c, and the value of energy differences and time intervals, in the case of $\hbar$). For a treatment of $c \to \infty$ limit in general relativity, see Malament (1986: pp. 192-193). For a treatment of the $\hbar \to 0$ case, see Landsman (2013: pp. 380-383). See also my example in Section 3.3.

it. This is often done by comparing some of the theory's *models*, i.e. the solutions of the theory (for an example of this, see Section 3.3.3). Such physical comparisons have correlates in the formalism, which can be implemented prior to interpretation, by comparing *different states or quantities* or different *models*, here considered as uninterpreted solutions of the equations (cf. Section 3.1.1), between the bottom and top theories. In these cases, $T_{\rm b}$ is the bare theory that describes the situation (or system, or configuration) one wishes to compare to, and $T_{\rm t}$ is the bare theory that describes the situation (or system, or configuration) that one compares. The linkage map is then the formal implementation of this comparison, e.g. the comparison between the formal, i.e. uninterpreted, models.

(iii) *Mathematical approximations* (whether good or poor). Here, one compares theories (or expressions within theories) mathematically: perhaps numerically, or in terms of some parameter(s) of approximation. For example, as when one cuts off an infinite sequence beyond a given member of the sequence (perhaps because the sequence does not converge, or perhaps because there is no physical motivation for keeping an infinite number of members of the sequence); or when, in a Taylor series, one drops the terms after a certain order of interest. This will implement the idea that "emergence need not require limits".

Let me make three comments about how these three ways, (i)-(iii), of linking a theory to another, can define the map **link** : $T_{\rm b} \to T_{\rm t}$. The first two comments are about how to apply (i)-(iii), while the third is about how linkage thus defined relates to emergence:

(1) Any of the three ways, (i)-(iii), of linking a theory to another, can define the map **link** : $T_{\rm b} \to T_{\rm t}$. A given linkage map almost always involves a combination of several of the above approximations (for examples of this, see De Haro (2019)). Also, the list is not meant to be exhaustive.

(2) *Bridge principles:* before (i)-(iii) can successfully define a linkage map, one may need to "choose the right variables", or make additional assumptions. For example, one may need to first recast the bottom theory in a more convenient or more general form (see an example of this in Section 3.3.1). This sometimes goes under the name of *bridge principles* or bridge laws, i.e. the map **link** : $T_{\rm b} \to T_{\rm t}$ between the bottom and top theories may involve additional assumptions such as choices of initial or boundary conditions, or defining new variables to link the two theories' vocabularies. However, since the comparison here is between *bare theories,* the bridge principles here meant are formal, i.e. uninterpreted, so that—recall the distinction at the end of Section 3.1.1—the 'language' just mentioned is the language of the bare theory, not the language we use to talk about the domain: thus it is mathematical rather than physical language.[11]

(3) The ways of linking, (i)-(iii), as here defined are formal i.e. not interpretative, as I stressed in (ii): they link the bare theories, but they are physically motivated, and they often have a correlate in the theory's interpretation (which serves as a constraining affordance on the kind of linkage map one is likely to adopt). Just *how* the theory's inter-

[11] Bridge laws or correspondence rules, which I here schematically dub 'bridge principles' (to avoid the physical connotations of 'correspondence' and 'laws') are discussed in Nagel (1949: p. 302) and (1961: p. 354). The latter reference discusses two roles (as 'conditions of connectability' and 'derivability') that bridge principles fulfill, and gives three types of bridge laws. My bridge principles here are more restricted, in that they connect formal, i.e. uninterpreted theories, as explained in the main text. For a discussion of "ontological reduction", see the end of Section 3.1.6, especially footnote 41.

pretation correlates with the formal linkage map is in fact the very question of emergence itself, as I will discuss in the next few Sections:

Emergence. We have emergence iff two bare theories, $T_{\rm b}$ and $T_{\rm t}$, are related by a linkage map, and if *in addition* the interpreted top theory has novel aspects relative to the interpreted bottom theory.[12]

The linkage map thus specifies the dependence part of the emergence relation. To characterise emergence, we still need to specify the novelty.[13]

Philosophers often distinguish between ontological and epistemic emergence. The intended contrast is, roughly, between emergence 'in the world' vs. emergence merely in our description of the world. Thus according to this distinction we can have two types of *novelty*: ontological or epistemic. Specifying novelty in the case of ontological emergence is the topic of the next Section.

3.1.3 Ontological emergence as novel reference

In this Section, I characterise the kind of novelty that is relevant to ontological emergence. First of all, let me point to an obvious question that comes to mind about the literature on ontological emergence. We saw, in Section 3.1.2, that it is reasonable to define emergence in terms of novelty: this being a more general notion than the stronger 'irreducibility'—and, as I mentioned, this also seems to reflect a consensus in the recent philosophy of physics literature. Thus one would expect that, when defining *ontological* emergence, one would try to further characterise the so-far unspecified notion of 'novelty' in terms that are relevant to the ontology of the entities or theories involved in a relation of emergence.

But the literature usually takes some highly *specific* metaphysical relation instead, whose relation to novelty is not explicitly addressed, and uses that to characterise the emergence relation between the top and bottom entities or theories. I think this shift to specificity is understandable taking into account some of metaphysicians' interests, but I will also argue that, lacking a general notion of ontological *novelty,* we run the risk of missing some more basic aspects of what we mean by 'ontological emergence'.

For example, Wong (2010: p. 7) defines ontological emergence as 'aggregativity generating emergent properties': 'Ontological emergence is the thesis that when aggregates of microphysical properties attain a requisite level of complexity, they generate and (perhaps) sustain emergent natural properties.' And Wimsatt (1997, 2007: Chapter 12) takes the *lack of aggregativity* of a compound as the mark of emergence.[14]

[12] This definition can also be adapted for the emergence of properties, entities, or behaviour within a single theory. I will use this in the example in Section 3.3.1.

[13] A general characterisation of novelty, as 'not being included in (the closure of) a domain', is given in Humphreys (2016: p. 26). Butterfield (2011: §1.1.1) adds 'robustness' (i.e. the top theory's independence from the bottom theory's details) as an *additional requirement* for emergence. While I agree that robustness is an important property to further analyse the *physical significance* of emergent behaviour, and I will implement it in my choice of the *emergence base* (i.e. the class of bottom theories with respect to which the top theory emerges) in Eqs. (3.3)-(3.5) below: I do not think robustness is a necessary requirement for a minimal ontological description of emergence: see footnote 29 for other choices of emergence base.

[14] I take 'lack of aggregativity' to be an example of ontological rather than epistemic emergence, since aggregativity is a property of entities rather than theories.

Further, in the *Stanford Encyclopedia of Philosophy* article quoted earlier, O'Connor and Wong (2002: §3.2) review various other uses of the phrase 'ontological emergence', none of which are obviously introduced as an explication of 'novelty': viz. ontological emergence as 'supervenience', as 'non-synchronic and causal', and as 'fusion'.

I will not criticise these accounts here, since my claim is *not* that they are incorrect,[15] but rather that the accounts—interesting as they are—are very *specific:* in fact, too specific to be able to deal with all the examples that physicists are interested in. For example, Wimsatt's (1997, 2007) 'failure of aggregativity' does not seem applicable in cases where one compares systems that are neither spatial components of each other, nor aggregates of components. In other words, the relevant relation between the levels is not always one of spatial inclusion or material constitution. Thus already my examples in De Haro (2019) are not covered by Wimsatt's criterion of failure of aggregativity. Also, it would seem that failure of aggregativity is most interesting *if it leads to novelty.* The same can be said about causation: it may play a role in important examples of emergence such as the mind, but it is hard to see how it plays a role in the kind of examples that I discuss in this thesis, which are, I believe, "garden variety" for the theoretical physicist, and involve *no causation* from the bottom to the top entities.

Another reason to be sceptical about too specific accounts of ontological novelty is the requirement of consistency with the general description of emergence reviewed in Section 3.1.2. As I said above, having settled for *novelty* as the general mark of emergence, one naturally expects an explication of emergence to point to novelty in the top theory's ontology. Thus I take it that the metaphysical accounts of emergence just discussed, in so far as they all capture aspects of emergence, aim to put forward a specific metaphysical expression of 'novelty'. This is apparent from Wong's mention of the 'generation of emergent natural properties': surely what makes these properties of the top theory 'emergent' (on pain of his own definition being circular) is that they are novel, relative to the properties of the bottom theory.

My account of ontological emergence is metaphysically pluralist in that, while it recognises that emergence is a matter of *novelty* in the world (as the general notion of emergence prompts us to say), it does not point to a *single* metaphysical relation (e.g. supervenience, causal influence or fusion) as constitutive of ontological emergence. And I believe that this is just right, for the reasons above: a single metaphysical relation does not seem to cover all the cases of interest. Indeed my main contention is that there is a more basic framework for 'ontological novelty' that needs to be developed before one fruitfully moves on to detailed metaphysical analyses. Furthermore, I take it as a virtue of the present framework that it allows for an explication of the phrase 'ontological emergence' without appealing to specific, and sometimes controversial, metaphysical notions. So it seems that an account satisfying these desiderata should exist (more on this in Section 3.2.2). In this sense, my account is "basic", and can perhaps best be seen as an attempt to state the "almost obvious" in a more precise way.

To discuss ontological novelty, then, I take my cue from Norton (2012), which is a perceptive discussion of the difference between *idealisation* and—what I shall call—*non-idealising*

[15] For a criticism of the supervenience account, see Butterfield (2011: pp. 948-956).

approximation. Roughly speaking, he proposes the following contrast:

> A [*non-idealising*] *approximation* is an inexact description of a target system.[16] An *idealization* is a real or fictitious, idealizing system, [possibly] distinct from the target system, whose properties provide an [exact or] inexact description of the target system' (p. 209).[17]

Norton summarises the main difference between the two as an answer to the question: 'Do the terms involve novel reference?' On Norton's usage, only idealisations introduce reference to a novel system—notice that the word 'system' here may refer to a real or a fictitious system. In the case of idealisation, there is an 'idealised system' that realizes the 'idealised properties'. In the case of a non-idealising approximation, an idealised *system* either does not exist, or is not accurate enough to describe the target system under study.[18]

I propose to define ontological novelty in terms of novel reference thus construed. Thus, this gives an additional condition on the linkage map defined in Section 3.1.2: namely, *the composition of the linkage map and the top theory's interpretation map,* $i_t \circ \mathbf{link}$*, must have an idealisation in its range*, in the sense that the top theory has a referent that is novel, relative to the bottom theory's domain of application.[19]

This is not *merely* the condition that the top theory must describe a physical system through its interpretation map, $i_{\rm t}$: for the top theory is linked to the bottom theory, and so this is an extra condition on the linkage map between the two theories.

Although this characterisation of ontological novelty as novelty of reference is, by itself, not a new construal of the notion, it will give an interesting criterion for *emergence*: one formalised in terms of maps, in Section 3.1.4.

There is 'novel reference' when the terms of the bare theory refer (via the interpretation maps introduced in Section 3.1.2) to new things in the world, i.e. elements or relations that are (a) not in the domain of application of the bottom theory, $T_{\rm b}$, that $T_{\rm t}$ is being compared to, but are (b) still in the world—they are in the domain of application of $T_{\rm t}$.

[16] Norton calls this an "approximation", but this I already used this word in Section 3.1.2 (iii) in the more common sense of a 'mathematical approximation', hence my addition of 'non-idealising' here. For it seems to me that doing an approximation does not, in general, prevent the approximation from being an idealisation, i.e. an idealising system could exist (and this is also the jargon adopted in the literature on reduction, see e.g. Schaffner (1967: p. 144) and Nagel (1979)). Therefore, I dub the more restrictive type (which does not play an important role in the rest of this Chapter) a 'non-idealising approximation'.

[17] Norton's use of the gerundive 'idealizing', here in his definition of 'idealization', is not circular, because 'idealizing', introduced as an adjective for 'system', is merely a label used to distinguish this system from the other system involved: namely, the 'target' system. See my use of these words in Section 3.2.1.

[18] An idealising system is sometimes called a 'limit system', the word 'limit' here echoing case (i) in Section 3.1.2. However, I deny that idealisations invariably require limits, because they only require a physical system described by a top theory (more on this in Section 3.1.5), and thus all of (i)-(iii) can correlate with an idealisation. For example, the case study in Sections 3.3.1-3.3.2, although naturally written as a limit, does not require one since the theory (as opposed to some of its models) contains no singularities. For this reason, I do not use the 'limit system' jargon.

[19] Humphreys (2016: p. 29) considers novelty within a single domain, while I here require novelty of one domain relative to another. The domain $D_{\rm t}$ is novel, relative to $D_{\rm b}$, if some of its elements or relations are not included in $D_{\rm b}$. For more details, see Section 3.1.5. For novelty of *entities*, see Humphreys (2016: p. 34).

Novelty is here not meant primarily in the temporal sense, since it is conceptual novelty[20] that counts (cf. Nagel (1961: p. 375)),[21] even allowing reference to other possible worlds (as Norton does when he refers to idealisations as referring to real or fictitious systems: p. 209; cf. also the definition of an interpreted theory in Section 3.1.1).

For example, Norton considers the eighteenth- and nineteenth-century theories of heat to be referentially successful, despite the fact that we now know that heat is not a fluid. According to Norton, the theory is nevertheless—

> referentially successful in that the idealizing system is a part of the same system the successor theory describes. The "caloric" of caloric theory refers to the same thing as the "heat" of thermodynamics, but in the confines of situations in which there is no interchange of heat and work.

I propose to read Norton's statement above as saying that the theory is referentially successful because it refers to an idealised world[22]—a possible world in which there is no interchange of heat and work—that approximates the target system of interest well, according to standard criteria of empirical adequacy: and that as such it is referentially successful.[23] I develop this suggestion in the next Section.

3.1.4 Ontological emergence, in more detail

In this Section, I propose an answer to the question of how to characterise ontological emergence for theories formulated as in Section 3.1.1. My proposal is based on a restatement of the notion of 'novel reference': namely as the failure of the interpretative and linkage map to mesh, i.e. to commute, in the usual mathematical sense of their diagram's "not closing". This idea is inspired by Butterfield (2011a: §3.3.2), though I here consider it for *interpretations* rather than for bare theories: a distinction which will turn out to be important (cf. Section 3.1.6).

[20] Humphreys (2016: p. 39) distinguishes ontological from conceptual novelty. His 'conceptual novelty' corresponds, roughly, to what I call 'epistemic emergence' (see Section 3.1.5). His 'ontological novelty' does not exactly match mine, which contains aspects of his ontological and his conceptual novelty. By 'conceptual' I here mean the qualitative notions involved in interpreting a theory (e.g. the notion of 'simultaneity' in classical mechanics as a relation between events in the world), rather than the uninterpreted notions of the bare theory. This is why 'conceptual', on my construal, pertains to the interpretation of the theory, and hence to its ontology, rather than to its descriptive apparatus.

[21] Of course, the linkage map does not exclude the temporal dimension: 'diachronic' emergence is obtained if the linkage map entails "evolving the time parameter". However, this will not be my focus, since it is a special case of the general linkage map. For interesting treatments of diachronic emergence, see Guay and Sartenaer (2016: pp. 301-302, 316-317) and Humphreys (2016: pp. 43-44, 66, 70-72).

[22] Norton's suggestion that novel reference characterises idealisation seems to rule out purely *instrumentalist* interpretations of theories: but only in so far as such interpretations can successfully avoid ontological questions (which is far from being an innocuous condition!). Thus from now on, the interpretations considered will not include strongly instrumentalist interpretations of theories. Other philosophical positions, such as empiricism, are of course supported by the framework: see footnote 46. For a critique of instrumentalism in physics, see Wallace (2012: pp. 24-28).

[23] While this Chapter does not assume a scientific realist position (see the preamble and Section 3.2.2), the above is of course compatible with extensional scientific realism. Namely, Norton's phrase 'confines of situations in which there is no interchange of heat and work' denotes the circumstantial physical conditions under which the extensions of 'heat' and 'caloric' agree.

I begin with two remarks which further specify the context to which these concepts apply:

(1) *Emergence of theories: and of models, entities, properties, or behaviour.* I will formulate the distinction in terms of *theories*, but the idea can be equally well formulated in terms of models (cf. Section 3.1.1), entities, properties, or behaviour (and I will adopt the latter perspective in the example in Section 3.3).

(2) *The space of theories.* Properly defining a linkage map, especially if it involves a mathematical limit of a theory (so that one can then discuss the sequence of corresponding interpretations), often requires introducing the notion of a space of theories related by the linkage map. Section 3.3 will take a simpler approach and consider sequences of states, quantities, and dynamics, without explicitly specifying the full details of the spaces in which the sequences are defined, so as to simplify the presentation.[24]

Section 3.1.5 contains the core of this Section and the main proposal of this Chapter, i.e. the explication of ontological emergence.

3.1.5 Epistemic and ontological emergence as commutativity and non-commutativity of maps

In this Section, I reformulate the notion of novel reference in terms of the non-commutativity of the interpretative and linkage maps, or their "failure to mesh": which I will use to give an explication of ontological emergence. I here mean non-commutativity of maps in the usual mathematical sense that, depending of the order in which the maps are applied, the image (value, output) that a given argument (input) is mapped to varies—the mismatch thus expressing the *novelty* of the reference.

Thus the starting point is a reformulation of the idea of novel reference, in terms of my notion of interpretation of a bare theory (Section 3.1.1). Recall that an interpretation is a surjective map, i, from the bare theory, T, to a domain of application, D, $i : T \to D$, preserving appropriate structures. Norton's idea can then be formulated as follows.

Consider two bare theories, a bottom theory, $T_{\rm b}$, and a top theory, $T_{\rm t}$. Their interpretations are corresponding surjective, structure-preserving maps (by our assumption, in the preamble of Section 3.1.1, that the interpretation is "sufficiently good"), from the theories to their respective domains of application. That is, there are maps $i_{\rm b} : T_{\rm b} \to D_{\rm b}$ and $i_{\rm t} : T_{\rm t} \to D_{\rm t}$.

Now, consider, as in Section 3.1.2, a linkage map, $\textbf{link} : T_{\rm b} \to T_{\rm t}$, mapping the bottom theory to the top theory. There are two interpretative cases, as follows.

(1) Allowing epistemic emergence. If the two theories describe the same "sector of reality", i.e. the same domain of application in the world, then the ranges of their

[24] Thus I refrain from a more ambitious technical aim, of trying to define general conditions of emergence (and, especially, the aim of characterising the robustness of the emergent behaviour—which, as I said in footnote 13, I do not think is required for emergence) in terms of the underlying topology, and choice of it. For an example of deploying topology in the "space of theories", cf. Fletcher (2016).

$$\begin{tikzcd}
T_{\mathrm{b}} \arrow[rr, "\mathrm{link}"] \arrow[dr, "i_{\mathrm{b}}"'] & & T_{\mathrm{t}} \arrow[dl, "i_{\mathrm{t}}"] \\
& D &
\end{tikzcd}$$

Figure 3.1: Possibility of epistemic emergence. The two interpretations describe "the same sector of reality", so that $i_{\mathrm{b}} = i_{\mathrm{t}} \circ \mathrm{link}$.

interpretations must coincide. (Alternatively, the top range must be contained in the bottom range, i.e. $D_{\mathrm{t}} \subseteq D_{\mathrm{b}}$: a condition that I will discuss separately, below). For the range of the interpretation is what the theory purports to represent: and these two theories represent the same sector of reality. This situation of identity is described by the commuting diagram in Figure 3.1. As is clear from the diagram, there is no novel reference here, because the domains of application in the world are the same.

Notice that, since the domains of application are the same, the two maps map to the same elements within that domain of application (the same experiments, interactions, correlations, etc.): even though, as interpretations, they are of course different, because they map from different theories (and so, one theory could describe water in terms of molecular dynamics of the single molecules, or in terms of the dynamics of groups of molecules, even though both describe the same empirical data). In such a case, the two interpretations are related as follows: $i_{\mathrm{b}} = i_{\mathrm{t}} \circ \mathbf{link}$.

Another type of epistemic emergence is if the bare theories, T_{b} and T_{t}, are equivalent bare theories. That is, we consider the commuting diagram in Figure 3.1, but weaken the linkage map **link**, allowing it to be injective. There is no novelty in the world, but only in the bare theory and the interpretation map (but not in its range).[25] Dualities in physics give vivid examples of this type of emergence:[26] see Castellani and De Haro (2019).

Since in Figure 3.1 there is no novel reference, there cannot be ontological emergence. However, there can be *epistemic emergence*: this will be the case when the two interpreted theories describe the domain of application differently. And so, the top theory T_{t} can have properties or descriptive *features* that are indeed novel, and striking, relative to the bottom theory, T_{b}. The novelty of these properties and descriptive features, however, is not a matter of *reference*: it does not rely on a difference in elements or relations between the bottom and top domains of application, since by definition we have a commuting diagram in Figure 3.1. Rather, the novelty lies in the different relations between the bare theory and the domain that i_{t} establishes, compared to i_{b}, i.e. in the kind of statements that the interpreted theory makes about the domains of application.

[25] Another interesting case is equivalent theories with *different* interpretations. Though there is ontological novelty here, there is no ontological emergence because the linkage map is injective: see van Dongen et al. (2020: footnote 52).

[26] As discussed in van Dongen et al. (2020: Section 4.4), De Haro (2015: p. 118), and Dieks et al. (2015: pp. 208-209), dualities of this type only give cases of epistemic emergence. The reason is that the linkage map between the two theories is an isomorphism, which is in tension with the asymmetry (non-injectivity) required for ontological emergence.

I now discuss the special case of proper inclusion, i.e. $D_{\mathrm{t}} \subset D_{\mathrm{b}}$. The reason for allowing this case is that there are surely cases of epistemic emergence in which the diagram "does not close". If i_{b} is a "very fine-grained interpretation", "respecting" the distinction of the many micro-variables of T_{b}, i.e. an injective map, while link is of course non-injective, then we have a diagram which cannot possibly commute on all arguments. But this could be a case of mere coarse-graining with no ontological novelty, while the failure of the diagram to close would seem to suggest ontological emergence. In such cases, empirical adequacy (see Section 3.1.1) on the domain D_{b}, so that it is a good range for i_{b}, requires D_{b} to "contain coarse-grained facts", since i_{b} needs to be surjective on D_{b}. Thus in such cases, "good" interpretations of T_{b} include a classification of microstates into macrostates.

To sum up: cases where the map **link** is *mere coarse-graining*, which results in the "dropping of certain fine-grained facts from the domain", do not count as ontological emergence because then D_{t} is a proper subset of D_{b}.

Let me spell out the possibility of epistemic emergence a bit more. Though the domains and ranges of i_{b} and $i_{\mathrm{t}} \circ \textbf{link}$ are the same, the interpretative maps i_{b} and i_{t} are different and have different theories as their domain of definition. Thus the novelty lies in the existence of a new theory, T_{t}, with its own novel interpretation map, i_{t}; and this novelty is relative to the bottom theory, T_{b}, with its interpretation map i_{b}. Therefore the emergence is epistemic, since it takes place at the level of the theory and its interpretation map (while the domains are the same): it is emergence in the theoretical description, rather than emergence in what the theory describes.[27]

Epistemic emergence should not be dismissed as "merely" a matter of "words", or even "taste". For theories, and interpretations of the kind here discussed, are not the kinds of things that one can choose as one pleases—there may be virtually no freedom in choosing an appropriate linkage map giving rise to a *consistent* theory T_{t} whose interpretation is *as good* as that of T_{b}—i.e. the ranges of their interpretations *must be* identical (cf. Section 3.1.6).

The above is, as it should be, only a *necessary condition of epistemic emergence*: as a difference in theoretical description. To give a sufficient condition for epistemic emergence, one needs a more precise characterisation of epistemic novelty. Epistemic novelty is often characterised in terms of computational or algorithmical complexity, chaotic behaviour, possibility of derivation only through simulation, etc. Alternatively, it is seen as the failure of prediction or of explanation.[28] But since epistemic emergence will not be my

[27] Epistemic emergence is a matter of novel description by the bare theory, and so an anonymous reviewer asked: 'How do we determine what a bare theory describes, prior to interpretation?' The answer is that we do not need to determine epistemic emergence in complete independence of the interpretation map. However, a bare theory, even if physically uninterpreted, already comes with a mathematical physics language that allows us to see the novelty in the bare theory: it is the linkage function itself that introduces this epistemic novelty (see Section 3.1.1). In the case of dual theories with the same domain of application, the two theories give different effective descriptions of this domain, because of the approximations introduced: for details, see Castellani and De Haro (2019: Section 2.2). In both cases, the novelty is in both the bare theory and in the interpretation map, but—crucially—the interpretation map's range is the same, and the diagram in Figure 3.1 closes: and so, the distinction with the case of ontological emergence, to be made below, *is* sharp.

[28] For more details, see, for example, Bedau (1997: pp. 378-379), Bedau and Humphreys (2008: p. 16),

main topic in this Chapter, I will leave such further characterisations aside: for an example of phonons, see De Haro (2019: pp. 29-38). For examples of dualities, see Castellani and De Haro (2019: §2.3).

(2) Non-commutation: ontological emergence. But when there is *novel reference*, the two interpretations refer to different things—they may even refer to different domains of application. For example, hydrodynamics describes the motion of water between $0^{\rm o}$ C and $99^{\rm o}$ C at typical pressures, while the molecular theory of water potentially describes it in a different domain of application, since it can e.g. explain the boiling of water as the breaking of the inter-molecular hydrogen bonds due to molecular excitations. Here, one should not confuse the fact that the interpretations of the two theories attempt to explain the properties of the same target system (e.g. turbulence in a fluid) with the distinctiveness of their ranges: each of which is subject to its own conditions of empirical adequacy, within its relevant accuracy.

In all such cases of novel reference, the domains of application are different:

$$D_{\rm t}\not\subseteq D_{\rm b}\ , \tag{3.1}$$

and therefore so are the interpretation maps, that is:

$$\begin{aligned} i_{\rm t}\circ\text{link} &\neq i_{\rm b} \\ \text{ran}(i_{\rm t}) &\not\subset \text{ran}(i_{\rm b})\ . \end{aligned} \tag{3.2}$$

Thus the range of the interpretation of the top theory, $T_{\rm t}$, is not the same as the range of the interpretation of the bottom theory, $T_{\rm b}$, nor is it a subset of it. This is what I call *ontological emergence*, and it is pictured in Figure 3.2.

There are cases in which $T_{\rm t}$ is not a good approximation to a single theory $T_{\rm b}$ (cf. (i) and (iii) in Section 3.1.2), but only to one of the members of a *family* of theories $T_{\rm b}$. Let us label this family by x, so that the bottom theory becomes a function of x, which we write as $T_{\rm b}(x)$. In other words, we take the emergence base (i.e. the class of bottom theories with respect to which the top theory emerges: cf. footnote 13) to be the *entire family* of theories, $\{T_{\rm b}(x)\}$, for any x in an appropriately chosen range. We will see this at work in the example of Section 3.3.

Since the emergence base is the entire set of theories, $\{T_{\rm b}(x)\}$, the bottom theories all have the same interpretation *map*, $i_{\rm b}$, but the *range* of $i_{\rm b}$ now depends on, i.e. varies with, x through the bottom theory, i.e. we evaluate $i_{\rm b}(T_{\rm b}(x))$.

Now take x to be bounded from below by 0 and consider, as a *special case,* a linkage map that is taking the limit $x\rightarrow 0$, i.e.:

$$T_{\rm t}:=\lim_{x\rightarrow 0}T_{\rm b}(x)\ . \tag{3.3}$$

In this case, the statement of ontological emergence, Eqs. (3.1)-(3.2), amounts to:

$$\text{ran}(i_{\rm t})\not\subseteq\text{ran}\,(i_{\rm b})|_{x=0} \tag{3.4}$$

Humphreys (2016: pp. 144-197).

$$\begin{tikzcd}
T_{\rm t} \arrow[r, "i_{\rm t}"] & D_{\rm t} \\
T_{\rm b} \arrow[u, "\text{link}"'] \arrow[r, "i_{\rm b}"] & D_{\rm b} \arrow[u, "/" marking]
\end{tikzcd}$$

Figure 3.2: The failure of interpretation and linkage to commute ($i_{\rm b} \neq i_{\rm t} \circ \text{link}$) gives rise to different interpretations, possibly with different domains of application, $D_{\rm b} \neq D_{\rm t}$.

or, perhaps more intuitively:[29]

$$D_{\rm t} := i_{\rm t}(T_{\rm t}) \nsubseteq D_{\rm b}|_{x=0} := i_{\rm b}(T_{\rm b}(x))|_{x=0} \; . \tag{3.5}$$

In other words, the domains that we get by interpreting the top theory directly, or by first interpreting the bottom theory and then setting $x = 0$, are different, as we will see in the example below.[30]

In view of the difference between Figures 3.1 and 3.2, we can now reformulate novel reference as *the linkage map's failure to commute, or to mesh, with the interpretation.* When linkage and interpretation do not commute, the two interpretation maps are different because the theories refer to different domains, and different systems—even if the underlying physical object, like the sample of water above, may of course be the same.

It is possible to formulate a *robustness* condition for the emergent behaviour—which I will assume in the rest of this Chapter—as the condition that the emergence base is a whole class of theories, i.e. $\{T_b(x)\}$ for a range of values of x, rather than a single theory

[29] The condition $x = 0$ in the subscripts on the right-hand side is motivated by the condition, mentioned above, that we take the *entire class* of theories, $\{T_{\rm b}(x)\}$, to belong to the *emergence base.* But there are other possible choices of this base. In particular, by analogy with the case of just two theories, it would suffice that $D_{\rm t} \neq D_{\rm b}(x_0)$, where x_0 is some reference value of x, for us to claim that the top theory emerges from the bottom theory (where the bottom theory is here identified with $T_{\rm b}(x_0)$). However, it is in most cases better motivated to regard the whole class of theories, for any x, as the emergence base, i.e. to take $\{T_b(x)\}_{0\leq x\leq x_0}$ as the emergence base, under the assumption that $\text{ran}\,(i_{\rm b}(x)) = \text{ran}\,(i_{\rm b}(x'))$ (where $i_{\rm b}(x)$ is $T_{\rm b}(x)$'s own interpretation map) whenever $x, x' \in [0, x_0]$. For then $T_{\rm t}$ is novel relative to the whole class of theories $\{T_b(x)\}_{0\leq x\leq x_0}$, all of which have the same interpretation, $i_{\rm b}$. It is *this* stronger requirement that I follow in this thesis, and it amounts to a *robustness* condition. The requirement can be easily weakened as needed, by evaluating the right-hand side of Eq. (3.4) at some other point x, or by adapting the emergence base in a way appropriate for the situation at hand. Also Humphreys (2016: p. 28) emphasises the relativity of 'emergence' to the emergence base: my choice $x = 0$, on the right-hand side of Eq. (3.4), is how my framework reflects this. Additional motivation for this choice is as follows. Assume that there were some $x' \in (0, x_0)$, such that $\text{ran}\,(i_{\rm b}(x)) \neq \text{ran}\,(i_{\rm b}(x'))$. Then the theory at $x = x'$ is already emergent: in other words, "we did not need to go to $x = 0$ to get emergence". So, it is best to take the emergent theory, $T_{\rm t}$, to be the one whose domain is distinct from the domains of all the theories at $x \in [0, x_0]$ (alternatively, to define x such that $D_{\rm t}$ is distinct from $D_{\rm b}(x)$ for the entire range of x).

[30] This also applies to cases where, instead of a continuous parameter x, we have a sequence of bottom theories, $T_{\rm b}, T'_{\rm b}, T''_{\rm b}, \ldots$, and a sequence of interpretations $i_{\rm b}, i'_{\rm b}, i''_{\rm b}, \ldots$ In such cases, ontological emergence lies in the sequence of interpretative maps not converging to, or being distinct from, the given approximation (cf. (i) and (iii) in Section 3.1.2), the top theory's interpretative map, $i_{\rm t}$.

for fixed x. Thus, emergence is robust if the interpretation does not change as we vary x over the base (see footnote 29).

To sum up: the meshing condition between the linkage and interpretation maps, formulated as the lack of commutativity or of closing of their joint diagram, Figure 3.2, should be seen as a reformulation of novelty of reference—a reformulation that gives a straightforward formal criterion that can be used in examples.

Let me briefly discuss the relation between reduction and emergence. I will endorse the philosopher's traditional account of reduction: namely, Nagel's view[31]—as, essentially, deduction of one (here, bare) theory from another, almost always using additional definitions or bridge-principles linking the two theories' vocabularies.[32] Not all linkage maps will be reductive. For the relations between physical theories are often much more varied than logical deduction with bridge principles allows (for example, the use of limits and related procedures, such as renormalization, often adds content to the theory). But when the linkage *is* reductive, the account explains at once *why* ontological emergence and reduction are independent of each other: for the former is a property of the theory's *interpretation* (i.e. the novelty in the domain), while the latter, understood as a formal relation, is a property of the linkage map only, i.e. a relation between *bare theories*.[33]

3.1.6 Is emergence ubiquitous? Regimenting the uses

In this Section, I point out two properties of the regimentation of 'emergence' I have proposed, that have the effect of reducing the number of putative cases of emergence; and I analyse the extent to which the framework gives a clear-cut criterion of emergence.

Understanding ontological emergence as novelty of reference, and epistemic emergence as novelty of theoretical description, gives a helpful regimentation of the uses of the word 'emergence'. One problem which seems to have plagued the literature is the apparent ubiquitousness of emergence, which would seem to be sanctioned by my logically weak conception in Section 3.1.1 (cf. Chalmers (2006: §3)).

In addressing this problem, the first point to note is that, as I stressed: given the wealth of examples in which physicists justifiably talk about emergence (and which the philosophical literature has also endorsed), pervasiveness is something one may, to some extent, expect and accept.[34] But, more important: my regimentation has two implications which amount to strengthening the conception of ontological emergence:

[31] Cf. footnote 6.

[32] By the 'vocabulary' of a bare theory, I here mean the mathematical rules and the words of ordinary or technical language used to express propositions within the bare theory, as well as its rules of inference.

[33] Notice that the formal reduction between bare theories discussed here contrasts with the more controversial notion of "ontological reduction". For example, Hendry (2010: pp. 184, 188) distinguishes between inter-theoretic reduction (which is what philosophers have traditionally meant by reduction: see the main text) and ontological reduction. The former sense will be my main meaning of 'reduction': and so, when I say that "emergence and reduction are compatible", I mean reduction in this sense. See also the end of Section 3.1.6, especially footnote 41.

[34] I agree with Humphreys (2016: pp. 54-55), who criticises the 'rarity heuristic', i.e. the idea that a construal of emergence must make it a rare property in order for it to be correct. Ontological emergence may well be pervasive: what I aim at here is clarifying how it is constrained. See also Bedau and Humphreys (2008: pp. 12-13).

(1) *The requirement that T_t, and a sufficiently good interpretation of it, must exist.* My conception makes emergence less pervasive than it might at first sight appear, because of the requirement that $T_{\rm t}$ must be a *bare theory*, presented in the same form as $T_{\rm b}$: usually as a triple, subject to the requirements in Section 3.1.1. Thus one cannot take an arbitrary bottom theory, apply to its elements $\mathcal{S}$, $\mathcal{Q}$, and $\mathcal{D}$ some arbitrary map one calls 'link', and claim that one has emergence: for one is not guaranteed to get as the range of the map labelled 'link' a *bare theory*! Likewise for the interpretation, which must be sufficiently good, in the sense of the preamble of Section 3.1.1: it must be empirically adequate and the interpretation map must be surjective. Again, some arbitrary map one calls 'link' will in general not secure empirical adequacy of $T_{\rm t} := \text{ran(link)}$, nor will the thus-obtained theory describe the whole domain.

(2) *The requirement of novel reference, for ontological emergence.* Interpretations that are genuinely novel (because they refer to novelty in the physical system) and, at the same time, empirically adequate to within required accuracies, need not be abundant. The reason is as follows:

Ontological emergence requires that we have a case of idealisation (see Section 3.1.3), so that there is a physical system to which the map refers. And this is, in fact, a considerable restriction on the linkage map: Norton (2016: p. 46) states that 'merely declaring something an idealization produced by taking a limit is no guarantee that the result is well-behaved.' In other words, 'the limit operations generate limit properties only. They do not generate a single process that carries these properties, for these properties are mutually incompatible' (p. 44).[35]

Ontological emergence thus crucially depends on the choice of ranges: for $D_{\rm t} \not\subseteq D_{\rm b}$ is the mark of novelty of reference. The ranges are to be determined by the theory's best interpretation, which is subject to empirical adequacy and the other constraints mentioned in Section 3.1.1 (see also below).

To what extent do we get a clear-cut criterion of ontological emergence, as I promised in the preamble of this Chapter? While I do not claim that formulating ontological emergence formally *automatically* dispels problems of ontology, which can be subtle (see Section 3.2.2): the criterion does give us a recipe for assessing whether ontological emergence obtains: it requires us to formulate theories, their interpretations, and the linkage

[35] Norton (2016: Abstract, and p. 46) has for example argued that there is no limit process that can be identified as a thermodynamically reversible process (as is standardly assumed by the notion of thermal equilibrium). According to Norton, the notion of 'equilibrium', as construed in thermodynamics, ascribes contradictory properties to the system. This amounts to denying that thermodynamically reversible processes are idealised processes at all. On this reading, thermodynamics (as standardly construed) could not emerge from statistical mechanics, because no actual physical system (real or fictitious) can be in thermal equilibrium, on the standard construal of the notion. Whatever one thinks of this particular example (for an alternative analysis, see Lavis (2017)): Norton's general argument indeed shows that my requirement, that there is ontological emergence only if the composition of the linkage map and the interpretation map, $i_{\rm t} \circ \text{link}$, maps to an idealisation, is a substantial requirement. Indeed the example shows one way in which the top theory's interpretation map can fail to refer: namely, if the linkage map is such that the domain of application of the top theory is ascribed contradictory properties.

relations, as formally[36] as possible. Once that work is properly done, the decision about ontological emergence follows as the non-meshing of linkage with interpretation, provided both theories refer. The challenge lies in adequately and formally formulating interpretations and linkage maps. To illustrate, let us return to the example of water. Consider the questions:

Is liquid water "more than" the sum of molecules and their interactions? Should we consider liquid water as different from its molecular basis? What distinguishes the relation between the liquid state and its molecular basis from its cousin state, ice, and *its* molecular basis?

My proposal does not do away with the need to ask such questions, at one point or another. These questions will get implicitly answered once sufficiently precise interpretations have been developed. But, on my account, these are not the important questions for the philosophical account of emergence. The recipe is to develop as accurate as possible interpretations of the theories, subject to empirical adequacy and the other requirements from Section 3.1.1, and to then allow those interpretations to answer the question for us.

Interpretations allow for comparisons: one interpretation is better or worse than another because it covers more or fewer cases, and is more or less empirically adequate and precise, etc. So, one now asks: is there, by the interpreted bottom theory's own lights, such a thing as 'liquid water'? Or does the range, $i_{\rm b}\,(T_{\rm b})$, only contain items such as 'the collection of interacting molecules'? Here, one should not try to judge, *independently of interpreting theories,* whether 'liquid water' is the same as, or can be empirically distinguished from, 'a collection of molecules'. The right question is whether the interpreted bottom theory, in its range, $i_{\rm b}\,(T_{\rm b})$, has the hydrodynamic substance 'liquid water' as an element—for recall, from Section 3.1.5 (1), that we could have epistemic emergence if the range of the top theory was a subset of the range of the bottom theory. Now if our bottom theory is a theory of molecules and their interactions, it does *not* have 'liquid water' in its range: and so water is, indeed, ontologically emergent in the top theory, $T_{\rm t}$, with respect to its molecular basis in $T_{\rm b}$. For on an interpretation containing only collections of molecules, there is never a continuous state of matter, however numerous the number of molecules, and however small the size of the molecules, may be. This is because a discrete set of molecules and a continuous medium are *referentially distinct* (recall that interpretation makes reference to concrete objects: it is not just more theory!). Such an interpretation may not need to distinguish, for its limits of accuracy, between 10^{40} and $10^{40}+1$ molecules (the number of molecules may only be defined up to 5% accuracy, say). But it *will* distinguish a discrete from a continuous medium, because these are different kinds of objects, and so they constitute different domains.

The above is best understood when formulated using intensional semantics. The hydrodynamic and molecular theories have the same extension but different intensions (liquid water vs. a large collection of interacting molecules). The intensions differ because the theories are different and, although they may share many common terms (such as 'volume', 'pressure', 'temperature', etc.), 'liquid water' is not one of the terms that they share. Nevertheless, they both refer to the same experimental phenomenon in our world

[36] Crucially, 'formally' is here taken in the official sense given in the preamble of this Chapter, as in 'mathematical philosophy': *not* in the sense of 'uninterpreted'.

and, within certain margins of accuracy, their predictions are the same.

Many examples of emergence are of this type: the theories have the same extension, but different intensions. Of course, it is also possible that both the extensions and the intensions differ.[37]

The distinction between extensions and intensions also further clarifies the criteria of identity of domains, from Section 3.1.1. For sameness of extension is chiefly (of course not solely!) determined by an empirical procedure, while sameness of intension is conceptual, and determined from within the theory[38] (again, not without empirical input!).

Finally, I stress that there can be no "fiddling" with the interpretation in establishing whether there is emergence. This is the strategy I adopted in case studies in De Haro (2019: pp. 35-38, 45-48): interpretations must be fixed (by other criteria) before one asks about emergence. This is especially true for *intensions.* In other words, ontological emergence can only be predicated relative to interpretations that are sufficiently good (on the criteria given in Section 3.1.1): and different interpreted theories must be assessed against each other employing the usual evaluation criteria given by the philosophy of science and in scientific practice. Thus my proposal is to decide for one's best interpretation once and for all using independent methods, to take the interpretation literally, and to stick to it when enquiring about emergence. Changing one's interpretation in the course of assessing emergence is liable to lead only to confusion.

Reduction and emergence. One should not confuse the question of emergence we are asking here with the question of reduction. We already agreed that, in this example, there *is* reduction between the bare theories: emergence and reduction are compatible! (see Section 3.1.5).[39] The point is precisely that reduction in the sense meant here, when it obtains, is a relation between the *bare theories*,[40] while ontological emergence is in the range of the *interpretation.* Thus, *the reduction of the bare theory cannot imply a putative reduction of the interpretation.*[41]

[37] This possibility gives rise to the various cases considered in Section 3.2.1. Theories with the same extension but different intensions were previously considered, in the debates between the Nagelian vs. Kuhn-Feyerabend positions on reduction, by Scheffler (1967: pp. 60-63) and Martin (1971: pp. 19-21). Nagel (1979: pp. 95-113), together with his (1961: pp. 342, 366-374) can be taken to hint at the relevance of this distinction for emergence. However, I *disagree* with these authors in several other aspects: not least, Scheffler's (1967: p. 57) statement that 'for the purposes of... science, it is sameness of reference [extension] that is of interest rather than synonymy [intension]'.

[38] See for example Martin (1971: p. 21).

[39] Butterfield (2011: §3.1.1-(4)) reminds us that, in cases of reduction (via definitional extension) of T_t to T_b, T_t does not extend the domain of quantification of T_b ("it has no new objects"). I of course accept, as everyone must, this verdict: but it does not bear on interpretation maps, which in putative cases of emergence are *not* related by reduction. The question of emergence is whether the pictures of the world, that we get from the two interpretations, mesh with the reduction relation between the bare theories.

[40] As Nickles (1973: pp. 183, 185, 193-194) emphasises, ontological issues are not central in the *reductions* done by physicists, and most cases they consider are not cases of ontological reduction. Rather, the issues are typically explanatory, heuristic, and justificatory. My weak notion of reduction agrees with this.

[41] Recall, from footnote 33, the distinction between reduction of bare theories (i.e. formal reduction regardless of interpretation, which is my default meaning of 'reduction') and the more controversial notion of "ontological reduction". The latter is controversial because, in all the cases of emergence here considered, the intensions differ. See, for example, Schaffner's (2012: p. 548) review, where he explicitly states that the bridge principles (which he calls 'connectivity conditions') are *extensional* not intensional, since the meanings differ.

3.2 Further Development

In this Section, I take up two further questions that my framework for emergence prompts.

3.2.1 Which systems?

In this Section, I will further specify the systems that are related by the emergence relation. This question is prompted by the condition (ii), in Section 3.1.2, that the linkage map relates bare theories for which appropriate interpretation maps exist, which map to physical situations or systems.

To answer this question, let us recall the following distinction, from Norton (2012: p. 209) at the beginning of Section 3.1.3: The *target system* is the system we aim to describe: the chunk of material in the lab, the gas of Hydrogen ions (composite particles!), which are fed to the LHC through its linear accelerator, etc. The *idealising system* is the system to which our idealising theories refer, and it may well be distinct from the target system.

The question I will ask is about the comparison between the idealising system(s) and the target system. I will address this question by looking at the empirical adequacy of the idealising sytem(s) relative to our target system, i.e. by asking how accurate, or exact, is their description of the target system.[42] Thus the question is simple: how empirically adequate, or exact, is the description of the target system given by the idealisation?

Suppose that we have a theory, $T_{\rm b}$, describing a certain target system within given limits of accuracy. Then the domain of the theory, $D_{\rm b}$, contains a system that is close to our target system. Typically, the system described by the domain of the theory will not be the target system itself, but a closely related system—an idealised system. Thus the domain $D_{\rm b}$ does not literally describe our world, at least not in full detail, but a physically possible world—it gives an inexact description of the target system (as both idealisations and non-idealising approximations do). In some cases, however, we may not be able to tell the difference, and we may just identify the target system with the idealising system—but this is not a generic situation, since future experiments may expose the inexactness of the description. The same can be said of $T_{\rm t}$ and *its* domain, $D_{\rm t}$: it gives an inexact description of the *same* target system, even though it refers to an idealising system that is, in principle, distinct from it.

As I said above, I will not here attempt to give an account of under what conditions the idealising system can be identified with the target system. Rather, my aim is to assess how accurate the descriptions are that are given by the bottom and top theories, i.e. about their empirical adequacy. So, we have three possible situations:

(1) The bottom theory gives the more accurate description of the target system.
(2) The top theory gives the more accurate description of the target system.
(3) The two theories are equally empirically adequate.

It is worth noting that, while (1) is often taken to be the default option, this need not be so. As we will see in the example of Section 3.3, if we are interested in describing massless particles, then $T_{\rm t}$ is the more accurate theory, and the role of $T_{\rm b}$ is largely

[42] By 'description', I do not here mean to talk about the 'truth' of the idealising theories in a deep sense nor, for the purposes of this Section, about their reference.

explanatory and unifying, thus a case of (2). Which of (1)-(3) is the more appropriate option is of course an empirical question.

This discussion sheds light on the nature of ontological emergence: ontological emergence is simply a matter of *comparing physical situations or systems*, whether real or fictitious, at our world or at another physically possible world, characterising the same target situation or system (although we may compare the target system under different conditions[43]).

3.2.2 Ontology and metaphysics

In this Section, I want to answer the following question: In what sense is ontological emergence *ontological*?

As I said in the Introduction, I construe 'ontology' here in the straightforward sense of 'the ontology of a scientific theory': which, crucially, I take to be more than mere semantics, since the theories in question are subject to the requirement of empirical adequacy: so in particular, they must describe the same target system (cf. Section 3.2.1). We are interested in physical, and not merely mathematical, possibilities. Thus I construe ontology, in a somewhat limited sense, as the subfield of metaphysics that is concerned with what is: but not in Quine's narrow sense of 'to be' as in 'to be the value of a bound variable'. My sense of 'ontology', on which I will expand further below, comprises both what Quine (1951: pp. 14-15) dubs the 'ontology' of a theory (the domain of objects that the theory quantifies over) and its 'ideology' (i.e. roughly: the meaning of the theory's non-logical vocabulary, here understood as elements and relations in the domain of application, which are part of the world).

A first comment to make is that the worlds that I have been describing are *not* to be (naively, and wrongly!) identified with *the world as it is in itself*—whatever that Kantian phrase might be taken to mean.

Second, there is an interesting metaphysical project of (a) studying *how* the entities postulated by scientific theories can be, even if (b) we have not yet decided whether and how they *exist*, i.e. even if we have not yet decided how the idealised systems characterise the target system. That is, we can assess the question of the being and properties of those entities as idealisations, according to the theory, before we ask about their existence at our actual world. The latter question, (b), of course requires commitment to a specific metaphysical position: the realist will say that those entities exist (approximately) as postulated by the theory, while the non-realist may say that these entities are fictions, or reduce to combinations of human perceptions.[44] Also, realists may disagree amongst themselves about a metaphysical construal of those entities: think, for example, of Quine's (1960: §12) *referential indeterminacy*, according to which the linguist, upon hearing the native utter the word 'gavagai' while pointing at a rabbit, might translate it as 'rabbit'—while still being at a loss whether to construe the objects as rabbits, or stages, i.e. brief temporal parts of rabbits, or mereological fusions of spatial parts of rabbits. My construal

[43] For example, the condition can be to take different values of the temperature: see the example of spontaneous magnetisation in ferromagnetism, in De Haro (2019: pp. 35-38).

[44] Recall, from the preamble of this Chapter, that to develop an account of emergence (and for the purposes of subsequent Chapters) I do not need to assume a scientific realist position.

of 'ontology' is limited in this sense, that 'rabbit' refers to the same element in the domain of application as 'gavagai', i.e. they refer to the same object even before we ask what further metaphysical categories the native appeals to in their understanding of 'gavagai'.

Another example of (b) is as follows: the quidditist will construe the novelty of the domains as a difference in the domains' *properties*: and the comparison will be based on the possibility of the primitive identity of fundamental qualities (for example, natural properties: cf. Lewis (1983: pp. 355-358)), across domains or worlds—irrespective of how the quidditist construes individuals in his or her ontology. On the other hand, the haecceitist will construe ontological novelty as a difference in the domains' objects, or individuals, irrespectively of how he or she construes properties in their ontology.[45]

While these differences could, to a large extent, be already built into the interpretations, I have chosen to keep things general and not give a specific metaphysical construal of the domains (other than saying that they are sets with elements and relations, and that they can be intensions or extensions).

Question (b), applied to our case of distinct domains of application, $D_{\mathrm{t}} \neq D_{\mathrm{b}}$, means that different metaphysicians can agree about the parts of D_{t} that are not in D_{b}, according to Eq. (3.1), since these are structured sets. But different metaphysicians will construe the differences differently—over and above the comparison given by the description of the top and bottom domains as structured sets.

But the ontological project in this thesis does not require that we answer question (b): for the realist and the non-realist alike must agree in constructing the ontology of the theory, before this ontology can be, for example, declared real or reduced to perceptions. For example, realists and constructive empiricists can agree about the interpretation of a theory, even if they have different degrees of belief in the entities that it postulates (van Fraassen (1980: pp. 14, 43, 57)). The idea here is that working out the ontologies of scientific theories, the way they are interconnected, and their logical structure, is a different project from explicating how the elements of that ontology actually exist.[46]

My position here bears some resemblance with what A. Fine (1984: p. 96-98) has called the 'core position' that realists and non-realists share: both accept the results of scientific investigations as, in some sense, 'true', even if they give a different analysis of the notion of truth.[47] It corresponds to the contrast that Schaffer (2009: p. 352), Corkum (2008: p. 76) and K. Fine (2012: pp. 40-41) have made, in the context of scientific theories, between

[45] See Black (2000: p. 92).

[46] An anonymous reviewer objected that the distinction between epistemic and ontological emergence is unimportant for a theory of emergence, because it depends on whether one adopts a realist or, for example, an empiricist or a conventionalist framework. However, I am arguing that the distinction *is* important, even before we address the debate between realism and its rivals. For even the empiricist has to admit that any interpreted scientific theory assumes certain ontological facts about the world: even if a world only of phenomena, to which the empiricist is minimally ontologically committed. For example, the empiricist assumption that all that our theories describe are regularities rests on some ontological assumptions: viz. that there are regularities in the world for our scientific theories to describe, and that those regularities do *not* point to any deeper ontological structure. Furthermore, Ladyman has argued that van Fraassen's belief in the empirical adequacy, rather than the truth, of theories requires an objective modal distinction between the observable and the unobservable: so that the empiricist cannot avoid modal talk. For a summary of that debate, see Ladyman and Ross (2007: pp. 99-100).

[47] Fine's position is criticised by Musgrave (1989) and Butterfield (1988). But remember: my account of emergence is independent of scientific realism.

a neo-Aristotelian metaphysical project (of enquiry into *how things are*) vs. a Quinean project of strict enquiry into *what exists* (in Quine's very narrow sense as domain of quantification), according to our best theory, appropriately regimented.[48] The former project permissively allows for things, and categories, to appear in our ontology, that we might one day come to reject as literal parts of our world. Those things are, in some sense: even if they do not exist in the literal sense in which the theory says they do (for example, for the reasons given in Section 3.2.1: the systems described by the theory are idealisations of the target system). Thus my position, which is based on a limited but straightforward reading of the ontology of a scientific theory, is closer to the 'jungle landscapes and coral reefs' (Richardson (2007)) of the neo-Aristotelian project than to the desert ecosystems that Quine's advocacy of first-order logic suggests.

3.3 Case Study: Masslessness in Relativistic Mechanics

In recent theoretical physics there is a pervasive idea that, for theories with external parameters, and taking some of those parameters to special values, one gets to a "critical, or massless, regime" of the theory. The idea is that, as the physical length scale of the massive theory disappears, the theory is in a scale-invariant, or massless, regime. This is the idea of conformal fixed points in both statistical mechanics and in quantum field theories. Thus emergence is often associated with the appearance of such a massless or scale-invariant regime within a massive theory: for it is commonly agreed that there is an important sense in which the massless properties that arise in the special regime—such as, for example, symmetries—are emergent.

In this Section, I will illustrate some of these ideas in a very simple model: namely, the emergence of a massless free classical point-particle, within the theory of relativistic classical mechanics for massive particles. It will concern the emergence of properties within a unified theoretical framework.[49] Since the example is simple, a short account can be detailed: thus exhibiting all the necessary ingredients of my account of ontological emergence.

For simplicity of the presentation, I will take the bare theory to be just the equations of motion, derived from a Lagrangian. This is justified for our purpose in this Section because, in a classical theory of the kinds I will discuss, the equations of motion allow us to construct all the states and quantities of the theory.

[48] I endorse this contrast, even if it may not be entirely faithful to the historical Quine and, especially, Aristotle (hence the word 'neo-Aristotelian').

[49] For an example of emergence between different theories, see De Haro (2019).

3.3.1 The bottom theory T_b for the point-particle

The equations of motion for a free classical point-particle of mass m in Minkowski space are our bottom theory, T_b. They are as follows:[50]

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(\frac{1}{e}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\right) &= 0 \qquad (3.6) \\ \eta_{\mu\nu}\,\frac{\mathrm{d}x^\mu}{\mathrm{d}\sigma}\,\frac{\mathrm{d}x^\nu}{\mathrm{d}\sigma}+e^2\,m^2 &= 0\;. \qquad (3.7) \end{aligned}$$

The variable $e(\sigma)$ is a real, positive-definite, Lagrange muliplier, introduced in order to deal with the massive and massless cases using a unified theoretical framework. Here, the Minkowski metric is $(\eta_{\mu\nu}) = \mathrm{diag}(-1,1,1,1)$, and σ is an affine parameter, viz. a monotonically increasing parameter, along the particle's world line. The equations of motion are invariant under reparametrisations of the affine parameter σ, which implies that $e(\sigma)$ can be transformed away (i.e. set equal to one, if $m \neq 0$) by a choice of the affine parameter, σ.

The equations of motion, Eqs. (3.6)-(3.7), actually form a one-parameter family of theories, where the parameter is the mass, m (though I will keep simply referring to Eqs. (3.6)-(3.7) as 'the theory'). Thus, we have a theory for each value of m, which I will denote as $T_\mathrm{b}(m)$, to indicate that the theory "fixes the mass". Accordingly, I take the mass to also be in the domain of the interpretation map, which is evaluated on $T_\mathrm{b}(m)$.

In the massive case ($m \neq 0$), it follows from Eq. (3.7) that the proper time is $\mathrm{d}\tau = mc\,e(\sigma)\,\mathrm{d}\sigma$, so that Eq. (3.7) simplifies to the condition that the velocity 4-vector is time-like.

When $m = 0$, Eq. (3.7) is the condition that the 4-velocity vector lies on the light-cone.[51]

3.3.2 Emergence as non-commutativity of linkage and interpretation

In this Section, I discuss the emergence diagram, in Figure 3.2, for the theory of the massive particle, Eqs. (3.6)-(3.7). I will do this by constructing the analogue of Eq. (3.5), i.e. the lack of commutativity, for a theory with a continuous parameter.

The point about emergence is this: as long as m is non-zero, the interpretation of the equations of motion, Eqs. (3.6)-(3.7), is the familiar one. Namely, $x^\mu(\tau)$ gets interpreted as 'the position of a massive particle, with mass m [some number fixed by the theory], moving freely in Minkowski space, as a function of the particle's proper time'. The interpretation is *the same* for all nonzero values of m, i.e. it is the same map: for, as we saw in Section

[50] These equations can be derived from a Lagrangian, which is discussed in detail in De Haro (2019b: §3.1).

[51] One may wonder whether it makes sense to take the $m \to 0$ limit for a *single,* free particle. Regardless of one's answer to this: the simple point-particle model considered here of course does not stand on its own, but is to be coupled to further physics, which will determine the scale relative to which the mass is taken to zero. In Section 3.3.3, we will consider collisions of two particles.

3.1.5, the interpretation map depends on m through the theory's dependence on it (here, m plays the role of the parameter x, in Eqs. (3.3)-(3.5)).

In the $m = 0$ case, the interpretation of $x^\mu(\sigma)$ is: the position of a massless particle moving at the speed of light in Minkowski space, as a function of the affine parameter of its worldline.

The main difference between the two interpretations is in how 'massive' and 'massless particle' are construed. I will discuss this in more detail in Section 3.3.3.

Here is what it means for the diagram in Figure 3.2, according to the interpretation in Eq. (3.5) for a one-parameter family of theories, to describe emergence: the linkage map is the process of taking the limit $m \to 0$. So, we start with the bottom theory, for any non-zero m:

$$T_{\rm b}(m) = \text{`the theory (Eqs. (3.6)-(3.7)) for the particle of mass } m\text{'} \ , \tag{3.8}$$

where such qualifications as 'free point-particle' are, for simplicity of presentation, already included in what 'the theory' is, namely in Eqs. (3.6)-(3.7). Taking the limit $m \to 0$ as the linkage map, we end up with the top theory:

$$T_{\rm t} := \text{link}\,(T_{\rm b}) = \lim_{m\to 0} T_{\rm b}(m) = \text{`the theory (Eqs. (3.6)-(3.7)) for the particle of mass 0'}. \tag{3.9}$$

The limit is smooth, and it is represented by the leftmost arrow in Figure 3.3.

Emergence occurs because *what is described*, in the limit, is different—indeed novel, as represented by the rightmost arrow in Figure 3.3. We have all the tools in place; so that this is now straightforward to show.

First, the interpretation of the bottom theory, Eq. (3.8), has as its range (in a simplified form that suffices for our discussion):

$$D_{\rm b} := i_{\rm b}(T_{\rm b}(m)) = \{\text{a free, massive point-particle of mass } m\} \ . \tag{3.10}$$

This reflects the fact that, as I mentioned in Section 3.3.1, the interpretation is a function of the mass, and the domain of application may contain particles of any mass.

The interpretation of the top theory, Eq. (3.9), has as its range:

$$D_{\rm t} := i_{\rm t}(T_{\rm t}) = \{\text{a free, massless point-particle}\} \overset{\text{Eq. (3.9)}}{=} i_{\rm t}\,(\text{link}\,(T_{\rm b})) \ . \tag{3.11}$$

But if we go along the lower horizontal arrow, $i_{\rm b}$, in Figure 3.3, and then take the value $m = 0$ in Eq. (3.10), we do *not* end up at $D_{\rm t}$ (i.e. Eq. (3.11)). For the range of the interpretation $i_{\rm b}$ is incorrect for a massless particle, namely:

$$D_{\rm b}\,|_{m=0} = i_{\rm b}(T_{\rm b}(m))|_{m=0} \overset{\text{Eq. (3.10)}}{=} \{\text{a free, massive point-particle of mass 0}\} \ . \tag{3.12}$$

This is different from the correct (because more accurate) interpretation: namely, Eq. (3.11).

Thus the two interpretation maps, $i_{\rm t}$ and $i_{\rm b}$, map to different domains, even though both have $m = 0$. The underlying point is that, as I will argue in Section 3.3.3:

$$\{\text{a free, massless point-particle}\} \neq \{\text{a free, massive point-particle of mass 0}\} \ , \tag{3.13}$$

$$\begin{tikzcd}
T_{\mathrm{t}} = \text{Eq.}\,(9)(m = 0) \arrow[r, "i_{\mathrm{t}}"] & D_{\mathrm{t}} = \text{massless particle} \\
T_{\mathrm{b}}(m) = \text{Eq.}\,(8) \arrow[u, "\begin{array}{c}\text{link}\\ m \to 0\end{array}"] \arrow[r, "i_{\mathrm{b}}"] & D_{\mathrm{b}} = \text{massive particle of mass } m \arrow[u, "\text{Emergence}"', "(\neq \text{massive particle of mass } 0)"' near end]
\end{tikzcd}$$

Figure 3.3: Emergence of the massless particle, analysed as the lack of commutativity between the linkage map, $m \to 0$, and the interpretation: $i_{\mathrm{t}} \circ \text{link} \neq i_{\mathrm{b}}$.

i.e. the massless particle is different from the massive particle 'with the mass set to zero' (and the former is also not contained in the latter as a proper subset). Thus, we have proven the conditions, Eqs. (3.1) and (3.2), that the linkage and interpretation maps do not commute. The interpretation of T_{t} as describing a *massless* particle is a novel interpretation, relative to the massive particle interpretation. This is why there is ontological emergence.

Notice that there is no requirement that the top and bottom domains, D_{t} and D_{b}, must "greatly differ", or that their differences must be striking, in order for there to be emergence. There is simply emergence if they differ. But agreed: the more they differ, the greater the novelty.

3.3.3 Massive vs. massless particles

In this Section, I argue that classical massive and massless particles are qualitatively different (thus substantiating Eq. (3.13)). Indeed, although they are both described by the equations of motion, Eqs. (3.6)-(3.7), their properties are distinct. I will discuss three related properties that differ between the massless and massive cases and that, together, characterise a massless particle:

(i) *Timelike vs. null geodesics.* The first difference is in the geodesics, namely in the time-like condition vs. the null condition, Eq. (3.7). In terms of the four-momentum vector, $p := m\,\dot{x}$, these conditions are: $p^2 = -m^2c^2$ in the massive case, vs. $p^2 = 0$ in the massless case.

In the limit $m \to 0$, the massive equation, $p^2 = -m^2c^2$, of course reproduces the massless equation, $p^2 = 0$. And thus the timelike geodesics converge to null geodesics. Yet this reduction does not prevent emergence, for it is reduction in the bare theory and not in the domain of application. Namely, the property of 'being a timelike geodesic' still figures in the interpretation of the bottom massive theory, even though we set $m = 0$. Thus the interpretation does not fit with the idea of a 'massless particle', in other words: the range of the interpretation is empty if the interpretation is an extension,

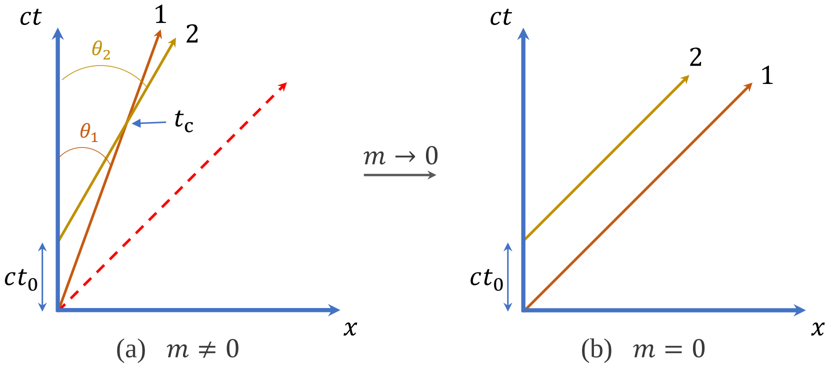


Figure 3.4: (a) Two particles with constant speeds, $v_1/c = \tan\theta_1$ and $v_2/c = \tan\theta_2$, colliding at $t = t_\text{c}$. (b) Massless limit, $v_1 = v_2 = c$: the particles *do not* collide at any finite time. As $m \to 0$, the lines in (a) are continuously deformed to the lines in (b) (but the *crossing point* is pushed to infinity!).

i.e. $i_\text{b}(T_\text{b}(m))|_{m=0} = \emptyset$.[52] One might at first sight think that this is only a matter of words, since it is clear from a Minkowski diagram that the $m \to 0$ limit of a timelike geodesic is a null geodesic. But it is, in fact, not a matter of words: for there are also consequences for the interpretation of the massive theory, of sending $m \to 0$, that plainly *contradict* the interpretation of the massless theory.

One such consequence is the absence of a rest frame for a massless particle, while there always exists one for a massive particle.

Another interesting consequence regards the *occurrence of events,* e.g. collisions. Consider two particles with mass m, travelling at different speeds in the same forward direction, but separated by an initial interval of time t_0, i.e. particle 2 is sent from the same location, but at a time t_0 later than, particle 1: see Figure 3.4(a). Take the second particle to travel faster than the first, so that they are due to collide when the second overtakes the first. But if we take the massless limit, these two particles will *never* collide, for they will both travel at the speed of light, and they will keep their mutual initial separation (see Figure 3.4(b)). In other words, an event that is possible, according to i_b, at non-zero m under the given initial conditions (namely, 'the particles will collide within finite time') becomes impossible, according to i_t, in the limit $m \to 0$, under the corresponding initial conditions (namely, 'the particles will not collide within any finite time').

To illustrate this distinction by a model, consider again Figure 3.4(a). The particles move along straight lines, at constant speeds v_1 and v_2, respectively, where $0 < v_1 < v_2 < c$. Their trajectories are given, respectively, by:

$$\begin{aligned} x_1(t) &= v_1\, t \\ x_2(t) &= v_2\,(t - t_0)\ . \end{aligned} \tag{3.14}$$

The time, t_c, at which the two particles collide is calculated by setting: $x_1(t_\text{c}) = x_2(t_\text{c})$.

[52] This concurs with Norton's (2012: pp. 208-210; 2016: §3.1) idea that limits of theories sometimes do not describe physical systems, real or fictitious, because they ascribe to them *contradictory properties.* More details follow below, in an example of a collision.

The result is:

$$0 \; < \quad t_{\rm c} = \frac{c}{v_2 - v_1}\, t_0 \quad < \; \infty \; . \tag{3.15}$$

By taking the $m \to 0$ limit of the trajectories, one finds that the two particles travel at the speed of light, i.e. $v_i \to c$ $(i = 1, 2)$, as in Figure 3.4(b): namely, the two particles' trajectories are taken to lie on separate, right-moving, null lines. The time of collision, Eq. (3.15), gets "pushed to infinity", i.e. $t_{\rm c} \to \infty$: for, in the limit, the speed of the first particle approaches the speed of the second particle from below, and they both travel at the speed of light (so that the condition $v_1 < v_2$ is *violated!*).[53]

There is an insightful way to understand why the maps fail to commute. Namely, we use the notion of a model of a theory, as being part of the interpretation map: we construct our interpretation map of $T_{\rm b}$ by constructing models like Eqs. (3.14)-(3.15) and Figure 3.4(a). The point, then, is that these massive models instantiate concepts such as 'timelike geodesic', 'particle's rest frame', and 'finite collision time': but the limit $m \to 0$, Figure 3.4(b), does not instantiate any of these notions. Thus there are *no models* of $T_{\rm b}$ of this kind in the limit.

For example, in the particle collision model it is true that, for any $m \neq 0$, the collision time is finite, as stipulated by Eq. (3.15). But the limit $m \to 0$ *violates* this condition, because it requires that the two particles travel at the same speed, viz. $v_1 \to v_2^-$, so that $t_{\rm c} \to \infty$: and so, there is in the limit *no model* of $T_{\rm b}$ with a collision of this kind.

Notice that, while this is closely related to Norton's idea of 'ascribing contradictory properties to the limit system', in this case there *is* a limit system, namely the one that can be constructed by interpreting the massless theory directly, through $i_{\rm t}$ (cf. Eq. (3.11)): so that we have a genuine case of idealisation, along the lines discussed in Section 3.1.3. This gives rise to a limit system that is arbitrarily close to the massive one (cf. Eq. (3.10)) in some aspects, but not in its novel aspects: and it is *not* obtained by taking the limit of the massive system.

Summarising, the interpretation of the bottom theory, $i_{\rm b}$, assigns contradictory properties to the particle in the $m \to 0$ limit: for it still contains concepts that it inherits from the massive case and which are contradicted by the massless limit, and which no longer refer. This is why, to get the top theory's interpretation, $i_{\rm t}$, one needs to go back to the $m \to 0$ limit of the bare theory, i.e. go back to $T_{\rm t}$, and build (at least some aspects of) its interpretation again. The interpretation of the limit theory is not the limit of $i_{\rm b}$.

(ii) *Symmetries of the solutions.* The states of the massive and the massless particles have different symmetries. To show this, we consider a generic solution of the equations of motion, Eqs. (3.6)-(3.7), and construct the subgroup of the Poincaré group leaving this solution invariant. Wigner (1939: p. 50) dubbed this the 'little group'.

First consider a massive particle, with given four-momentum vector $p = m\,\dot{x}$, obtained by solving the equations of motion, Eqs. (3.6)-(3.7). To find the subgroup of the Lorentz

[53] Malament (1986: pp. 192-193) discusses a somewhat similar case, namely the geometric (Cartan) version of Newtonian gravitation, as the $c \to \infty$ limit of general relativity. In this limit, the light-cones "open up" and become flat, thus violating the four-dimensional general covariance of the theory.

transformations leaving this vector invariant, go to the particle's rest frame, where $(p^\mu) = mc\,(1, \mathbf{0})$. The subgroup of the Lorentz group leaving this vector invariant consists of the spatial rotations, plus in addition time reversal symmetry. This group is isomorphic to O(3).

In the massless case, there is of course no rest frame for the particle: but we can orient our coordinate system in such a way that the symmetries are easy to see. The following is a useful choice: $(p^\mu) = p\,(1, 0, 0, 1)$, where p is the particle's momentum. The group preserving this vector is the group of Euclidean motions of the plane, E(2) ≅ ISO(2) (see e.g. Gilmore (2008: §13.3)). It contains three generators: one rotational (corresponding to rotations around the x^3-axis) and two translational, along the x^1 and x^2-directions.[54] Therefore, the little group is isomorphic to E(2).

Notice that E(2) is *not* a subgroup of O(3): it has different generators, and so the symmetries of the two cases are different, which expresses the *novelty* of the symmetry group in the massless case, compared to the massive case.[55]

Recall, from Section 3.2.1, the three possible cases, (1)-(3), for comparison between the idealising and the target systems. In so far as the massless theory is taken to give a description of photons or other massless particles, we have a case of (2), i.e. the top theory gives a good description of massless particles, while the bottom theory does not. The usefulness of emergence is then that it exposes the interpretative differences between the massless and the massive cases, while it gives a unified description of the two cases through the linkage relation.[56]

3.4 Emergence: Discussion and Further Work

This Chapter has analysed ontological emergence as novelty of reference, in terms of two conjuncts: a formal, i.e. not interpretative, condition—linkage—and an interpretative condition—ontological novelty, which is a difference in intension, and sometimes also extension. One typical case of emergence is when the theories have the same extension but different intensions. I illustrated the idea of ontological emergence as the non-commuting diagram in Figure 3.2, namely the failure of the interpretation to commute with the linkage map, Eq. (3.2).

I illustrated ontological emergence in a simple example, of the emergence of masslessness (or of a massless regime) for classical point-particles. We saw that the interpretation of the massless particle theory is different from the interpretation of the massive particle theory with the mass set to zero. Namely, there are differences between their causal prop-

[54] For a discussion of their physical interpretation, see Han et al. (1981: §II).

[55] O(3) and E(2) are related by group contraction (see Inönü and Wigner (1953), Gilmore (2008: §13.3.1)), which has a geometric interpretation in terms of "projection to a flat surface, around the North pole of the sphere" associated with O(3), i.e. a "large radius" limit. See Kim (2001: §4).

[56] One might also want to apply the massive theory to describe massive, but very light, particles. Pitts (2011: p. 277) argues that the massless case is empirically distinguishable from any *particular* massive theory, but only by observations probing length scales of the order of the inverse mass. Pitts goes on to argue that 'there exists a range of sufficiently small photon masses such that the massive... theories are empirically indistinguishable from the massless theory'.

erties and symmetries, as well as in their predictions about whether certain collisions can take place.

My account of novelty is logically weak in that *any difference* in the description (epistemic novelty) or in the domain of application (ontological novelty) counts as novelty. This is as it should be: for 'novelty' does not, by itself, carry a connotation of scientific or philosophical importance. And it is also *metaphysically* weak. For I have focussed on the basic question of when we are entitled to claim that ontological emergence obtains—a claim on which I have argued, in Section 3.2, that any interpretative strategy must agree, except for scepticism and radical forms of instrumentalism: which will dismiss ontology from the outset (although it is hard to do so consistently!). Further work, on the metaphysics of emergence, will have to further characterise the kinds of ontological novelty which arise in each case—namely properties, individuals, causal powers, etc.

Unlike other accounts, which seek the mark of emergence in some technical property of bare theories (see (I) below), my explication formalises a conception of emergence that is intrinsically *interpretative* (which should not be confused with: subjective, or arbitrary!). The mark of novelty is the non-meshing of the linkage map with the interpretation. The main idea illustrated in Section 3.3 was that the interpretation of the bottom theory differs from the interpretation of the top theory. Typically, the linkage map gave rise to an interpretation with properties that do not fit the top theory: for example, as a 'massive particle with mass 0', which differs from a genuine 'massless particle' interpretation.

I make two further comments characterising the properties of the emergence relation which we have obtained:

(I) *"Non-singular"*. We saw that the limit in Section 3.3 is not singular, in the mathematical sense: *pace* various philosophers' emphases.[57] There are no singularities, in the mathematical sense, in that case: for the limit is smooth.

(II) *"Top-accuracy"*. A related aspect of emergence that the example in Section 3.3 (and other examples treated in De Haro (2019)) illustrate is that the bottom theory does not always give a better (in the sense of: more accurate, both quantitatively as well as conceptually) description of the target system than the top theory that is being compared to. This is because, for any non-zero value of the mass, the bottom theory's interpretation does not give the correct properties of a massless particle, in particular: its geometric properties and its symmetries.

Further work on emergence is as follows. The relation and contrast between ontological emergence and duality is discussed in detail in Dieks et al. (2015), De Haro (2017), and De Haro, Mayerson, Butterfield (2016: pp. 1417-1421), especially for gauge-gravity dualities. The relation between duality, fundamentality, and emergence is discussed, in various examples, in Castellani and De Haro (2020). The examples discussed in this paper are cases of epistemic emergence.

De Haro (2019: pp. 35-38) discusses the relation of the present proposal for emergence to other accounts in the literature, including Franklin and Knox (2018) and Crowther

[57] Singular limits are essential for e.g. Rueger's (2000: p. 308) and Batterman's (2002: pp. 80-81, pp. 1424-126) senses of emergence. On the other hand, Butterfield (2011a: pp. 1073-1075) has argued that singular limits are not essential to have emergence.

(2016). It also develops further the topic of idealisations and approximations, from Sections 3.1.2-3.1.3, and discusses spontaneous magnetisation. Phonons are described as a case of epistemic emergence in De Haro (2019: pp. 27-34). The emergence of masslessness, including a discussion of the ultrarelativistic limit of solutions in general relativity, is further developed in De Haro (2020d).

For the emergence of spacetime in the context of holographic scenarios, see Dieks et al. (2015). The emergence of space in a random matrix model is treated in De Haro (2020e: pp. 45-48). Emergence and correspondence for stringy black holes are discussed in De Haro, van Dongen et al. (2020) and van Dongen et al. (2020).

Part II. Further Development

Chapter 4

On Theoretical Equivalence and Duality

> *A science can never determine its subject-matter except up to an isomorphic representation* (Hermann Weyl, 1934).

The simple conception of scientific theories and models, from Chapter 1, is a good framework to analyse both dualities and theoretical equivalence in physics. The conception of a duality will be simple: *a duality is an isomorphism of models of a single bare theory* (Section 4.1). Then, in Section 4.2, I will discuss theoretical equivalence, which has a formal condition (which I will dub 'weak theoretical equivalence') and an interpretative condition. I will argue that dualities are specific cases of theoretical equivalence. Then, in Section 4.3, I will discuss the conditions under which cases of duality give cases of theoretical (or, in the physical sciences, 'physical') equivalence.

As I mentioned in the previous Chapter, the rest of this thesis does not assume a scientific realist position, except when explicitly discussed (viz. in Section 4.2.4).

4.1 A Schema for Dualities

In this Section and the next, I develop the treatment of duality that forms what I will dub the 'Schema' for dualities, and which allows us to analyse the question of the theoretical equivalence of duals.

4.1.1 Duality's scientific importance

Recall from Chapter 1, especially 1.2.1 and 1.2.2, my overall proposal. A bare theory can be realised, or modelled, in various ways, like the different representations of an abstract algebra. These models are in general *not* isomorphic, since they differ from one another in their specific structure. But when they are isomorphic, we have a duality.

To develop this proposal, I begin with four clarifying remarks. Each remark leads in to the next. The first three defend my taking duality as a notion that is both logically weak and independent of a theory's interpretation. The first is, in effect, just the point that 'duality' is a term of art; so one can choose how to use it. But the second and third are substantive—about the scientific importance of dualities. The fourth remark is a contrast with the notion of *gauge*.

(1): *A logically weak but physically strong definition*:— I agree that at first sight, it looks profligate to say that there is duality whenever two models are isomorphic. For it means there are countless dualities. For example: if a group or an algebra, endowed with a set of rules for evaluating quantities and a dynamics, can be a bare theory, any two isomorphic representations will yield a duality, as long as the isomorphism preserves the values of the quantities and the dynamics. Accordingly, the notion of duality is sometimes narrowed by adding physical conditions, e.g. by requiring that dual models describe 'a single quantum system that has two classical limits' (Polchinski (2017: p. 7)).[1]

But I will maintain in (2) and (3) below that it is best to leave 'duality' broadly defined, as I have done: with such extra conditions being articulated in individual cases as the need arises. The strengthening will be given by the kind of physical degrees of freedom that one wishes to describe.[2] And so, my notion of duality will be physically strong. In particular, it cannot be argued that two given models which share some structure are dual, unless the common structure is exactly equal to what the models regard as physical. In short: this apparently profligate verdict can be accepted.

(2): *Duality as surprising*:— So far we have said what a duality is, but not how surprising and fruitful it can be. The examples in Sections 5.2.1, 5.2.2, and 5.2.3 will of course bring out these issues. For example, in some cases the isomorphism links the weak and strong coupling regimes of the two models, so that calculations that are difficult to do in one theory are easy to do in the other. For the moment, I note three clarifying comments: each comment leads in to the next.

(i): We usually discover a duality in the context, not of a bare theory, but of an interpreted theory (cf. §1.2.3 and 1.2.4); for of course we work with interpreted theories.[3]

(ii): Indeed, we usually work with what I have called 'a model of the theory', indeed an interpreted model. That is: usually, before the duality is discovered, we have two interpreted models (usually called 'physical theories'!) which we do not believe to be isomorphic in any relevant sense.

(iii): Usually, we do not initially believe the two models are models of any single relevant theory (even of a bare one: i.e. even if we let ourselves completely suspend our antecedent interpretation of the models). The surprise is to discover that they are such models—indeed are isomorphic ones.

The word 'relevant' in (ii) and (iii) signals the fact that of course 'isomorphism', 'model' and 'theory' are very flexible words. For example: almost any two items can be

[1] Such a notion would clearly be too restrictive, since very well-known dualities such as Kramers-Wannier duality (1941) would already not qualify as dualities by this criterion.

[2] See De Haro and Butterfield (2018: pp. 339-341).

[3] Agreed, pure mathematicians sometimes work with uninterpreted theories; and duality is a grand theme in mathematics, just as it is in physics. But although comparing duality in mathematics and in physics would be a very worthwhile project, I set it aside. Cf. Corfield (2017).

considered isomorphic, i.e. as having a common structure, under a weak enough construal of 'structure'. Thus physicists might well in some specific context notice that the two models in question are both groups, or both algebras. But they rightly do not announce this as discovering a duality: not even if they also notice that the two groups (or algebras) are isomorphic. They set it aside as irrelevant, since the abstract notion of group or of algebra is so general that having it identified as a bare theory in common between the models is scientifically useless.

On the contrary, what is surprising, and scientifically valuable, is to find very specific structures in common between different models: especially when

(a) the models as presented (so: as interpreted) are very disparate, and-or

(b) the common structure is not only detailed (like '10-dimensional semisimple Lie group', as against 'group') but amounts to an isomorphism of that detailed structure (like 'isomorphic as 10-dimensional semisimple Lie groups').

(c) the common structure is rich and general (like 'a two-conformal field theory with conformal charge c at level k', as against 'a specific solution of the theory').

As noted above, what will give physical theories their specificity, thus making duality a more powerful tool than its logically weak definition might make it seem, is the fact that physical theories, even bare ones, come with sets of maps from groups and algebras to appropriate fields (in the mathematical, not physical, sense!), i.e. maps that assign values to the physical quantities. These maps are defined at the level of the abstract structure, but must also be instantiated in each of the models (according to the relevant sense of instantiation, as either 'representation' or 'realization': cf. Section 1.2). And this set of maps is usually so rich, that it often suffices to reconstruct a model. And so, the fact that duality preserves these maps can be very non-trivial, and surprising, especially when combined with (a)-(c) above.

This discussion of (a)-(c) returns us to (1) above. I doubt that there can be a general characterization of when the models as presented are disparate enough, and-or the discovered isomorphism is rich and detailed enough, for scientific importance. Instead, one can only articulate in any specific case how the disparity and-or the details *are* enough. So it is not worth trying to tighten the *definition* of 'duality' with conditions beyond the logically weak ones I advocate. One just needs to use one's judgment about which cases count as scientifically important enough to analyse.

(3): *Examples*:— The conclusion of (2) is supported by some famous examples of duality in physics. It is worth illustrating this with two other examples.

(A): Gauge-gravity duality. In this case, the models differ in the dimensions they assign to spacetime (d in the gravity model, $d-1$ in the gauge model), in their field content and classical equations of motion, and in much more. In this case, the common core consists only in a class of asymptotic operators and a conformal class of $(d-1)$-dimensional metrics. Of course, it is very surprising to learn that a gauge theory model in $d-1$ dimensions, and a model of quantum gravity in d dimensions, despite their very disparate guises, nevertheless have the same common core, and represent the same theory. See De Haro (2020) for a discussion in the context of this Schema.

(B): Electric-magnetic, or S-duality. This relates two models by mapping the electric charges of one model to the magnetic charges of the other. Furthermore, it does so by mapping a small electric charge to a large magnetic charge. Nevertheless, the common structure is the same in the two models, i.e. the quantum theory is invariant under the replacement of one gauge group by its dual.

(4): *A contrast with 'gauge'*:— This discussion of dualities' scientific importance brings out a contrast between my treatment of duality, and the notion of gauge. Physicists sometimes make remarks like: 'two dual theories are like different gauge formulations of a single theory'. I agree that this remark is *analogous* to my view: indeed, in two ways.

(i): A gauge formulation of a theory has specific structure (viz. the gauge variables) going beyond that mandated by the ideas (gauge-invariant ideas!) of the theory; just like for us, a model has specific structure going beyond that mandated by the bare theory.

(ii): The idea of gauge as 'descriptive redundancy' means that two gauge formulations of a single theory must 'say the same thing'; just like we say that in a duality, two models are isomorphic, and so (if interpreted: could) 'say the same thing'.

But I submit that this is *only* an analogy. There are two differences. First, we want to allow for cases where the two duals are not physically equivalent (as in Kramers-Wannier duality, mentioned in footnote 1): *pace* the suggestion in (ii). Second (and more importantly), the extra structure in a model is usually *not* gauge, i.e. descriptively redundant: think of how the extra structure in a representation of a group usually carries physical information (e.g. a representation of the Poincare group carrying mass and spin information). Again, as stressed in (2) above: the surprising and scientifically important discovery is that in two models, with apparently very disparate structures, there is in fact an exact correspondence of structures.

4.1.2 Duality as isomorphism

I turn in this Subsection to my proposal that a duality is an isomorphism of models of a bare theory. To be precise: on the account, in Section 1.2.2, of theories and models as triples, a duality is *an isomorphism of model triples* of a bare theory. Indeed, after all the stage-setting in Chapter 1, the proposal is straightforward. I first give its details, using the notations we have established.

The basic idea is that a duality is an isomorphism between two triples, each comprising a state-space, a set of quantities, and a dynamics. 'Isomorphism between triples' is of course short for a triple of maps: an isomorphism between the two state-spaces, and isomorphism between the sets (almost always: algebras, cf. footnote 5) of quantities, and an equivariance condition on the dynamics.[4] In addition, the isomorphism must commute with the symmetries of the theory.

More important is the question of *which* kinds of triples are related by duality. Recalling the distinction between bare theories and their more specific models, the answer is clear: *a duality relates two model triples of a single bare theory.*

[4] My proposal does not depend on the formulation of models as triples. A model root can be presented in many different forms, and the isomorphism should then preserve the corresponding structure. Even for triples, one can envisage isomorphisms which do not respect the triple structure, though they map the model roots isomorphically. Cf. De Haro and Butterfield (2018: pp. 339-341). But it will suffice for our purposes to restrict to model roots defined as triples, whose structure is preserved by the duality.

The crucial point here is that the model triple is separated from the model's own specific structure, and expresses only the model's realizing (typically: representing in the mathematical sense) the bare theory. Recall the notation from Eq. (1.1): $M = \langle \mathcal{S}_M, \mathcal{Q}_M, \mathcal{D}_M, \bar{M} \rangle =: \langle m, \bar{M} \rangle$, where $m := T_M := \langle \mathcal{S}_M, \mathcal{Q}_M, \mathcal{D}_M \rangle$ is the model triple. Model triples, as representations of a bare theory, are in general of course *not* isomorphic to each other, nor to the bare theory. So the assertion of duality is substantive: it asserts that two model triples are in fact isomorphic.

But this is *not* to say that the two *models*, each 'considered in their entirety', are isomorphic. They each have their own specific structure, and are (in almost all cases) *not* isomorphic. Recall my other notation (Eq. (1.2)) for models 'considered in their entirety': $M = \langle \bar{\mathcal{S}}, \bar{\mathcal{Q}}, \bar{\mathcal{D}} \rangle$. Indeed, their being non-isomorphic is usually part of what makes the duality surprising and (if Nature is kind to us!) empirically fruitful, i.e. of scientific importance. And 'the more non-isomorphic'—i.e. the more disparate the two models, considered in their entirety, are—the more surprising, and (one hopes) empirically fruitful, is the duality (cf. (2) and (3) in Section 4.1.1).

I now introduce some notation for dualities as isomorphisms between model triples. This will require first giving:

(1) some new notation for the value of a quantity on a state, and

(2) a more detailed discussion of dynamics (in both the 'Schrödinger' and 'Heisenberg' pictures).

Both (1) and (2) can be given wholly independently of my distinctions (i) between theories and their models, and (ii) between interpreted and uninterpreted theories. So for the moment, please consider a generic triple of a state-space, a set of quantities, and a dynamics: $\langle \mathcal{S}, \mathcal{Q}, \mathcal{D} \rangle$.

(1): Suppose we are given a set of states $\mathcal{S}$, a set of quantities $\mathcal{Q}$ and a dynamics $\mathcal{D}$: $\langle \mathcal{S}, \mathcal{Q}, \mathcal{D} \rangle$. I will write $\langle Q, s \rangle$ for the value of quantity Q in state s. This prompts two further general points.

First: it is common to think of a state $s \in \mathcal{S}$ as a maximal specification of the instantaneous properties of the system in question; and a quantity $Q \in \mathcal{Q}$ as a numerically measurable property of it. In effect, this makes states and quantities nothing but assignments of values to each other. Second: for classical physics, one naturally takes quantities as real-valued functions on states, so that $\langle Q, s \rangle := Q(s) \in \mathbb{R}$ is the system's possessed or intrinsic value of the quantity; and for quantum physics, one naturally takes quantities as linear operators on a Hilbert space of states, so that $\langle Q, s \rangle := \langle s|\hat{Q}|s \rangle \in \mathbb{R}$ is the system's Born-rule expectation value of the quantity. But for quantum physics it is often important to consider the non-diagonal matrix elements of a given quantity/operator $\hat{Q}$, without requiring this to be adequately encoded in the Born-rule expectation values of various other quantities. So for a quantum theory we should understand a value written schematically as $\langle Q, s \rangle$ to also represent all the matrix elements $\langle s_1|\hat{Q}|s_2 \rangle$. Thus $\langle Q, s \rangle$ is a short-hand for an expression like $\langle Q; s_1, s_2 \rangle := \langle s_1|\hat{Q}|s_2 \rangle$, i.e. Q is regarded as a map: $\mathcal{S} \times \mathcal{S} \to \mathbb{C}$.

(2) I turn to the dynamics $\mathcal{D}$, i.e. a specification of how the values of quantities change over time. I will keep the discussion very simple. First, I assume the dynamics is de-

terministic: also in quantum theories, despite the threat of Schrödinger's cat. Then it can be presented in two ways, for which we adopt the quantum terminology, viz. the 'Schrödinger' and 'Heisenberg' pictures. But I shall not need to distinguish otherwise between the different detailed formalisms for dynamics, such as Hamiltonian vs. Lagrangian, and the path-integral. Besides, I will adopt for simplicity the Schrödinger picture.

So we say: D_S is an action of the real line $\mathbb{R}$ representing time on $\mathcal{S}$. There is an equivalent Heisenberg picture of dynamics with D_H, an action of $\mathbb{R}$ representing time on $\mathcal{Q}$. The pictures are related by, in an obvious notation:

$$D_S : \mathbb{R}\times\mathcal{S} \ni (t,s) \mapsto D_S(t,s) =: s(t) \in \mathcal{S} \text{ and } D_H : \mathbb{R}\times\mathcal{Q} \ni (t,Q) \mapsto D_H(t,Q) =: Q(t) \in \mathcal{S} \tag{4.1}$$

where for all $s \in \mathcal{S}$ considered as the initial state, and all quantities $Q \in \mathcal{Q}$, the values of physical quantities at the later time t agree in the two pictures:

$$\langle Q, s(t)\rangle = \langle Q(t), s\rangle \ . \tag{4.2}$$

With the notations and notions of remarks (1) and (2) in hand, I can now present the notation for dualities as isomorphisms between model triples. Let M_1, M_2 be two models, with model triples $m_1 = \langle \mathcal{S}_{M_1}, \mathcal{Q}_{M_1}, \mathcal{D}_{M_1}\rangle$ and $m_2 = \langle \mathcal{S}_{M_2}, \mathcal{Q}_{M_2}, \mathcal{D}_{M_2}\rangle$. We can suppose that M_1, M_2 are both models of a bare theory T. Or we can proceed in the 'opposite direction': that is, we can suppose that M_1, M_2 are given independently of a bare theory T, but their model triples (model roots in the more general language of Section 1.2.2) are isomorphic. Either way, the notation for dualities is as follows.

To say that the model triples m_1, m_2 are isomorphic is to say, in short, that: there are isomorphisms between their respective state-spaces and sets of quantities, that (i) make values match, and (ii) are equivariant for the two triples' dynamics (in the Schrödinger and Heisenberg pictures, respectively). I now spell this out. Though retaining the Ms in the subscripts is cumbersome, I will do so, in order to emphasise my main conceptual point: that duality is a relation between model triples in my sense—it is *not* between theories, or between generic triples $\langle \mathcal{S}, \mathcal{Q}, \mathcal{D}\rangle$ as in remarks (1) and (2).

Thus we say:— A duality between $m_1 = \langle \mathcal{S}_{M_1}, \mathcal{Q}_{M_1}, \mathcal{D}_{M_1}\rangle$ and $m_2 = \langle \mathcal{S}_{M_2}, \mathcal{Q}_{M_2}, \mathcal{D}_{M_2}\rangle$ requires:

an isomorphism between the state-spaces (almost always: Hilbert spaces, or for classical theories, manifolds):

$$d_s : \mathcal{S}_{M_1} \to \mathcal{S}_{M_2} \ \text{ using } d \text{ for 'duality' ;} \tag{4.3}$$

and an isomorphism between the sets (almost always: algebras) of quantities

$$d_q : \mathcal{Q}_{M_1} \to \mathcal{Q}_{M_2} \ \text{ using } d \text{ for 'duality' ;} \tag{4.4}$$

such that: (i) the values of quantities match:

$$\langle Q_1, s_1\rangle_1 = \langle d_q(Q_1), d_s(s_1)\rangle_2 \ , \ \ \forall Q_1 \in \mathcal{Q}_{M_1}, s_1 \in \mathcal{S}_{M_1}. \tag{4.5}$$

and: (ii) d_s is equivariant for the two triples' dynamics, $D_{S:1}, D_{S:2}$, in the Schrödinger picture; and d_q is equivariant for the two triples' dynamics, $D_{H:1}, D_{H:2}$, in the Heisenberg picture: see Figure 4.1.

$$\begin{tikzcd}
\mathcal{S}_{M_1} \arrow[r, "d_s"] \arrow[d, "D_{S:1}"] & \mathcal{S}_{M_2} \arrow[d, "D_{S:2}"] \\
\mathcal{S}_{M_1} \arrow[r, "d_s"] & \mathcal{S}_{M_2}
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathcal{Q}_{M_1} \arrow[r, "d_q"] \arrow[d, "D_{H:1}"] & \mathcal{Q}_{M_2} \arrow[d, "D_{H:2}"] \\
\mathcal{Q}_{M_1} \arrow[r, "d_q"] & \mathcal{Q}_{M_2}
\end{tikzcd}$$

Figure 4.1: Equivariance of duality and dynamics, for states and quantities.

Eq. (4.5) appears to favour m_1 over m_2; but in fact does not, thanks to the maps d being bijections.

4.1.3 Duality and interpretation

So far, the discussion of interpretation has concerned a *single* theory or model. Thus recall that Sections 1.2.3-1.2.4 introduced interpretation maps i_{Int} and i_{Ext} in a rather informal way, as mapping from a bare i.e. uninterpreted theory or a bare model, to the realm of intension (which I will label 'Sinn'), or to the realm of extension ('Bed'), respectively. But again, everything in Section 1.2.4 concerned a *single* theory or model.

Since duality is about relations between theories/models, there is, at first sight, little to say about duality and interpretation. That is: interpretation should simply proceed independently on the two sides of the duality—for example, we just require the interpretation-symmetry commuting diagram on both sides of the duality. Indeed: I said already in Section 4.1.1 that in some cases of duality, the two sides were clearly not—nor intended to be—physically or semantically equivalent: e.g. the high and low temperature regimes in Kramers-Wannier duality. And my definition of duality as formal (viz. an isomorphism of model triples) certainly allows this idea of 'distinct but isomorphic sectors of reality'—namely as the codomains of the interpretation maps on the two sides of the duality.

This verdict—'there is little to say'—is true, so far as it goes. And of course, it does not forbid the other sort of case: where the two sides of the duality *are* physically/semantically equivalent, i.e. do describe 'the same sector of reality'. In the Schema, this would be modelled by the interpretation maps on the two sides having the same images/values in their codomain—so as to give a *triangular*, rather than *square*, commuting diagram. I shall spell this out as regards the interpretation of (bare) quantities: similar diagrams could of course be drawn for states.

For (bare) quantities being mapped by i_{Int} into the realm of intension 'Sinn', the two sides of a duality describing 'the same sector of reality' amounts to the diagram in Figure 4.2.

Similarly: for (bare) quantities being mapped by i_{Ext} into the realm of extension 'Bed'—relative to some given possible world W with a context rich enough to determine references, of course—the two sides of a duality describing 'the same sector of reality' amounts to Figure 4.3.

So far, so straightforward. But the above verdict is a bit quick: there are two further

$$\begin{tikzcd}
\mathcal{Q}_{M_1} \arrow[rr, "d_q"] \arrow[dr, "i_{\text{Int}}"'] & & \mathcal{Q}_{M_2} \arrow[dl, "i_{\text{Int}}"] \\
& \mathcal{Q}_{\text{Sinn}} &
\end{tikzcd}$$

Figure 4.2: The two sides of the duality describe 'the same sector of reality', in the realm of intension.

$$\begin{tikzcd}
\mathcal{Q}_{M_1} \arrow[rr, "d_q"] \arrow[dr, "i_{\text{Ext}}"'] & & \mathcal{Q}_{M_2} \arrow[dl, "i_{\text{Ext}}"] \\
& \mathcal{Q}_{\text{Bed}} &
\end{tikzcd}$$

Figure 4.3: The two sides of the duality describe 'the same sector of reality', in the realm of extension.

points to make.

(1): *What determines equivalence?*:— First, there is the question what determines whether the two sides of a duality are physically/semantically equivalent, i.e. describe the same 'sector of reality'. Dieks et al. (2015) and De Haro (2020) have argued that the choice between these two options should, at least in part, depend on whether the models in question are given an internal or an external interpretation (see Section 1.2.4). The idea is that for the ranges of the interpretation maps to be distinct, there must be other facts, external to the triples themselves and our use of them, that determine the distinct ranges. Typically, these facts will be other pieces of physics to which the system described by each model triple is coupled—with different pieces of physics on the two sides of the duality. This coupling 'breaks the symmetry' between the two sides, and secures that the two model triples are about distinct, albeit isomorphic, subject matters ('sectors of reality'). In the proposed jargon: the coupling provides an 'external interpretation' of the model triple. On the other hand: sometimes we propose a physical theory as a putative theory of the whole universe, i.e. as a putative cosmology, so that according to the theory there are no physical facts beyond those about the system (viz. universe) it describes. If in such a case, there is a duality—which means there are two isomorphic model triples, each putatively describing the whole universe—then there can be no such coupling to other pieces of physics. (Gauge-gravity duality provides, of course, a putative example of such a duality between theories of the universe.) An interpretation of each triple must therefore be an internal interpretation; and this prompts the conclusion that the two triples can describe the very same 'sector of reality'. That is: the interpretation maps can have the same range; and there can be a triangular diagram, as in Figures 4.2 and 4.3.

(2): *Interpreting the specific structure*:— Second, there is more to say about the interpretation of a model's specific structure, especially in the latter sort of case, i.e. two sides of a duality describing the same 'sector of reality'.

Recall that a model M is more than the model triple m, by which it realizes a bare theory, and which relates to another model triple in a duality. M also has a specific structure $\bar{M}$: as I stressed at the start of Section 4.1.2, this structure is *not* related by the duality to 'the other side'. But the specific structure $\bar{M}$ *does* get interpreted—it supplies arguments for the interpretation maps i_{Int} and i_{Ext}—just as much as the model triple m gets interpreted.[5] This was emphasised by the other notation for models introduced in Eq. (1.2): viz. a model is itself, like a bare theory, a triple. M's states and quantities, being specific to M, are in general 'bigger'/'more structured' than the states and quantities of the bare theory that M models/realizes. I wrote them with a 'bar': thus $M = \langle \bar{\mathcal{S}}, \bar{\mathcal{Q}}, \bar{\mathcal{D}} \rangle$.

So the first point to make is: my discussion of duality has so far ignored the specific structures $\bar{M}_1$ and $\bar{M}_2$ on the two sides, even though they do get interpreted. This silence is presumably no problem in a case where we agree that the two sides are not physically or semantically equivalent. In such a case, the interpretation of $\bar{M}_1$ and $\bar{M}_2$ just means there are physical facts on each of the two sides, additional to the facts that are isomorphic with (a subset of) facts on the other side.[6] Thus for the case of Kramers-Wannier duality, the obvious examples of such non-matching physical facts would be facts about the value of the temperature: high on one side, and low on the other. In short: these additional facts (in the realms of intension and extension, respectively) are: on one side, in the ranges $i_{\text{Int}}(\bar{M}_1)$ and $i_{\text{Ext}}(\bar{M}_1)$; and on the other side, in the ranges $i_{\text{Int}}(\bar{M}_2)$ and $i_{\text{Ext}}(\bar{M}_2)$.

But what about the other sort of case: where the two sides *do* describe 'the same sector of reality'? Is it really satisfactory to say that there are physical facts that:

(a) are additional to those facts described by the isomorphic model triples, i.e. those 'caught' by the duality/the common bare theory; yet also

(b) fall into two such disparate subsets: one subset expressed by $\bar{M}_1$ and the other subset expressed by $\bar{M}_2$?

In short: this world-picture, combining (i) a set of facts expressed by the two sides in the same way (though this sameness may be not obvious—the duality can be surprising), and (ii) two other sets of facts expressed in very different ways by the two sides, is surprising: and maybe it is odd, or unsatisfactory...

We saw examples of this in comment (3) of Section 4.1.1, for example in example (A), gauge-gravity duality. Here, the set of facts that are the common core, *à la* (i), consists only in a class of asymptotic operators and a conformal class of $(d-1)$-dimensional metrics. And the sets of facts *à la* (ii) include, on the bulk side, gravity (such as: Einstein's equations coupled to matter) in d dimensions expressed by $\bar{M}_1$, and on the boundary side, a conformal field theory (such as: the Yang-Mills equations) in $d-1$ dimensions expressed

[5] I am here discussing general interpretations, i.e. external ones (in the terminology of Section 1.2.4). Internal interpretations were defined as interpreting only model roots, and such internal interpretations will be important in Sections 4.2 and 4.3.

[6] In De Haro and Butterfield (2018: pp. 339-341), we emphasised the fact that only the model triples, and not the specific structure, are physically significant. When we now consider external interpretations that do give a physical meaning to the specific structure, we have to say that these interpretations change the physical content of the model (its physical degrees of freedom). This is correct, because external interpretations do not need to preserve the structure of the model as a quadruple.

by $\bar{M}_2$. See De Haro (2020: Section 2.1) for a discussion in the context of the schema.

4.2 Theoretical Equivalence and Duality

Recall that theoretical equivalence and duality are both taken to be, roughly, a matter of "two theories saying the same thing, in different words". Thus the question arises how these two notions are related: and, in particular, whether discussions of duality bear on discussions of theoretical equivalence. This Section aims to address this question. I will argue that the here presented Schema for duality leads to a proposal for theoretical equivalence that, furthermore, suggests interpretative constraints on the use of theoretical equivalence in the physical sciences. (Sections 4.2.2-4.2.3 are about weak theoretical equivalence).

4.2.1 Theoretical equivalence

Theoretical equivalence is an old and venerable topic in the philosophy of science, which goes back to the logical positivists: and it finds its roots even earlier, for example in the philosophy of spacetime (viz. in the Leibniz-Clarke correspondence, and in Poincaré's conventionalism). Two influential accounts are Quine (1975) and Glymour (1970). The recent discussion of theoretical equivalence, with which I will engage, includes: Halvorson (2012, 2013), Glymour (2013), van Fraassen (2014), Coffey (2014), Weatherall (2015, 2016, 2016a), Barrett and Halvorson (2016), Lutz (2017), Teh and Tsementzis (2017), Barrett (2018), and Hudetz (2018).

Theoretical equivalence is, in fact, something of a term of art, without a fixed meaning. It usually combines two conditions:

(A) some sort of "formal or mathematical requirement"; and
(B) some sort of "interpretative requirement".

Individual authors may of course stress one aspect over the other, or even altogether reject one of the two conditions.[7]

Although the recent literature on dualities has of course engaged with the broad philosophical discussion of theoretical equivalence,[8] it has for the most part not attempted to answer in detail the question of how duality *bears* on the notion of theoretical equivalence. This Section aims to fill this gap, by developing a duality-inspired criterion of theoretical equivalence, which I will dub *physical equivalence*.

Recent discussions of theoretical equivalence have focussed on an important question: *what is the best formal account of equivalence between physical theories?* In an influential

[7] For example, Coffey (2014: p. 837) proposes an account in which questions of theoretical equivalence ultimately reduce to questions of interpretation, and formal considerations are only relevant in virtue of how they shape interpretative judgments.

[8] Earlier work relating duality and the recent literature on theoretical equivalence is in De Haro (2020: Section 1.4), Butterfield (2018), Read and Møller-Nielsen (2018: Section 4.1). See also Weatherall (2019).

series of papers, Weatherall (2015, 2016, 2016a) has argued that categorical equivalence provides a good standard of theoretical equivalence for examples in physics. He also proposes an interesting category-theoretic criterion for "when a theory has excess structure": namely, when an appropriate functor between the two empirically equivalent theories "forgets structure".[9] He has argued that categorical equivalence gives correct verdicts in a number of important cases: Newtonian gravitation vis-à-vis Newton-Cartan theory, various versions of electromagnetism, Yang-Mills theory, etc. On the other hand, Barrett and Halvorson (2016: p. 556), and Hudetz (2018: §2.1), have argued that categorical equivalence is "too liberal", in that it rules as equivalent theories that we would not expect to count as equivalent. Thus it is worth exploring other formal conceptions of theoretical equivalence in connection with Weatherall's examples, to see whether they make the same judgments.

But the search for a formal account should not overshadow the importance of *interpretation*. I will argue that the formulation of the formal account is itself dependent on interpretation, since theoretical equivalence usually requires a suitable "translation". And so, I will argue that an account of theoretical equivalence requires having an account of semantic interpretation. I will argue that (i) the Schema's distinction between internal and external interpretations gives just such an account, and thus (ii) that one can draw interpretative lessons for theoretical equivalence more generally.

I will propose, roughly speaking, the following construals of two crucial terms:

(i) *Weak theoretical equivalence:* isomorphism of models (of a single theory) and matching of interpretations (where 'matching of interpretations' means that the domains of application used in the interpretations are isomorphic).
(ii) *Physical equivalence: this is the duality-based account of theoretical equivalence in the physical sciences:* sameness of the interpretations of weakly theoretically equivalent models (where 'sameness of the interpretations' means the lack of a difference between the elements and relations in the domains of application of the two models).[10]

Thus the Schema's notion of theoretical equivalence, namely physical equivalence, is an isomorphism criterion of equivalence. Such criteria have been criticised, in the context of the semantic view of theories, because they supposedly make distinctions without a difference—and this criticism is often accompanied by repudiation of the semantic view (see e.g. Halvorson (2012: p. 187-188)). These criticisms have motivated the search for other criteria of theoretical equivalence (see the authors cited at the beginning of this Section).

However, Lutz (2017: p. 335) has recently argued that the arguments against the isomorphism criterion can be blocked: and furthermore, that the recent syntax-semantics debate does really not capture any significant differences (ibid., p. 347). Likewise, Hudetz (2018: p. 18) has introduced a new criterion, definable categorical equivalence, which proposes to find a middle-ground between the category-theoretic approach and the earlier

[9] For related work, see Barrett and Halvorson (2016), Halvorson and Tsementzis (2015), Barrett (2018).

[10] 'Physical equivalence' is, in this Section, my duality-based proposal for *theoretical equivalence:* and so, it is not restricted to physics, and I will use the two phrases interchangeably. But in Section 4.3, the argumentation will be mostly restricted to the physical sciences.

definability-theoretic approach exemplified by Glymour (1970, 1977) and Quine (1975).

In De Haro (2019: Section 2.1.4), I argued that some of the criticisms of the isomorphism criterion have as their target only 'naïve' notions of isomorphism, and do not impugn more sophisticated notions such as the one provided by the Schema. And so, some of the motivations recently adduced for abandoning the isomorphism criterion were found to be wanting.

I do not mean to be dogmatic about the Schema's formal adequacy for dealing with all possible theories. There may be more generally valid, or more precise, mathematical conceptions of theoretical equivalence than the one I will work out based on the Schema. But it is worth exploring how far the Schema can go, and we will see that it fares very well: for it is able to give undoubtedly reasonable judgments about the case study presented in De Haro (2019) (Maxwell's electromagnetic theory) and it also gives correct judgments about previous case studies in De Haro and Butterfield (2018) (bosonization and gauge-gravity duality).

I am however uncompromising about the conceptual and methodological importance of semantic interpretation in any account of theoretical equivalence: a feature endorsed by the Schema. Errors in the formalism are generally easy to spot, while interpretative oversights often remain hidden in the background—and they do bear on judgments of theoretical equivalence. And so, I think that one important way in which the literature about dualities can contribute to general discussions of theoretical equivalence is through the philosophical analyses of the interpretation of duals that it has developed in recent years.

4.2.2 Locating the Schema's contribution to an analysis of theoretical equivalence

This Section and the next contain the thesis' main proposal for theoretical equivalence, based on the Schema. I first give some background on the notion of theoretical equivalence in recent philosophy of science discussions, in terms of two conditions.

Recall, from Section 4.2.1, the two conditions, (A) and (B), for theoretical equivalence: there is a formal and an interpretative requirement.

Two influential proposals for theoretical equivalence are due to Quine and Glymour. On Quine's (1975: p. 320) proposal, two formulations count as formulations of the same theory if, besides being empirically equivalent, the two formulations can be rendered identical by a *reconstrual of the predicates* in one of them (for example, by switching some of the predicates). And Glymour (1970: p. 279) requires, besides empirical equivalence, *inter-translatability* as a necessary condition for theoretical equivalence.[11]

[11] Thus Quine and Glymour agree that empirical equivalence is a necessary condition for theoretical equivalence, but construe the formal requirement (reconstrual of predicates vs. inter-translatability) differently. And both do seem to have in mind a notion of theoretical equivalence in which both theories in some sense "say the same thing about the world", i.e. a notion that is also interpretative. However, one should note that Quine's and Glymour's interpretative attitudes are quite different, because Quine was sceptical about meaning and ontology (see Quine's (1960: Section 2) discussion of referential indeterminacy) while Glymour (1977: p. 228) is not so sceptical.

About *empirical equivalence:* there are two main construals of the requirement of this criterion, one of which is syntactic, and the other semantic. Quine (1970: p. 179, 1975: p. 319) says[12] that two theories are empirically equivalent if they imply the *same observational sentences,* also called 'observational conditionals', for all possible observations—present, past, and future. Van Fraassen (1980: p. 64) says that two theories, T and T' are empirically equivalent if for every model (not the Schema's model!), M, of T there is a model, M', of T' such that all of M's *empirical substructures are isomorphic* to empirical substructures of M', and vice-versa.[13] The criterion of theoretical equivalence that I will give in Section 4.2.4 entails empirical equivalence as a special case (for a detailed discussion, see De Haro (2020b)).

The recent discussion of theoretical equivalence by Halvorson (2012), Barrett and Halvorson (2016), Halvorson and Tsementzis (2015), Barrett (2018), and Weatherall (2015, 2016, 2016a) has emphasised the formal aspects of equivalence over the interpretative. Weatherall mentions empirical equivalence in all of his papers on the topic, but for example Barrett and Halvorson (2016) do not mention it at all[14]—nor do those papers mention matters of meaning or interpretation (other than interpretation using formulas in a formal language). Thus one could easily get the impression that some of these authors "are not interested in interpretation". However, the situation is more subtle.[15]

I will endorse this recent (though usually unstated) consensus, that there is an interesting part of the project of theoretical equivalence that is a *largely, but not solely, formal matter.*[16] Thus I will first develop, in Section 4.2.3, a notion of 'weak theoretical equivalence' (I will discuss physical equivalence in Section 4.2.4). Indeed I believe (contra Coffey (2014)) that there is an interesting philosophical project of:

(EE) *Explicating equivalence in formal terms.*

This consensus position in the recent discussion of theoretical equivalence can be qualified as a *quietist* position. Namely, it is the position of authors engaged with parts of the project of theoretical equivalence, and for whom interpretative equivalence is a minimal requirement that their project needs, but on which they do not wish to focus. And so, such authors, when confronted with dual models, discuss formal equivalence but typically refrain from discussing ontological questions. Quietism is the position that my account

[12] Glymour (1970 : p. 277) holds a similar view.

[13] For a comparison between Quine's syntactic and van Fraassen's semantic conceptions of empirical equivalence, see De Haro (2020a: Section 4.1).

[14] Barrett (2018) and Rosenstock et al. (2015) also mention empirical equivalence.

[15] In conversation, Weatherall and Barrett say both that empirical equivalence is a prerequisite for their analyses of theoretical equivalence, and that a suitable notion of theoretical equivalence should (at least) respect (some) meanings, i.e. it is constrained by the interpretation. However, the latter requirement is not mentioned in their papers, and it is not clear what guiding role it plays. This contrasts with the older literature on the topic. For Glymour's (2013: p. 289) and van Fraassen's (2014: pp. 278-281) comments about interpretation already go some way towards the kind of interpretative project I have in mind here. Their comments of course go back to their own accounts of theories and of theoretical equivalence, in Glymour (1970, 1977) and van Fraassen (1970).

[16] Butterfield's (2018: Section 1.1) *Remark*, namely, 'in physics, two theories can be dual, and accordingly get called 'the same theory', though we interpret them as disagreeing', emphasises the distinction between the formal and the interpretative aspects. In the context of theoretical equivalence, Butterfield's *Remark* has also been voiced by van Fraassen (2014: pp. 278-279); and, in the context of duality, by De Haro (2020: §1.3-§1.4, §2.2).

of weak theoretical equivalence attempts to articulate, by including an interpretative requirement that is as minimal as possible—namely, mere matching of interpretations.[17] The account can be suitably strengthened in the second step, as I will do in Section 4.2.4. Such authors have nothing to say about physical equivalence—for whatever reason: perhaps because they are sceptical about the notion, or not interested in it, or think it is too difficult to articulate. This is, to a large extent, a tenable position:[18] and in a moment I will add some of my own reasons why (EE) is an interesting philosophical project.

In light of the literature discussed above, the challenge in the first step of defining weak theoretical equivalence lies, for the most part, in how one fills in its being 'a largely, but not solely, formal project'. That is, the difficulty lies in striking an appropriate balance between the interpretative and the formal: in such a way that, once one takes the interpretative conditions to have been established for a pair of theories, as a kind of "boundary condition", then the project can concentrate on the formal issues.

In the next Section I will propose how the Schema suggests this can be done. But it will clarify that proposal if I first give three reasons why I think 'weak theoretical equivalence', in contrast with 'theoretical i.e. physical equivalence', should be construed as being *mostly* formal or mathematical:

(i) The notion of weak theoretical equivalence will easily lead in to a notion of theoretical i.e. physical equivalence, which can then be further analysed.

(ii) The second reason for letting weak theoretical equivalence be formal and not require sameness of interpretation (in the sense of the interpretation map's having the *same range*, i.e. what in Section 4.1.3 I called describing 'the same sector of reality') is that establishing sameness of interpretation is never a simple matter: for it requires a notion of identity of domains of application, as I will discuss in Section 4.2.4, which leads in to issues of metaphysics. And we do not want to be too worried about metaphysics in our project of theoretical equivalence.

I take this second reason to underlie the recent consensus mentioned above: that there is a significant conceptual and technical project of finding criteria for when two theories are equivalent, without the need for all the details about the interpretation to be fleshed out, nor to commit to a metaphysical account of how the terms of the theory refer. And such a project strikes me as sensible.

This then prompts me to break up the overall project of explicating equivalence between theories into two tasks. The first task, namely explicating *weak theoretical equivalence*, cares as little as possible about interpretation: and it corresponds to (EE) above. The second task, namely explicating *theoretical i.e. physical equivalence*, takes interpretation fully on board.

(iii) The third reason for keeping weak theoretical equivalence largely formal is that, as

[17] See also Roberts (2014: Section 5).

[18] The reason for the qualification 'to a large extent' is that it is not automatic that the various extant formal proposals—definitional equivalence, Quine equivalence, Morita equivalence, categorical equivalence—will give physically equivalent theories. However, in Section 4.2.4 I will argue that, given weak theoretical equivalence, one can construct interpretations that deliver physical equivalence.

the Schema for duality has shown, this is indeed also how physicists think about equivalence. Physicists are happy to say that the high-temperature Ising model is equivalent to a low-temperature one, without worrying about the ontological status of 'temperature' under such a duality map. It suffices that two variables that are interpreted as 'temperatures' are mapped to each other, and that the values (low vs. high) of 'temperature' are also mapped to each other. Even less do physicists worry whether they are e.g. realists or empiricists about temperature under this duality. And this is of course a legitimate practice. Indeed I of course maintain that the metaphysical questions are important; but they can be addressed in the next step, as part of the project of *theoretical i.e. physical* equivalence. Thus a project like (EE), that only takes interpretation minimally into account, agrees with a widespread, and justified, scientific practice.

To sum up: regardless of one's preferred use of 'theoretical equivalence': it is an interesting project to establish criteria of formal equivalence that minimize interpretative issues, and so avoid being committed to the fully-fledged project of showing that the theories have the same interpretation (for they may *not* have the same interpretation!), i.e. the same range of their interpretation maps: and this is what I shall call the project of weak theoretical equivalence.

But if one endorses this recent consensus—that weakly theoretically equivalent models are *in some (weak!) sense* interpretatively equivalent: but that establishing weak theoretical equivalence does not require one to have a complete understanding of the interpretation, so that the project is largely formal—then it is clear that achieving what the consensus aims for requires a delicate balance.

So, it is now crucial to fill in condition (B)—the interpretative requirement—in such a way that we can strike this balance. I will propose that the models must have *matching internal interpretations.*

4.2.3 Conditions (A') and (B'): isomorphism of model roots and matching internal interpretations

I first fill in condition (A) in Section 4.2.2: namely, the formal or mathematical requirement for weak theoretical equivalence. My proposal here is simple: namely, to read the Schema's notion of duality as a *formal condition of theoretical equivalence,* applicable to formal theories in the physical sciences:[19]

(A') **Isomorphism of model roots of a single bare theory.**

Recall that this was precisely the definition of duality, in Section 4.1.2. And it is indeed natural to regard duality as prompting the weak theoretical equivalence of two models (in my sense of 'model').

[19] Notice that this proposal can be discussed independently of whether one endorses my explication, (B'), of (B): the proposal could be combined with any other explication, (B").

Next I fill out the second condition for weak theoretical equivalence: namely, the interpretative requirement, as prompted by the Schema's analysis of duality.

We have seen that the spirit of weak theoretical equivalence is that it should capture when two interpreted models (in the sense of model roots, i.e. often called 'theories') are equivalent, both formally and interpretatively, without being committed to a strong notion of sameness of interpretation. This is the 'delicate balance' that I mentioned at the end Section 4.2.2.

My proposal, then, for filling in condition (B), namely filling in the "interpretative" part of the definition, is to require "matching internal interpretations". So, let me *first* define the notion of matching interpretations. This will then lead us to the restriction to *internal* interpretations (as defined in Section 1.2.4).

(B') **Part 1: Matching interpretations:** two models, M_1 and M_2, have matching interpretations when the ranges of their interpretation maps are isomorphic, i.e. when:

$$\mathrm{ran}\,(i_1) \cong \mathrm{ran}\,(i_2)~. \tag{4.6}$$

I use the word 'matching' here, in order to avoid the already over-used word 'equivalence'. Recall, from Section 1.2.4, that an interpretation is a structure-preserving partial map that maps a model to a domain of application. For the two models M_1 and M_2, we thus have maps $i_1 : M_1 \to D_1$ and $i_2 : M_2 \to D_2$. Here, the domains of application D_1 and D_2 are structured sets, and the interpretation maps are structure-preserving. Thus, the condition Eq. (4.6) is an isomorphism with respect to that structure, which is induced from the model's own structure.

For simplicity in the discussion in the rest of this Section, I will take the interpretations to be surjective maps, so that $\mathrm{ran}\,(i_1) = D_1$ and $\mathrm{ran}\,(i_2) = D_2$, i.e. the interpretation maps map "to the whole domain of application": every element in the domain of application is mapped to by at least some element of the model. (This restriction does not affect the content of the discussion but does simplify the notation.) Thus, we can now restate the condition Eq. (4.6) as follows: $D_1 \cong D_2$. Let us denote the corresponding isomorphism between domains of application by $\tilde{d}$, so that:

$$\tilde{d}:~ D_1 \to D_2~. \tag{4.7}$$

I shall call this map, induced from the duality and interpretation maps, the *induced duality map*.

Combining the duality map, d (defined in Section 4.1.2), with the two interpretation maps, we get the diagram in Figure 4.4, where the interpretation maps, i_1 and i_2, are evaluated on the model roots, m_1 and m_2 (see the discussion two paragraphs below). Here, d and $\tilde{d}$ are both isomorphisms, while the interpretation maps of course need not be. Figure 4.4 thus contrasts with the diagram in Figure 4.5, where the duality *commutes* with the interpretation—in the sense that the three maps d, i_1, i_2 form a commuting diagram. In Figure 4.4, the three maps d, i_1, i_2 do not form a commuting diagram. Rather, we need to co-vary the duality map, thus getting the induced duality map, $\tilde{d}$, in Eq. (4.7), as we change the interpretation.

$$\begin{tikzcd}
m_1 \arrow[r, leftrightarrow, "d"] \arrow[d, "i_1"] & m_2 \arrow[d, "i_2"] \\
D_1 \arrow[r, leftrightarrow, "\tilde{d}"] & D_2
\end{tikzcd}$$

Figure 4.4: Weak theoretical equivalence and the induced duality map, $\tilde{d}$, between domains of application.

The notion of co-variation of the interpretations with the duality map is natural for dualities.[20] For example, take Kramers-Wannier duality, i.e. the Ising model with high, respectively low, temperature as its two models. These two models are isomorphic, and furthermore we can also match their interpretations by everywhere replacing a lattice at high temperature with a lattice at low temperature. Thus the isomorphism, d, between the models induces an isomorphism, $\tilde{d}$, between the domains of application, that takes 'high temperature' to 'low temperature' and vice versa, while respecting the syntax. The induced duality map between the domains of application, thus construed, is analogous to translating one language into another. Of course, such a map does not in general *preserve* meanings—rather, it *maps* them into each other in a non-trivial manner!

Matching internal interpretations. Weak theoretical equivalence, as discussed here, is in the context of duality. That is: it is an isomorphism account of equivalence. Thus in particular, the diagram in Figure 4.4 implies that $i_2 \circ d = \tilde{d} \circ i_1$. Now, in order for i_2 to be defined on, or "match", the range of d (i.e. the model root, since the duality map only maps the model roots, see Section 4.1.2), i_2 must be restricted to the model root, and likewise for i_1 (as shown in Figure 4.4). This suggests that, for the purpose of theoretical equivalence between dual theories, we should restrict our interpretations to the *internal interpretations,* as I anticipated in Section 1.2.4. For such a commuting diagram does not exist for external interpretations, as I now show.

In principle, weak theoretical equivalence could also be defined for external interpretations. We would then replace the model roots m_1 and m_2 in Figure 4.4 by the full models, M_1 and M_2 (cf. Eq. (1.1)), and i_1 and i_2 would be their external interpretations. We would then still require the map, $\tilde{d}$, between the domains of application to be an isomorphism, as in Eq. (4.6). But notice that the diagram is no longer commuting: in general, M_1 and M_2 are not isomorphic and, if their cardinalities are unequal, they *cannot* be isomorphic.[21]

[20] Read and Møller-Nielsen (2018) actually define *theoretical equivalence* in this way, i.e. as empirical equivalence plus duality.

[21] If the full models, M_1 and M_2, including their specific structure, do happen to be isomorphic, then one could define weak theoretical equivalence for external interpretations as an isomorphism of the full models, i.e. $M_1 \cong M_2$. But then we are changing the notion of duality, which was tied to the notion of the common core theory, and we lose the connection between the common core theory and the duality. In such a case, we might say that the full models, $M_1 \cong M_2$, *are* the common core, and that consequently there is no specific structure—but then we are back to the original case in the main text, of two isomorphic model roots. And then the external interpretations simply collapse to internal interpretations. And so,

Thus, although weak theoretical equivalence *could* be defined for external interpretations as the isomorphism condition Eq. (4.6), there is no commuting diagram criterion for it, so that this sort of weak theoretical equivalence is unrelated to the existence of duality. Thus in this thesis we hold on to the connection between weak theoretical equivalence and duality, by restricting to internal interpretations.

To sum up: I believe that making (B) precise as (B'), i.e. matching of internal interpretations, gives a minimal requirement of interpretative equivalence—as required for the 'delicate balance' of the previous Section—that is useful both for dualities and for discussions of theoretical equivalence. In short: meanings are simply co-varied as we map one model to the other, without the requirement that the meanings stay "the same"—the latter requirement is only introduced in the next step. Since the recent literature on theoretical equivalence has, as I mentioned, not given a conception of (B), I have here endeavoured to give a minimal one, based on the Schema. In the next Section, I will refine this to a criterion of theoretical equivalence.

Notice the difference between duality and weak theoretical equivalence thus construed: namely, condition (B') is required for weak theoretical equivalence, but not for duality. We follow the physicists in defining duality as isomorphism of model roots—period. Weak theoretical equivalence, on the other hand, is constrained by the philosophical tradition to require, in addition, an interpretative condition that I have here interpreted weakly, in the sense of Eq. (4.6) and Figure 4.4. Thus weak theoretical equivalence is an equivalence—an isomorphism—of interpreted model roots, while duality is an isomorphism of model roots, irrespective of their interpretation.

4.2.4 Theoretical equivalence of duals

In this Section, I spell out the notion of theoretical equivalence, or—in the context of physics, especially as suggested by dualities—*physical* equivalence: and discuss the relation between weak theoretical equivalence and physical equivalence.

Recall that I have explicated the weak theoretical equivalence of models as the matching of their interpretations. I will now define physical equivalence as weak theoretical equivalence plus, in addition, *sameness* of interpretation, i.e. of the ranges of the interpretation maps: thus as sameness of the domains of application, D_1 and D_2. Thus the induced duality map, $\tilde{d}$, between the domains of application, is the identity map: so that we have the triangular commuting diagram between the duality and the two interpretation maps, see Figure 4.5. (This is of course a version of Figures 4.2 and 4.3, for internal interpretations.)

But recall that, after making explicit the unstated consensus of the project of theoretical equivalence in Section 4.2.2, I noticed that this consensus aimed to stay formal and to minimize matters of metaphysics. And I justified this consensus as sensible, by using a notion of interpretation that assigned a basic ontology to the models, but remained quiet

the only gain in defining duality as an isomorphism between full models M_1 and M_2 (in cases where such an isomorphism obtains), while still maintaining that the model roots are smaller than M_1 and M_2, is that the common core comes out to be smaller—thus, as far as I can see, no real gain. And so, the analysis in the main text is enough.

$$\begin{tikzcd}
m_1 \arrow[rr, leftrightarrow, "d"] \arrow[dr, "i_1"'] & & m_2 \arrow[dl, "i_2"] \\
& D &
\end{tikzcd}$$

Figure 4.5: Physical equivalence. The two interpretations describe "the same sector of reality", so that the ranges of the interpretations coincide.

about other metaphysical matters. And that quietism is, in its turn, justified by intensional semantics' appropriately modelling the interpretative practices of both realists and empiricists: as in the interpretation maps that I defined in Section 1.2.4. Namely, recall the discussion in that Section: realists and constructive empiricists can agree about the interpretation of a theory or model, in other words about its basic ontology ('the picture of the world drawn by the theory', to use van Fraassen's (1980: pp. 14, 43, 57) words), even though they have different degrees of belief in the entities that the ontology of the theory postulates.

I will now argue in a "contrary direction", since the project of theoretical equivalence that I have in mind goes beyond quietism. For example, realists may disagree amongst themselves about a metaphysical construal of the entities postulated by the theory: think, for example, of Quine's (1960: §12) referential indeterminacy (described e.g. in Section 3.2.2).

In other words, whatever the formally coinciding domains of application are, one has to impose on both sides the same ontological construal of terms, e.g. 'energy is a property and not a relation', etc.

Thus, if *numerical identity* of the elements and relations in the domains of application is what we want, we need an agreed philosophical conception of the ontology, i.e. of the domains of application: which means that we now need to make our metaphysical commitments explicit, e.g. about realism or empiricism and referential indeterminacy. This then amounts to a deeper explication of what we mean by an 'interpretation' than just saying that it is a map to items in the world. Only then does the requirement, that the induced duality map, $\tilde{d}$ (Eq. (4.7)), between the domains of application be the identity map, secure the same interpretations.

Thus I will say that the domains of application are (numerically) *the same* if we have an *agreed philosophical conception of the interpretation.*

In conclusion: *two models are physically equivalent iff (i) they are weakly theoretically equivalent, and (ii) they have the* same (internal) interpretation, i.e. their interpretation maps have the same range: where 'sameness' requires that an agreed philosophical conception of the interpretation has been supplied. This is the notion of theoretical equivalence that is suggested by dualities.

We can now be more precise about how weak theoretical equivalence can lead to physical equivalence, by comparing the diagrams in Figures 4.5 and 4.4. Evidently, we obtain the diagram for physical equivalence, Figure 4.5, if the induced duality map, $\tilde{d}$, in Figure

4.4 happens to be the identity map (and provided the above-mentioned philosophical explication is also given). But this is not the generic situation, i.e. $\tilde{d}$ is in general *not* the identity map, as we saw in the example of Kramers-Wannier duality: since it maps high temperature to low temperature, and vice-versa.

There is a second way to obtain a situation of theoretical, i.e. physical, equivalence from weak theoretical equivalence, i.e. Figure 4.4. The idea is to define a new pair of interpretations using the induced duality map, as follows: $i_1' := i_1$ and $i_2' := \tilde{d} \circ i_2$, where both models now have D_1 as their domain of application, i.e. $i_1'(m_1) = i_2'(m_2) = D_1$. This means that, in effect, we change m_2's domain of application, from D_2 to D_1. (Although this definition treats the two models asymmetrically, there is an alternative definition that maps to D_2 rather than D_1). While this may at first sight look ad hoc, it is in fact natural if i_2 is an *external* interpretation, while i_2' is now an *internal* interpretation:[22] and it is common practice among the physicists working on dualities.

Consider, for example, gauge-gravity duality, where m_1 is a gauge theory model and m_2 is a quantum gravity model (more precisely, a version of string theory), and take i_2 be the (external!) string theory interpretation of the quantum gravity model, while i_1 is a sophisticated, internal interpretation of the gauge theory model. What we then learn about the interpretation of the quantum gravity model is that an *internal* interpretation, i_2', of it can be constructed: and this is natural, given the existence of a duality. This "change of interpretation", from the external interpretation i_2 to the new internal interpretation i_2', in fact models the practices of the physicists who work on dualities: and the recent literature on dualities has tried to articulate in what sense such interpretations give natural interpretations of quantum gravity and quantum field theories. This is a large topic that I cannot discuss here, but it has been central to the recent literature on dualities: see, for example, Dieks et al. (2015), De Haro (2020, 2019), Huggett (2017), De Haro and Butterfield (2018), Read and Møller-Nielsen (2018).[23]

4.3 From Weak Theoretical Equivalence to Physical Equivalence

In this Section, I will go in more detail into the question of Section 4.2.4: *When does duality amount to physical equivalence?* More precisely, I will now ask whether we are *justified* in adopting an internal interpretation, hence judging two models to be physically

[22] If i_2 is external, then it maps the whole model, M_2, and not just the model triple m_2. In that case, the existence of an isomorphism $\tilde{d}$ between the domains of application is very non-trivial, because the cardinalities of the models considered might be no longer the same. However, it can be achieved if i_2 is defined on m_2 by, in effect, restricting to the model root, i.e. the external interpretation stripped of its "external connotations". (This can be done using the forgetful map from M_2 to m_2: we then have on m_2 the pullback of i_2 by the forgetful map).

[23] Recently, Weatherall (2019) has discussed a similar view in the context of classical electric-magnetic duality, where he introduces an 'empirical significance functor' that relates the empirical substructures of the domains of application of the dual models. Weatherall, too, discusses the empirical *inequivalence* of electric-magnetic duals on their ordinary (in my language, 'external') interpretations vs. their empirical *equivalence* once an empirical significance functor is introduced (in my language, on an 'internal' interpretation, restricted to the observational conditionals).

equivalent.

4.3.1 Having and losing external interpretations

In this subsection, I further develop the external and internal interpretations, and in particular two cases: (i) cases of external interpretations, in which *theoretical equivalence fails to obtain*, despite the presence of a duality; (ii) cases in which an external interpretation is not consistently available (where 'consistently' will be qualified below), so that one can only have an internal interpretation, and hence there is *theoretical i.e. physical equivalence.*

Let me illustrate the external interpretation with an elementary example that should make clear the difference between duality as a case of weak theoretical equivalence, and theoretical equivalence. Consider classical, one-dimensional harmonic oscillator "duality": an automorphism, $d : \mathcal{H} \to \mathcal{H}$, defined by $d : \mathcal{H} \ni (x, p) \mapsto (\frac{p}{m\omega}, -m\omega x)$, from one harmonic oscillator state to another, leaving the dynamics $\mathcal{D}$ invariant—namely, the Hamiltonian $H = \frac{p^2}{2m} + \frac{1}{2}kx^2$ and the equations of motion that H defines. So, it is an automorphism of $T_{\text{HO}} = \langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$.[24] But this automorphism of $\langle \mathcal{H}, \mathcal{Q}, \mathcal{D} \rangle$ does not imply a theoretical equivalence of the states:[25] the two states are clearly distinct and describe different physical situations: since the map d relates an oscillator in a certain state of position and momentum, to an oscillator in a *different* state.

This difference is shown in the fact that there is an independent way to measure the 'position' of the oscillator at a given time: one sets the oscillator and a standard rod side by side, observes where on the rod the oscillator is located, and so carries out a measurement of the former's position. We can picture this as coupling harmonic oscillator theory, T_{HO}, to our theory of measurement T_{meas}, and interpret the measurement as measurement of the oscillator position. Clearly, such an interpretation of T_{HO} is an example of an *external interpretation.* For it is obtained by inducing the interpretation of T_{HO} from an already interpreted theory T_{meas}, or by extension to $T_{\text{HO+meas}}$. And I call a theory, that can be coupled or extended in this way, an *extendable* theory.

But there are cases—such as cosmological models of the universe, and models of unification of the four forces of nature—in which these grounds for resisting the inference from duality to theoretical equivalence—a resistance based on the possibility of finding an external theory T_{meas}—are *lost.* For the string theories under examination—even if they are not *final* theories of the world (whatever that might mean!)—are presented as candidate descriptions of an *entire* (possible) physical world: let us call such a theory T. So, there is *no* independent theory of measurement T_{meas} to which T should, or could, be coupled, because T itself should be a closed theory (an *unextendable* theory: see §4.3.3). In the next subsection, I will use these ideas to spell out the conditions under which two dual models are physically equivalent.

[24] I have specified the states and dynamics of the harmonic oscillator: so I should here add that the quantities $\mathcal{Q}$ include e.g. any powers (and combinations of powers) of x and p.

[25] In a world consisting of a single harmonic oscillator *and nothing else*, the two situations could not be distinguished, and one might invoke Leibniz's principle to identify them. This amounts to adopting an internal interpretation, in the sense of Section 1.2.4. (For more on "a lonely oscillator" and on "and nothing else", see Section 4.3.2).

4.3.2 Internal interpretation and unextendability allow sameness of reference, and so physical equivalence

We return to dualities as isomorphisms of models; and so, we consider, specifically, the interpretation of two models, M_1 and M_2, "of the whole world". In this Section, I propose a condition that will, together with the internal interpretation, secure theoretical, i.e. physical, equivalence, in the sense that one is justified in taking duals to be physically equivalent: and I give the arguments to that effect.

The leading idea of an internal interpretation, as now further characterised in the context of *theory construction,* is that the interpretation has not been fixed a priori, but will be developed starting from the duality. Here, 'developed' does not mean 'logically entailed by the duality', since we are in the context of theory construction. Rather, it should be interpreted as "not mandated but conceptually suggested or natural". (Or, if by some historical accident, an interpretation has already been fixed, one should now be prepared to drop large parts of it.) The requirement that, I propose, justifies the use of the internal interpretation such that uniqueness of reference is secured, is as follows:

Unextendability: roughly, 'the interpretation cannot be changed by coupling the theory to something else or by extending its domain'. Unextendability replaces the somewhat vague phrase 'of the whole world' in the previous paragraph, and I will expound it in §4.3.3.

But even before that detailed exposition, one can argue that unextendability plays a key role in inferring physical, i.e. theoretical, equivalence. For it ensures that there is "no more to be described" in the physical world, and that the models cannot be distinguished, even if their domains of application were to be extended (since no such extension exists). And so, it ensures that the internal interpretation *can be trusted* as a criterion of theoretical i.e. physical equivalence (cf. §4.3.3), as I now argue.

Starting, then, from two such dual models, M_1 and M_2 of T, the duality map lays bare the invariant content $\langle\mathcal{H},\mathcal{Q},\mathcal{D}\rangle$, as that content which is common to M_1 and M_2, through the duality map (cf. Section 4.1.2). This is the starting point of the *internal interpretation*, for both the theory and the models. I now propose that an internal interpretation of a theory, satisfying the two conditions below, is the same for the two models (in the sense of §4.1.3), and in particular its reference is the same:

(i) the formalisms of the two models say the same thing: for they contain the same states, physical quantities and dynamics (i.e. the domain of the maps is identified by the isomorphism), and (ii) their physical content is also the same: for the interpretation given to the physical quantities and states is developed from the duality *and nothing else*: so, the codomains of the maps are the same, and they coincide with the entire world (thus assuming unextendability).

I am thus here proposing that *the domains of the worlds described by two dual models, and the worlds themselves are the same*: this is because, on an *internal* interpretation as further characterised here in the context of theory construction, the two worlds, and all their physical facts, are 'constructed, or obtained from', the triples.

There is a way in which this inference, from dual models with internal interpretations, to identical worlds, might fail: there might be more than one internal interpretation, and therefore more than one codomain D_W described by the theory. For in that case, despite the isomorphism of the two models, one might be tempted to think that one model could be better interpreted in one way, and the other better interpreted in another way.

But I take it that this is after all not an objection: the point is that, even if there is more than one internal interpretation, the reference of a given internal interpretation is the same for any two isomorphic models. Remember that, by definition, an internal interpretation cannot discern between models, because it starts from the theory, as a triple, *and nothing else*. So, there can be no reason for the interpretation to distinguish one model from the other—they both describe the world equally well and in the same way, according to that internal interpretation. In other words: *even if a single common core admitted several internal interpretations, each of them would refer to a single possible world, which would be the single reference of the corresponding internal interpretation of all the models isomorphic to the common core.*

Recall, from Section 1.3.2, the slogan 'extensions are determined by intensions, circumstances, and context'. The inference from an internal interpretation and unextendability to sameness of reference, in this Section and the next, can be seen in terms of this slogan (alternatively, Frege's "sense determines reference"). Namely, since the internal interpretation is developed from the duality and nothing else, it is an intension, i_{Sinn}. The unextendability condition then guarantees that two duals also have the same reference, i.e. i_{Bed} is determined by i_{Sinn}, given that we have an unextendable theory.

Let me spell out in more detail this inference from the isomorphism of models, to the identity of the internal interpretations and identity of the worlds described (hence theoretical equivalence), under the unextendability condition. There are two ways to make this inference. The first argument, from the unextendability of the models, will be given in §4.3.3: it is a version of Leibniz's principle of the identity of indiscernibles: which can be applied here, because the two models describe the entire world. The second argument, given in the previous paragraph, simply follows from the definition of the internal interpretation, in Section 1.2.4: an internal interpretation is constructed from the triple of the theory *and nothing else* (so, the specific structure of a model is not to find a counterpart in the world, since the interpretation must be invariant under the duality map). Thus, given an internal interpretation of the theory, the codomain of that interpretation, mapped from the two models, is the same by definition: *since the internal interpretation is insensitive to the differences in specific structure between the models, its reference must be the same* (thus again echoing Frege's slogan).[26] Explicitly, $i_1 = i_2 \circ d$, where $d : m_1 \to m_2$ is the duality map. Thus, such thorough-going dualities can be taken to give *theoretical i.e. physical equivalence* between apparently very different models.

It is important to note that this second argument for the identity of the codomains follows from the *definition* of an internal interpretation, given in Section 1.2.4 (together with the two stated conditions). What is surprising about a duality that is a theoretical equivalence, then, is not so much that two very different models describe *the same world*,

[26] The internal interpretation i is a partial surjective map from the theory to the world. But using the forgetful map from the model to the theory (the map which strips the model of its specific structure), we can construct an internal interpretation of the model, as the pullback of the interpretation map i by the forgetful map. It is in this sense that I here speak of internal interpretations of the models as well as of the theory.

but rather that *there is an internal interpretation* to be constructed from such minimal data as a triple:[27] and so, what is surprising is that there is a (rich) world for such a triple to describe! Admittedly: it would be hardly surprising if the internal interpretation described something as simple as the real line. But, in the examples we are concerned with here, the internal interpretation describes far richer worlds!

I do not claim to have established theoretical i.e. physical equivalence, under the conditions of internal interpretation and unextendability, as a matter of logical necessity, just from the notions of theory, model, and interpretation. Doing so would require a deeper analysis of the notion of reference itself, and the conditions under which it applies to scientific theories: which is beyond the scope of this thesis (Lewis (1983: pp. 361-365) contains a discussion of duplicates of worlds). What I claim to have argued is that, making some natural assumptions in particular about how terms refer in ordinary language and in scientific theories, the adoption of an internal interpretation and the assumption of unextendability does imply physical equivalence. This is also what I meant, in Section 4.2.4, by the need to have 'an agreed philosophical conception of the interpretation'.

4.3.3 Unextendability, in more detail

I turn to consider theories, such as theories that aim to unify the four forces, T_{QG} say, for which there is no extra physics to which T_{QG} can be coupled or extended. Being a description of the entire physical universe, or of an entire domain of physics,[28] I will take the interpretation i_{QG} to be *internal* to T_{QG}. Thus, as a sufficient condition for being justified in the use of an internal interpretation (in the sense meant at the beginning of Section 4.3), I have required, in §4.3.2, that T_{QG} be an unextendable theory. An interpretation map i_{QG} only requires the triple $T_{\text{QG}} = \langle \mathcal{H}, \mathcal{Q}, D \rangle$ as input, and it only involves the triple's elements and their relations—it does not involve coupling T_{QG} to other theories.

In such a case, duality preserves not only the formalism, but necessarily also the structure of the concepts of two complete and mathematically well-defined models: if one model is entirely self-consistent and describes all the relevant aspects of the world, then so must the other model. And so, duality becomes physical equivalence. Thus, in other words, we are really talking about different formulations of a *single theory*.

Let me spell out the (sufficient) condition, suggested by this discussion, for a theory to have an internal interpretation, such that the conclusion of physical equivalence of two sides (two dual models) is justified, in the sense of Section 4.3.2. A bare theory T in a domain D_W of a possible world W is *unextendable* iff three conditions hold:

(i) T is a complete theory in the domain of applicability D_W at W; and

(ii) There is no other theory T'' for the possible world W (or another possible world that includes it) and domain D_W, such that: for some T' isomorphic to T, $T' \subset T''$ (the meaning of the proper inclusion here is that there are more worlds in which T'''s

[27] I call the triple 'minimal' data because it does not contain specific structure, which we normally think of as giving a model, and its interpretation, its particular features. So, the internal interpretation constructed from a triple may be rather abstract: yet, the claim is that it is an entire world!

[28] The idea here is that (a) the theory or model is advocated as a cosmology, (b) the interpretation includes philosophical or metaphysical aspects.

propositions are true than there are for T', and consequently T'' is logically weaker);[29] and

(iii) The domain of applicability D_W coincides with the world described, i.e. $D_W = W$.

Note that the possible world W is fixed by the interpretation. Unextendability is thus a relation between bare theories and worlds, and is thus a property of *interpreted theories.*

By a 'complete theory', in (i), I mean a theory that is well-defined, consistent, and encompassing all the empirical data in a certain domain D_W (i.e. a partial surjective map). Condition (ii), in addition, requires that there is no extension of the theory at W: or, in other words, the theory already describes all the physical aspects of the relevant domain at W. Since the relation of isomorphism in (ii) is formal, (ii) is a sort of "meshing" condition between (i)—or, more generally, between the idea of "not being extendable"—and the formal relation of isomorphism between bare theories.[30] Condition (iii) is the usual condition that T is a theory "of the whole world", i.e. its domain of application is the entire world. Notice that (i) and (iii) do not suffice for theoretical, i.e. physical, equivalence: one needs something like the technical requirement (ii) (see the brief discussion of M theory below).

Notice that, for a theory to be unextendable, it is not enough for it to describe an entire possible world in full detail. One needs, in addition, some argument to the effect that the theory is in some sense unique or, better, isolated in the space of related theories, and that in that sense the theory cannot be extended. For if it could be extended, the interpretation could thereby map to a different possible world, thus invalidating the former interpretation and hence theoretical equivalence. As I will argue in a moment, symmetries can secure unextendability, since in effect they give arguments that a theory is unique within a set of theories based on a given set of fields, and requiring a given set of symmetries.

In what follows, I will often use 'unextendability' in the weaker sense that a model might be extended in some cases, but only in such a way that the interpretation in the original domain of application does not change, and the models continue to be physically equivalent after the extension.

The example of position-momentum duality in elementary quantum mechanics can be interpreted internally once we have developed quantum mechanics on a Hilbert space. The two models are then representations of a single Hilbert space, and we can describe the very same phenomena, regardless of whether we use the momentum or the position

[29] T' is a fiducial theory that may well be *identical* to T. But in general, it may be the case that $T \subset T''$ is not true but $T \cong T' \subset T''$ is. In other words, $T \subset T''$ may only be true up to isomorphism.

[30] I have argued that unextendability is a sufficient, though not a necessary, condition for the justified use of an internal interpretation in the sense of the beginning of Section 4.3. The condition is not necessary because one can envisage a theory (e.g. general relativity without matter) receiving an internal interpretation (e.g. points are identified under an active diffeomorphism, taken as the lesson of the hole argument). This interpretation does not change when we couple the theory to matter fields: and I will say that such an internal interpretation is *robust* against extensions. If all possible extensions of a theory preserve an internal interpretation, then such an interpretation is justified in the sense here meant. If the extensions suggest diverging interpretations, then one needs to specify the domain of the extension before one is justified in interpreting the theory internally. In other words, (i)-(iii) can be weakened, if what we want is not sameness of reference but of descriptive abilities.

representation. And quantum mechanics is an unextendable theory in the weaker sense (namely, with respect to this aspect with which we are now concerned, of position vs. momentum interpretations) because, even though we do not have a "final" Hamiltonian (we can always add new terms to it), the position and momentum representations keep their power of describing all possible phenomena equally well, whatever is the Hamiltonian. Namely, because of quantum mechanics' linear and adjoint structure, unitary transformations remain symmetries of the theory regardless of which terms we may add to the Hamiltonian: so that our interpretation in terms of position or momentum will not change.

Dualities in string and M theory are also expected to be dualities between unextendable theories, at least in the weak sense. For example, they have the maximal amount of supersymmetry possible in ten dimensions; and the number of spacetime dimensions is determined by their internal quantum consistency. Also, their interactions are completely fixed by the field content and symmetries. If we imagine, for a moment, that these theories are exactly well-defined (since the expectation is that M theory gives an exact definition of these theories), then they are "unique" in the sense meant earlier: they are picked out by the field content, the amount of symmetry, and the number of spacetime dimensions.[31]

As De Haro (2020: Section 1.4) discusses in the example of the effective field theory programme, claims of theoretical equivalence between incomplete or inconsistent theories are not justified in the sense here meant.[32] Namely, if one is dealing with an effective theory, one must always be prepared to see its interpretation change, precisely at those points at which the theory becomes inconsistent—and that may call for a major overhaul of the theory's interpretation. In such cases, talk of the 'possible worlds' described by the theory cannot be literal, since there are no such worlds. At best, there are limited domains of application within larger worlds that are described by the theory. Thus general considerations of 'toy cosmologies' or "theories of everything", without regard for the finiteness and consistency of the theory, are insufficient. Unextendability is a natural condition that is sufficient to fix this. I consider this to be one major contribution of dualities in string theory and quantum field theory to analyses of theoretical equivalence in physics.

De Haro and Butterfield (2018: pp. 348-373) propose the common core theory involved in bosonization as example of an unextendable theory. For other examples of unextendable theories, see De Haro (2020: Section 2.3).

4.4 Equivalence and Duality: Discussion and Further Work

This Chapter has summarised what I have called the Schema for dualities. The Chapter necessarily had to be brief on a number of issues, of which I now mention several. For

[31] See also Dawid (2006: pp. 310-311), who discusses a related phenomenon of 'structural uniqueness'. Huggett (2017) argues that, although the radius of the compact dimension is *prima facie* indeterminate, one can nevertheless assign a radius to it phenomenologically, in such a way that the observers in both theories agree about the radius of this dimension. Such a phenomenological interpretation can of course be part of an internal interpretation, in so far as it is developed using objects described by the theory.

[32] De Haro (2019b: Section 2.3.2) presents other details about this argument.

more details about the Schema itself, see De Haro (2020) and De Haro and Butterfield (2018). De Haro (2020: Section 1.4) also compares dualities to other work on theoretical equivalence, most notably by Glymour. For more about the relation between theoretical equivalence and duality, see De Haro (2019b). This paper also includes a discussion of the relation to other work, in particular Butterfield (2018) and Read and Møller-Nielsen (2018) (De Haro (2020b) also discusses the relation to e.g. Le Bihan and Read (2018)). The relation between duality and empirical equivalence (in Quine's syntactic, and van Fraassen's semantic versions) is the topic of De Haro (2020b). This is then applied, in De Haro (2020b), to the empirical under-determination argument against scientific realism. For examples of unextendable theories, see De Haro (2020: Section 2.3).

De Haro and Butterfield (2019) is entirely about the relation between symmetry and duality (see also (2018: pp. 328-335)). For the relation between dualities and gauge symmetries, see De Haro, Teh, and Butterfield (2016, 2017). The role of diffeomorphisms in dualities is discussed in De Haro (2017a). Other papers illustrate the notion of duality in various examples. Bosonization is discussed in De Haro and Butterfield (2018: pp. 348-373). For different formulations of Maxwell's theory of electromagnetism, see De Haro (2019b: Section 3). For gauge-gravity duality, see Dieks et al. (2015), De Haro (2017), and De Haro (2020: Section 2). (A philosophically motivated review of gauge-gravity dualities is in: De Haro, Mayerson and Butterfield (2016)).

De Haro (2020a, 2020b) relate duality to other philosophical themes, in particular empirical equivalence and the under-determination argument against scientific realism.

Chapter 5

The Heuristics of Theory Construction

> *If you will not let me treat the Art of Discovery as a kind of Logic, I must take a new name for it,* Heuristic, *for example, only that, as you know, I do not assert such an art to exist* (William Whewell, 1860).

There is a use of duality and of theoretical equivalence that seems to have gone largely unnoticed in most of the literature, and on which this Chapter aims to zoom in. It is the distinction between what I shall call the theoretical vs. the heuristic functions of both dualities and of theoretical equivalence.[1] By applying the Schema for dualities to the heuristic function of duality, the Chapter aims to shed light on the practical use of duality and theoretical equivalence to *construct new theories* out of approximately dual models—which is the task of the heuristic function of dualities. If one is lucky, that is! For heuristics, of course, never lead mechanically, or with deductive certainty, to novel theories. In other words, I will regard dualities as *tools, or methodologies, for theory construction.* To do so, I will place dualities against a philosophical background discussion about tools for theory construction, and about the different functions that those tools have. The discussion should be useful for understanding different approaches to theory construction in physics more generally.

Analysing the role of dualities as tools for theory construction should help us to conceptualise some dominant practices in physics: and, in particular, it should be instrumental in explaining the importance that scientists ascribe to dualities in the overall string theory and M theory programmes. As such, a good conceptualisation of dualities, as elements of scientific practice, could also be instrumental for broader questions of theory assessment and of the progressiveness of the programmes—a question which I will nevertheless not take up here.

[1] The only exceptions I am aware of are: Rickles (2011, 2013, 2017), Dieks et al. (2014), and De Haro and Butterfield (2017), in the case of duality, and Coffey (2016), in the case of theoretical equivalence. I will discuss Coffey's views in Section 5.2.4 (cf. also De Haro (2018: §3.4.1)).

We can see my distinction between the theoretical and the heuristic functions of dualities or, more generally, theoretical equivalence, in a widespread difference in how they are discussed. On the one hand, we are told (sometimes with the addition of several exclamation marks!) that dualities imply that very different theoretical descriptions give rise to the same physics, or *equivalent physics*: and so, a theory of gravity in D dimensions is dual to a gauge theory in $D-1$ dimensions (which goes under the name of 'gauge-gravity duality'). Or a theory in a very large volume, V, is dual to a theory in a very small volume, $1/V$, in appropriate units (this is called 'T duality'). And so on. But, on the other hand, we are also told (sometimes adding even more exclamation marks!), that dualities point to *new physics*: and so, gauge-gravity dualities 'point towards a new definition of string theory', or the dualities between the five different 10-dimensional string theories and one supergravity theory (some of which are related by the *same* T duality mentioned before) 'point to the existence of a, so far unknown, 11-dimensional M theory'. And so on.

Although the philosophy of dualities is now thriving,[2] the recent philosophical literature has—apart from the occasional mention—failed to analyse in detail this second aspect: i.e. physicists' claims that dualities 'point to new physics'. The literature also does not seem to have noticed the different kinds of claims that physicists make about dualities, and the kinds of expectations that are associated with such claims. In fact, the recent philosophical analysis of dualities has almost unanimously sided with the former interpretation of dualities: as cases of weak theoretical (and sometimes also physical) equivalence. This is understandable for the still young literature on dualities: after all, there is a venerable tradition in the philosophy of science of analysing equivalence, which is rooted in studies in logic and mathematics—and philosophers are less prone to mingle with physics that is not settled, or with theories that are still under construction. Thus 'heuristics' is often left to the physicists.

The general philosophy of science literature on theoretical equivalence, on the other hand, *has* noticed—and discussed in some depth—a similar distinction. But it is only similar, and not the same; because in the general context of theoretical equivalence, the issues that have been discussed so far are slightly different. Coffey (2016: Sections 3.1-3.2) has noticed an 'asymmetric treatment' of theoretically equivalent cases, which is similar to the distinction I am drawing here. But his treatment differs widely from mine, as I will discuss in Section 5.2.4.

Thus, by siding with the 'equivalence' account of duality and its corresponding use, which—I will agree—is indeed the correct account if one wishes to explicate the *nature* of duality (cf. Chapter 4), the philosophical literature has left unanalysed the main use that physicists make of dualities: namely the construction of *new* theories, as in for example the influential M theory programme.

Thus: the distinction between the theoretical and the heuristic functions is this. On the one hand, dualities describe *equivalent* theories (i.e. they make new connections be-

[2] See, for example, a recent special journal issue edited by E. Castellani and D. Rickles (2017). For some recent philosophical discussions of dualities in string theory, see e.g. De Haro et al. (2015), De Haro (2020), Dieks et al. (2015), Fraser (2017), Huggett (2017), Read (2016), Rickles (2011, 2013, 2017), Matsubara (2013).

tween the physics described by different-looking, but *given*, theories, e.g. by describing the common core that is shared between two theories). They assume we have almost complete control over those theories, so that duality conjectures can be used to develop a common core theory. And on the other, dualities are used to develop *new theories*, which apparently go *beyond* that common core: i.e. the duality is supposed to lead us to the new theory.

One might think that there is a tension here: if duality only expresses the equivalence between already existing theories, then it is not entirely clear how duality can help develop a theory that *supersedes* the two existing theories, or develop a candidate theory that will succeed them—thereby invalidating the preceding theories, and their duality relation.

But I will argue that the tension is only apparent: it corresponds to two different *functions* of duality in scientific practice. Namely, duality-as-theoretical-equivalence assumes that the two theories are well-defined, and requires that their theoretical descriptions be exactly equivalent; while the accounts that we are given, for how new physics arises from dualities, invariably assume that both the duality and the theories involved are *not exact*, and in fact *cannot* be rendered exact, not only with current knowledge, but even in the context of the theory yet to be developed, which only instantiates duality *approximately.* In fact, the physics literature sometimes moves seamlessly back and forth between these two views of duality: and only confusion can ensue from the mixing of these two functions.

In this Chapter, I will use the Schema for duality to clarify the distinction between these two different functions and to develop in detail the heuristic function. As such, the Chapter can be seen as an application of that Schema, thus further supporting the Schema's applicability.

The plan of the Chapter is as follows. In Section 5.1, I give some background about the idea of tools for theory construction, and define the two functions of these tools that I will consider. In Section 5.2, I expound the basic distinction between the two functions of duality. In Section 5.3, I expound the heuristic function of dualities.

5.1 Two Functions of the Tools for Theory Construction

5.1.1 The theoretical function

In this Section, I briefly introduce the *theoretical function* that tools for theory construction can have.

There are of course many kinds of theoretical tools used in physics: for example symmetry arguments, analogies, approximative relations, and indeed dualities. These tools can take on different functions, i.e. they can be put to use in different ways, under different constraints, and for different purposes (even if the general aim, as we assume throughout this Chapter, is invariably taken to be 'theory construction').[3] The theoretical

[3] There is a lively debate, in the philosophy of technology, about the correct notion of 'function' (see Houkes and Vermaas (2010: Chapter 3)). I will not need to develop a function theory, although I believe that the analogy between artefacts and scientific theories can be used to specify more precisely what one means by a 'function' of a scientific theory. Functionalism is of course a large topic in the philosophy of

function I have in mind is aimed at developing a *given* theory, i.e. not a novel theory, according to constraints. It is the aim of extracting the content of a theory "that is somehow already there", even if only implicitly, using a set of rules or a recipe. The set of rules or recipe is then the tool in question, though use of the tool of course never by itself guarantees success.

The phrase 'the content of the theory is "already there"' should be understood by analogy to how, in elementary logic, the conclusion of a syllogism is implied by its premisses. (J. S. Mill (1882: Bk II, Ch III) argues that there is a *petitio principii* in all deductive reasoning). Here, however, we are dealing with theories in the physical sciences, and so not all recipes or sets of rules will be cases of logical deduction.

For example, once one knows the Hamiltonian (i.e. the energy function) describing a system in classical mechanics, one can use Hamilton's variational principle to derive the equations of motion for that system. In doing so, one may encounter problems (e.g. the difficulty of how to choose appropriate boundary conditions for a given situation), so success is not guaranteed. Nevertheless, the equations of motion are, in essence, "already there" once the Hamiltonian is given, since there is a recipe or set of rules which lead from a Hamiltonian to the equations of motion, partly deductively. That recipe, i.e. Hamilton's variational principle, is the tool in question, used in its theoretical function of finding the equations of motion, i.e. of extracting the full theory (of which the equations of motion constitute the dynamics).

The theoretical function, as just presented, comes with a partly deductive procedure to formulate the theory, T: and so, the theory is, in a way, already there from the start.[4] Thus the method is not aimed at finding physical novelty: even if in some cases it may find some novelty—precisely in those cases in which complete deduction fails. Rather, it is aimed at making more perspicuous the conceptual and-or mathematical presentation of the theory T. This is what I call the theoretical function of a tool.

One might object: why call *duality* a 'tool', and the use made of this tool a 'function'? Is it not simply a case of providing a mathematical proof of duality, i.e. is duality not the thing we wish to prove, rather than the tool to achieve a goal?

The answer is No, and for three reasons. First: a proof of duality, given a set of models, only requires the *existence* of a common core theory, of which the models in question are representations, and of a proof of an isomorphism between the models: the proof does not require the actual *construction* of the common core theory, T, which is the aim of the theoretical function (cf. the definition of duality in Section 4.1.2, does not explicitly mention the theory, but only its models). Thus the aim of *constructing* a common core theory, T, is more ambitious than the aim of proving duality.

Second, the theory thus constructed need not be unique, as emphasised in De Haro

mind: see Lewis (1972) and the brief report of it in Section 2.2.2.

[4] A theory, T, thus formulated may of course have unforeseen consequences, such as the existence of new (non-isomorphic) models. (Think of the axiomatizations of first-order arithmetic—any possible ones, thanks to Gödel 1931). Elucidating the physical interpretation of those new models will be the task of the heuristic function. Notice that the theoretical function of course has an element of heuristics of its own (cf. the remark above about boundary conditions): but the method of the theoretical function is a largely *deductive*—sometimes even a mechanical—one, i.e. it follows prescribed rules.

and Butterfield (2017: §2.4). Thus, there is judgment involved in deciding which theory, T, is the most appropriate one, in a specific situation.

Third, there is also choice that scientists can make between attempting to try to construct the common core, i.e. the theory T, or not constructing it. For the construction of theories exhibiting a duality is not a necessary aim of scientific theories: a theory with a duality need not necessarily be better than a theory without it. In other words, theorists faced with a duality are free to construct the theory T or to not construct it, and there can even be a choice of the theory T among a number of competitors, e.g. how much of the structure of the models M_1 and M_2 to deem "specific" and not being a homomorph of the yet-to-be-formulated common core theory T.[5]

An analogy with symmetries and with approximative relations is helpful here. Imagine a rule that produces, given an input state, an output set of states, according to a symmetry principle (e.g. given a wave-function with given energy, symmetry considerations are used to produce other wave-functions with the same energy). This is done by the theoretical function.

But now imagine a more adventurous use of symmetry in which, given an equation of motion that does not display a given symmetry (e.g. because it is written in a specific gauge or coordinate system), one writes the equation in a manifestly symmetric way (e.g. in a gauge-invariant way, or as a covariant equation). Once again, one can here give rules for such a procedure: of symmetrisation, or covariantization. This use of symmetry also falls under the theoretical function, for two reasons:

(i) there is a well-defined general rule, saying 'for any equation A of a certain kind that does not exhibit a symmetry, there is an equation B that is manifestly symmetric',

(ii) it does so in such a way that the number of degrees of freedom is not modified, in particular it is required that no new physical degrees of freedom are introduced.

Thus, though not strictly deductive in the logical sense, this use of the theoretical function still operates according to a general rule, and it does not introduce "new physics".

Similarly for an approximative scheme: this theoretical 'function' is a matter of a reduction, or linkage, between two theories, as discussed in Chapter 3. Given a basic, or 'bottom', theory $T_{\rm b}$, the approximation scheme is a rule that produces a new, 'top', theory $T_{\rm t}$. Again, though the success of the application of this rule is not guaranteed: if it succeeds, then one ends up with a theory $T_{\rm t}$ that is obtained from $T_{\rm b}$. Because there is reduction, or at least linkage, there is a sense in which the degrees of freedom of $T_{\rm t}$ can be taken to be derived from those of $T_{\rm b}$ (under suitable assumptions about the approximation).[6]

[5] This is discussed further, under the heading of 'abstraction', in De Haro and Butterfield (2018: pp. 325-327, 367-369).

[6] I am assuming that there is no novelty in the interpretation of the top theory, relative to the basic theory, and thus excluding cases of emergence as discussed in Chapter 3. For I take a procedure that helps develop a *new interpretation* of a theory to be part of a heuristic function as well. The example of approximation here is what Radder (1991) has called 'correspondence from L [$T_{\rm b}$] to S [$T_{\rm t}$]. It is the kind of heuristics considered by Post, Krajewski, and Fadner (mentioned in Radder (1991)). With Radder, and contra these authors, I do not take heuristics to be applicable only, or even mainly, at the formal level.

5.1.2 The heuristic function

In this Section, I briefly introduce the *heuristic function* that tools for theory construction can have.

Whewell (1860: p. 480) described 'heuristic' as the 'art of discovery', which, he admitted, was not to be understood as 'a kind of Logic'. A narrower conception of heuristics is as a set of 'efficient rules or procedures for converting complex problems into simpler ones' (Hey (2016: p. 472)). It is the former conception which I have in mind in this Chapter: a tool as used in the art of discovery, and in the construction of new theories.[7] It is a *tool*, rather than a *rule*, because success in theory construction is never guaranteed; nor can the tool be applied mechanically, as the phrase 'efficient rule' would suggest.

Indeed, whenever general, and more or less mechanical, *rules* are involved in theory construction, I will take the corresponding use of the tool to belong to the theoretical function, as discussed in Section 5.1.1, rather than to the heuristic function. The heuristic use of a tool involves craftmanship and creativity, and should lead to the formulation of *new theories*, which contain new physics.

The heuristic function will obviously have some rules of its own (having to do with the sorts of constraints that the new theories should satisfy), but the defining mark lies in the theoretical and physical novelties that are its aims. Novelty in the theory's formalism includes: novelty in the number and nature of states and quantities (or 'physical degrees of freedom'), the dynamics, and the rules for calculating physical quantities (cf. Section 1.2.2). Novelty in the interpretation is novelty in the theory's reference to worldly items, which includes cases of ontological emergence.

Let me illustrate this in the examples given in the previous Section. There, we considered a system for which the Hamiltonian was given. The theoretical function was a rule that gave us the equations of motion for this system. Now consider a system for which the Hamiltonian needs to be found. The main difference between the two cases is that, in the former, there is a recipe, indeed a partly deductive procedure. In the second case, there is no such procedure, and the arguments required are of a different kind. Scientists indeed use heuristics when, given a physical system, they try to find a Hamiltonian describing this system. The procedure in question may involve writing down parts of a Hamiltonian (or limits of it) which they already know from similar systems: but it also involves educated guesses about those parts of the Hamiltonian which they do not yet know, e.g. because they describe some of the system's novel, or even unique, features. Such tentative guesses are usually informed by different kinds of arguments: symmetry arguments, combined with arguments about the number and kind of degrees of freedom to be described, assumptions or constraints about the admissible kinds of interactions, etc. But even if physicists are able to come up with fairly systematic rules constraining the admissible classes of Hamiltonians (though this usually only works for a class of *similar problems*), in the end there is no mechanical, or indeed general, rule for writing down the Hamiltonian describing a physical system. It is ultimately always a matter of creativity, craftmanship and, often, luck; and the best one can do is verify that it describes the target

[7] The heuristic function is of course not limited to the construction of new theories. It in fact plays an important role as well in the application of known theories, in theory confirmation, etc. However, I will here restrict the discussion to theory construction, which is relevant to Section 5.3's main example.

system accurately, in specific situations.

Recall the example, in Section 3.1.2 and at the end of Section 5.1.1, of an approximative scheme (e.g. $\hbar$ small compared to the action) relating the top theory, $T_{\rm t}$, to the bottom theory, $T_{\rm b}$, especially in cases of emergence. The heuristic relationship between $T_{\rm b}$ and $T_{\rm t}$ now goes in the *opposite* direction to that discussed in Sections 3.1.2 and 5.1.1. Given an approximative theory, $T_{\rm t}$, and given an approximation scheme from which one believes $T_{\rm t}$ (or a theory close to it) is obtained, physicists' job is now to try and guess, or to somehow reconstruct, the basic theory, $T_{\rm b}$. Again, such educated guesses are subject to constraints, but $T_{\rm b}$ can ultimately only be justified if it describes more phenomena than $T_{\rm t}$ does.[8]

5.2 The Two Functions of Duality

In the previous Section, I discussed two of the functions that tools for theory construction can have: a theoretical function and a heuristic function. In this Section, I will discuss those two functions for dualities in string theory, and show how they differ.

5.2.1 Motivating duality: string theory and the M theory programe

I first briefly introduce, in this Section, the main ideas behind the string- and M theory programme: and, in particular, the role of duality within that programme.

String theory is a candidate theory for the unification of general relativity and quantum field theory. Its basic assumption is that matter is made of strings, i.e. extended, one-dimensional objects that can vibrate, move around in spacetime, and interact by joining or by splitting.

For string theory to be mathematically consistent, 10 spacetime dimensions are required for the strings to move in (6 of which are thought to be curled up, so that they are inaccessible to current experiments). In the low-energy limit, string theory is well-approximated by supergravity theories, i.e. supersymmetric extensions of Einstein's theory of general relativity, which are also 10-dimensional, and compactified down to four dimensions.

Initially, five different string theories were known, differing over the precise details defining the strings. However, significant dualities were found relating them to one another. T duality, for example, relates one type of string theory on a circle of radius R, to another type of string theory on a circle of radius $1/R$. And electric-magnetic duality (so-called S duality) relates some other string theories.

In 1995, Witten conjectured that the five known string theories, plus in addition a sixth known, 11-dimensional, supergravity theory, were all different limits of (approximations to) a single 11-dimensional theory, which he dubbed M theory. Witten assumed the eleventh dimension to be a circle, which could be of one of two kinds. He identified the

[8] There is of course no claim here that $T_{\rm b}$ is descriptively strictly better than $T_{\rm t}$, in each and every respect. Kuhn losses may well be our fate! (see Section 2.3.4, especially Figure 2.4). But at least $T_{\rm b}$ does explain the success of $T_{\rm t}$. This possibility of a 'loss', when going from $T_{\rm t}$ to $T_{\rm b}$, is of course the counterpart of emergence, in the reverse direction.

radius of this circle with the coupling constant ruling the joining and splitting interactions of the strings. For a small circle of the first kind, the string coupling is weak, so that one of the five known 10-dimensional string descriptions of weakly-coupled strings (the so-called *perturbative string theory*), is accurate. For a small circle of the second kind, another of the five known versions of perturbative string theory is accurate. The other three string theories are related to these two by T and S dualities.

But, at strong coupling, the eleventh dimension opens up, and the perturbative string descriptions are no longer valid. Eleven-dimensional supergravity provides a semi-classical description in 11 dimensions, valid at strong string coupling but only as long as the length of the fundamental string is small, i.e. in the point-particle limit of the string (or whatever replaces it in eleven dimensions). The challenge is then to find a theory valid away from the point-particle limit: this should be the sought-for M theory.

Since Witten's conjecture, two main approaches to M theory have been taken. The first is the conjecture by Banks, Fischler, Shenker, and Susskind (1997) that M theory is a theory of matrices, with eleven-dimensional supergravity as its low-energy limit.

The second main approach is AdS/CFT, which is a series of conjectured dualities between string theory or M theory in asymptotically anti-de Sitter space (AdS, i.e. a manifold of negative curvature), and a specific quantum field theory at the boundary of this space (where CFT stands for 'conformal field theory'). Compactifying M theory on e.g. an internal seven-dimensional manifold of positive curvature, the remaining four dimensions have negative curvature, and are dual to a three-dimensional CFT, for which exact treatments exist. This approach is more generally called 'gauge-gravity duality', because it relates a theory of gravity to a quantum field theory with gauge symmetry.

Details aside, M theory is the main unifying conjecture behind the various versions of string theory, and dualities play a key role in the attempt to formulate M theory. What remains unclear is the precise status that dualities are supposed to have in M theory, once a non-perturbative version for it is found. Should M theory exhibit duality, or should dualities be superseded by the final theory—are they merely "ways towards the formulation of a new theory"?

This question is, of course, here intended, not as about trying to peek into the future of theories that do not yet exist, but as about the heuristic paths of investigation that one may reasonably take dualities to suggest. We will explore the role of dualities within this programme in Sections 5.2.2 and 5.2.3. Here I anticipate by saying that the answer to this question will come down to a different function of duality.

The conjectural status of most dualities in string theory, and of M theory itself, should not be a reason to dismiss the programme as philosophically irrelevant, or as mere speculation. There are four reasons for this, which I here briefly list:—

First, the programme is very influential in physics: and, in the last thirty years or so, it has spawned a large number of new ideas and technical developments which (arguably) no other research programme in high-energy physics has been able to produce. Second, and more importantly, being conjectural does not mean being physically and mathematically unmotivated.[9] For the evidence that is available for some of the string theory dualities is strong and compelling. Third, there are also rigorous results, at various levels of mathematical and physical rigour: especially about the conformal field theories, random matrix models, and quantum field theories involved, fairly rigorous mathematical results exist. Finally, it is of course simply false that philosophy should limit itself to studying theories that are already in final form and that are mathematically completely rigorous: for not only would philosophers then quickly run out of a job, but also because it is their task to clarify and assess whatever fragments of theory are available (cf. Huggett and Wüthrich (2013: p. 284)). This is especially true in areas of research where direct observations are so far absent, and so the main guidance is the—apparently very strong—requirement that general relativity and quantum field theory should be reproduced in suitable approximations, and in addition one has the requirement of mathematical consistency and the tools of conceptual analysis at one's disposal (besides what little available evidence there is from experiments and analogue experiments). Rather than making these theoretical attempts uninteresting for philosophers, these four reasons make philosophy relevant, even indispensable, to the programme of string theory.

5.2.2 Duality as exact equivalence: duality's theoretical function

In this Section, I discuss within string theory the theoretical function of duality, in the sense of Section 5.1.1, where duality is construed as in the Schema from Section 4.1.

Recall, from Section 5.1.1, the idea that the theoretical function extracts the content of a theory "that is already there". Thus the theoretical function, as applied in this Section and the next to *dualities,* never tries to be more general than, or to describe physics or degrees of freedom beyond, the dual models: while the heuristic function does.

The Schema construes duality as an isomorphism between models. This isomorphism relates the common core that the two models deem physical (i.e. the triple of states, quantities, and dynamics). As such, duality is a formal notion, i.e. a definite relationship between uninterpreted, but physical, models. It relates triples of states, quantities, and dynamics on the two sides, preserving the structure of the models (including the values of the quantities, evaluated on the states). Thus duality is not *merely* a formal relation, because it deals with *physical* models, but by itself it makes no reference to interpretation—the latter is the question of what, in Chapter 4, I called 'physical equivalence'.

Both physicists and philosophers tend to construe duality this way. Therefore, the theoretical function of dualities, i.e. the function that follows from the *nature* of duality, as outlined in Section 5.1.1, is to establish theoretical relationships (more specifically: to establish isomorphism) between models. These relationships typically entail relating states and quantities in one model, to states and quantities in another model, and also relating the dynamics of one model to the *different*, but isomorphic, dynamics of the other.

[9] The first two reasons of course run into large debates about the present state of, and future planning for, physics. See Dawid (2013) and Ellis and Silk (2014) for opposite sides of these debates.

Thus dualities are very strong relationships between two models, since they relate everything that the models deem physical (namely, the model root m that is within the model M in Eq. (1.1)). Establishing a duality between two models thus presupposes precise knowledge of the elements of the two models (the sets of all the states and quantities, and the complete dynamics), as well as knowledge of the relations in which these elements stand (i.e. there are not only bijections between each of the elements of the triples of the two models, but all physical structure must also be preserved). Thus establishing a duality requires a formulation of a model that captures all of those details, even if perhaps only implicitly. (Full transparency of the model, or full understanding of it or perfect computational power, are of course not (and cannot be) required). I will say that such a model (i.e. one where all the states and quantities, and the complete dynamics, as well as the complete rules for calculations, are known and are consistent, within the domain of application of the model) is *exact*.[10]

Notice that this notion of being mathematically well-defined, within a domain of application, is much weaker than the requirement that a model gives a non-vague, good, or successful description of the domain—the former is a formal requirement, while the latter is interpretative and empirical.

Exactness can be proven for a number of significant dualities in physics. Simple examples are the Fourier transformation in elementary quantum mechanics, harmonic oscillator duality, and electric-magnetic duality in electrodynamics. For more sophisticated dualities in quantum field theory and in string theory, the only case, so far as I know, in which the philosophical literature has proven a duality to be exact is the example of boson-fermion duality in two dimensions (De Haro and Butterfield (2017)), though in the physics literature there are other cases. Most dualities in string theory are cases of dualities which are *conjectural*. Nevertheless, it is an important aspect of duality that *all* dualities are exact—as they must be, according to the above definition.

The physics literature confirms the claim that dualities must be exact: i.e. that the *definition* of duality entails that they are cases of exact, and not approximate, isomorphism, within a domain of application. Also, the physics literature confirms that duality is a case of weak theoretical equivalence, i.e. of a formal, or mathematical, relationship between two physical models, as in Section 4.2. I will now substantiate this consensus with some quotations from the physics literature, which also illustrate how physicists think about dualities.

The literature quoted below of course also emphasises the following aspects: seemingly different physics and difference of description, but equivalence (or sameness) of theory; and the exactness of the duality, and of the theories involved, is also denoted as the theory's being 'non-perturbative', i.e. its formulation goes beyond, or does not require, perturbation theory.

(A) In the Glossary of his textbook on string theory, Polchinski (1998, p. 367, my em-

[10] Exactness is a prerequisite of duality, because it is part of the isomorphism condition. My use of the phrase 'exact duality' simply emphasises duality's exactness. It is not meant to suggest that 'inexact dualities' exist, which would be contradictory. Rather, dualities can be inexact, or approximate, in a different sense: namely, in the sense that they are not instantiated by the models in an exact manner. Whenever I use the adjective 'inexact' in connection with dualities, it will be in this sense.

phasis) defines duality as: ‘the *equivalence* of seemingly distinct physical systems. Such an equivalence often arises when a single quantum theory has distinct classical limits.’

He describes one specific duality (T duality) as a case of sameness of theory, but difference of description: ‘T-duality is just a different description of the same theory’ (p. 268). ‘[T-]duality is a symmetry not only of string perturbation theory but of the *exact* theory (p. 248, my emphasis).

(B) In an influential paper putting forward a specific definition of M theory, Banks et al. (1996: Abstract, my emphasis) also regard duality as an exact equivalence. Thus they write:

> We suggest and motivate a *precise equivalence* between uncompactified eleven dimensional M-theory and the $N = \infty$ limit of the supersymmetric matrix quantum mechanics’. ‘If our conjecture is correct, this would be the first *nonperturbative formulation of a quantum theory* which includes gravity’ (p. 2, my emphasis).

And later they say:

> Our conjecture is thus that M-theory formulated in the infinite momentum frame is *exactly equivalent* to the $N \to \infty$ limit of the supersymmetric quantum mechanics described by the Hamiltonian (4.6) (p. 11, my emphasis).

(C) In an influential review on gauge-gravity duality, Aharony et al. (1999: p. 57, my emphasis) formulate duality in terms of sameness of theoretical description, or theory:

> Thus, we are led to the conjecture that... Yang-Mills theory in 3+1 dimensions is *the same as (or dual to)...* superstring theory on $\text{AdS}_5 \times S^5$’.

They extend this conjecture to a full equivalence between string theory and gauge theory:

> The strong form of the conjecture, which is the most interesting one and which we will assume here, is that *the two theories are exactly the same* for all values of g_s and N [i.e. the string coupling constant and number of colours, respectively].’ (p. 60, my emphasis).

The common thread is clear: these are all cases of conjectured, but *exact*, equivalences of the theoretical structures (sometimes, in a limit of the physical parameters that is relevant to the theories involved). This is in agreement with the Schema’s definition of duality, given in Section 4.1.2, and it grounds the theoretical function of duality: namely, duality thus construed is a relationship between models that are already there and which were previously thought to be unrelated.

In light of the discussion in Section 5.1.1 on the theoretical function of a tool, we can now understand a conjectured duality as a help in finding more perspicacious formulations of a given model. This is for example the case when physicists use the better-known side of the duality to investigate the lesser-known side. This is akin to solving a quantum mechanics problem (even: formulating a model description of a system) in momentum space, and then doing the Fourier transformation back to position space. This use of the Fourier transform, which is a deductive rule that by itself does not add any new degrees of freedom, is a translation of one model description to another, and so it belongs to the theoretical function. Unless the model description A was already known, the Fourier transform would be of no help in getting the model description B via duality. It is only when the model description A is already worked out, that we can find out more about the model description B, in a quasi-mechanical way, using the Fourier transform. The same remarks go through for other dualities, in this kind of use.

But notice the assumption behind string theory dualities: within the theoretical function, the duality relation itself *will not change*, once the two dual models are formulated to our satisfaction. Rather, the search for a satisfactory formulation of two dual models is a search for two structures that stand in precisely the relation that is described by the duality conjecture, which is assumed to be *exact*. On this view, duality is not to be superseded in the theory one is aiming to construct: rather, *establishing duality* is the aim of the proof of the duality conjecture, since this gives insight into the theory. The duality is to be instantiated by the final pair of models: perhaps in a manifest and completely obvious way, on a sufficiently perspicacious formulation of them. Following my earlier usage, I will call the theory T, thus obtained, the *common core theory*: for this core structure is assumed to be isomorphic between dual models, i.e. it is the model root of each of the models.

5.2.3 Duality and approximation: duality as a heuristic for theory construction

In this Section, I discuss within string theory the *heuristic function* of duality, in the sense of Section 5.1.2, and give some quotations from the physics literature supporting the existence, and even the essential role, of this function.

The physics quotations below also emphasise the lack of exactness of the theories involved (viz. they are perturbative) and the use of dualities as heuristics for finding new unifying theories (or new formulations of old theories, describing more physics). The heuristic function, in the context of this literature, is then seen to be strongly linked with the aim of *unification*. The examples are as follows:

(A) In a review paper about dualities, Dijkgraaf (1997: p. 120, my emphasis) connects the approximate nature of dualities to the suggestion of the existence of new theories:

> The insight that all perturbative string theories are different expansions of one theory is now known as string duality... *It is one of the amazing new insights following from string duality that these "theories" are all expansions of one and the same theory around different points in the moduli space of vacua.*

'Expansion... around a point' should here be taken in the sense of, for example, a Taylor series expansion of a function about a particular point: which is captured by the notion of 'approximation', discussed at the end of Section 5.1.2. Dijkgraaf also emphasises the 'perturbative' nature of the dual models, i.e. their lack of validity beyond a certain order in such an expansion. Thus, Dijkgraaf's picture of dualities is one which regards models as *inexact*, and dualities as only approximately instantiated, i.e. the dualities are valid only within a limited range of parameters, but are to be *superseded* by a better theory, namely what he calls 'one and the same theory', of which the mutually dual models are expansions, i.e. approximations.

(B) In the paper in which Witten put forward the influential M theory conjecture, he wrote (1995: p. 2, my emphasis):

> S-duality between weak and strong coupling for the heterotic string in four dimensions... really ought to be *a clue for a new formulation of string theory.*
>
> ...in this paper, we will analyze the strong coupling limit of certain string theories in certain dimensions. *Many of the phenomena are indeed novel, and many of them are indeed related to dualities...* (p. 4).

These quotes by Dijkgraaf and Witten underline a related aspect of dualities: they use terms like 'amazing', 'new insights', 'clue for a new formulation', 'novel phenomena'. The emphasis here, unlike the quotes from §5.2.2, is not on the conjectured equivalence between already existing models: but on the *novelty of theory* which can arise once a duality between such models is understood.

They also emphasise duality's pointing to 'a new formulation of string theory': where I take it that 'a new formulation' is more than just a '*re*formulation': for a new formulation contains something extra, not only in terms of the mathematical formalism, but also in terms of the physics that is associated with that formalism—as the papers confirm, when they talk about novelty of phenomena.

Thus, dualities here point to the existence of new theories, but are ultimately bound to be superseded: the new theory, once found, will explain these dualities as being the result of certain approximations, which can be done in different ways, but lead to identical results, as articulated in the duality. But once that new theory is reached, the duality is no longer needed, except for practical purposes: for the resulting theory is a *single*, complete theory. In other words, establishing duality is here not the goal: rather, it is an intermediate step towards finding a new theory.

In what follows, I will dub that new theory, the one that supersedes the dual models and of which they are particular limits, the *successor theory*, T_{S}.[11]

These two viewpoints thus lead to different uses of duality in string theory. On the view discussed in Section 5.2.2, the goal is to look for a common core theory, T, that realises the dualities as manifestly as possible. On the view in this Section, the goal is to find the successor theory, T_{S}, that is "behind" the dualities, and which reveals them to be approximations. As I will argue in more detail in the next Section, even if they lead to two different research programmes, the two ideas need not contradict one another, and one could pursue both. But it is important to clearly distinguish the two functions: for otherwise, confusion easily ensues about the nature of duality, and about what one is entitled to expect from a duality conjecture.

[11] The fact that there is a theory, T_{S}, which succeeds the dual models, does not imply any specific stance about questions about scientific realism or referential stability across theory change, which was the topic of Chapter 2.

5.2.4 Does the distinction imply a tension?

In this Section, I argue that the distinction between the two functions does not necessarily imply a tension.

At first sight, the previous quotes might suggest the distinction as a tension: in the first case (Section 5.2.2), string theory and M theory instantiate the dualities exactly, while in the second case (Section 5.2.3) dualities are perturbative clues towards finding a new theory, which will not instantiate duality exactly. However, one should interpret these quotations with some care, since they are not very precise (for example, these articles do not even include definitions of what is meant by 'duality') and they involve quantum field theories and string theories which are still being developed. Therefore, some of the central questions, viz. whether the models as formulated are exactly valid, or whether dualities are exactly instantiated by the models, cannot be answered at this stage.

Nevertheless, I argue that the tension does not simply come down to lack of knowledge about the models involved: for the same tension exists for dualities and models which are exact, and well-known.[12]

Here are two important reasons why the two accounts, duality as exact equivalence, and duality as an approximately instantiated equivalence and pointing to new physics, might be thought to be in tension. First, they do not refer to two different levels of explanation or of ontology. Namely, being 'two dual models of a single theory' or being 'approximate dual models of a new underlying theory' both operate at the level of the formal structure: therefore, the potential resolution 'the two accounts operate at different levels, and so they do not contradict one another' is not available. Second, they might be seen to be in tension because the former sense assumes an exact duality, and being an exact instantiation of a theory; while the latter necessitates dualities which are not exactly instantiated, thus pointing to a new (unifying) theory, of which the two models are only approximations.

Nevertheless, I claim that, when made explicit in a language sufficiently precise using the Schema from Section 4.1, the tension turns out to be only apparent, and can be resolved. Namely, one distinguishes two *different theories*, corresponding to two different ways in which the theory to be constructed can relate to the given duality. Duality is then recognised as having two different *functions*, which aim at the construction of different kinds of theories, as I will analyse in §5.3.[13]

[12] This tension is analogous to that between *emergence* and duality (cf. De Haro (2015) and Dieks et al. (2014)). In that case, there exist two mechanisms that can make duality and emergence compatible. First, there can be emergence *independently*, on the two sides of the duality (rather than across the duality relation). Second, there can be emergence if, for whatever reason, the duality is either not exact, or broken. My resolution, in Section 5.3, of the tension between the two functions of duality will be similar.

[13] Coffey (2016) has noted an analogous tension in the context of theoretical equivalence between classical physical theories. Though my tension differs from his (his focuses on asymmetry, and especially asymmetry of ontology; while mine focuses on theory construction), his questions are similar to mine: 'One, how can a symmetric relation of theoretical equivalence accommodate an asymmetry of reformulation? Two, why does this asymmetry only occur in some cases and not others?' (p. 832). But our resolutions are very different: Coffey opts for an interpretative solution; but I think this is a red herring, since the tension is there even for uninterpreted theories.

5.3 The Heuristic Function of Duality and Theoretical Equivalence

In this Section, I come to the central question, of how the Schema for dualities, reviewed in Section 4.1, bears on the heuristic function of duality and theoretical equivalence.

5.3.1 How to use dualities heuristically

In this Section, I will give examples of the use of dualities according to the heuristic function, i.e. for constructing new theories.

Our question is whether the successor theory, $T_{\rm S}$, which one constructs from a duality between models, could be 'bigger' than its models: what we would like the outcome of such a construction to be is a more general and precise theory, which comprises the models as *approximations* to specific physical situations—so, they are approximate representations of the theory.

It is not hard to suggest how this may happen, and I have illustrated this in an explicit example in De Haro (2019a: pp. 5189-5193), which I here summarise. The example is a model of a classical point particle moving on the real line, with a local (canonical) decomposition of position and momentum variables on the cotangent bundle. Different choices of such decomposition, preserving the Poisson bracket, give rise to isomorphic models. The set of transformations relating these different decompositions are given by the group of area-preserving linear transformations, viz. $\mathrm{SL}(2,\mathbb{R})$.

This ability to decompose the phase space in different ways has a parallel at the quantum level, where the $\mathrm{SL}(2,\mathbb{R})$ is now a symmetry of the Heisenberg commutation relations. However, in a theory of quantum gravity the Heisenberg commutation relations are only an approximation to a more general algebra. This general algebra in general does *not* possess the $\mathrm{SL}(2,\mathbb{R})$ symmetry, although it reproduces it in a limit. Here, we regard $\mathrm{SL}(2,\mathbb{R})$ as being the "duality group", i.e. the group that relates various isomorphic models of the same theory.

This argument suggests that there is a successor theory describing more degrees of freedom, of which the given models, that possess the symmetry, are only *approximately* models, i.e. representations. Each of the models should somehow have an embedding in the successor theory (and perhaps even an *extension* into that theory), so that duality does *not* hold exactly in the entire new theory, and it does *not* hold exactly between the extensions of the models, if such extensions are given. The idea is indeed to break the duality.

In conclusion, what the example illustrates is that one can start from a duality group $\mathrm{SL}(2,\mathbb{R})$ to then make $\mathrm{SL}(2,\mathbb{R})$ arise as an approximation of a physical system. This suggests generalisations to a successor theory that contains more possibilities than the ones strictly postulated by duality.

Notice that the successor theory, $T_{\rm S}$, also gives an *explanation* for the duality: namely, in the duality, what are in fact *different models*, or different limits of $T_{\rm S}$, look like isomorphic models M_i, because of the approximations (to $T_{\rm S}$) they introduce.

The model notation $M_i = \langle m_i, \bar{M}_i\rangle$ (for $i \in I$ in an appropriate index set I) from

Eq. (1.1) should make this clear. Isomorphic models are obtained by $\mathrm{SL}(2,\mathbb{R})$ transformations, and so the set indexing the models is $I = \mathrm{SL}(2,\mathbb{R})$. (The specific structure can be written down in terms of the $\mathrm{SL}(2,\mathbb{R})$ transformation used and the representative of the $\mathrm{SL}(2,\mathbb{R})$ orbit). In this example, all the models are isomorphic to each other, i.e. $m_i \cong m_j$ $(i, j \in I)$. The models $\{M_i\}_{i\in I}$ are then isomorphic representations of the common core theory, T.

The heuristic construction of the successor theory, T_S, is then envisaged to proceed in two steps, as follows:[14]

Initially, i.e. for exact dualities, only the model triples m_i $(i \in I)$, i.e. the models stripped of their specific structure, are interpreted as being physical. More precisely: the specific structure, $\bar{M}_i$, gives each model its specificity: it is like the choice of $\mathrm{SL}(2,\mathbb{R})$-representative, and it is not physical.

But subsequently, some new physics modifies the relation that allowed us to stipulate a choice of specific structure, so that (typically!) more variables are needed to describe the problem: which prompts us to interpret part of the specific structure as actually being physical, in the modified model. (Of course, the successor theory might have bits which are utterly novel, i.e. alien to the specific structures of the given models). In the example above, the new physics is given by the quantum gravity effects experienced by the point particle. And so, a choice of specific structure—a choice of representative of $\mathrm{SL}(2,\mathbb{R})$—is no longer innocuous; it is physical. The successor theory, T_S, describing these corrections differs from T, because it incorporates the symmetry as a *physical* symmetry of its triple, at least in a suitable approximation (such as: gravitational effects being negligible), which now incorporates some of what used to be the specific structure. In effect, we have changed the models and the theories: some of the specific structure has now become part of the triple, giving rise to new states (new physical situations), new quantities (which make a distinction between those situations) and new dynamics (accounting for new interactions).

5.3.2 Successor theories and the heuristic function

The heuristic function of dualities is the ability to use dualities in the constructive way just discussed, i.e. for building new theories: starting with exact dualities, viz. isomorphic models, building successor theories that implement the duality as approximate symmetries (or as other constitutive parts of the theory's structure leading to approximately isomorphic models), and then reinterpreting the symmetries, or parts of structure, as special properties of a limit or approximation to a specific physical system. Away from the limit, the number of degrees of freedom of the theory (i.e. number of the states and-or quantities) is typically not reduced, but increases: at any rate, the physical interpretation changes.

Thus there is no longer a duality, but only a theory with one or several approximations or limits: duality holds only approximately, but there is a self-consistent regime in which duality obtains. The use of T_S, from the point of view of duality, is that it explains the physical origin of the duality, and exhibits how duality is implemented. So, duality ends

[14] There is no claim here that the two-step procedure outlined below is somehow unique or necessary. One can think of slightly different procedures that amount to the same thing: namely, a successor theory T_S which is approximated by both T and by T's models.

up being a property of idealised models, but not a property of the physical successor theory. This is what Radder (1991) has called heuristics 'from the old theory to the new theory', i.e. the heuristic function helps one find a successor, more accurate, theory, starting from a given set of models.

The conceptual picture arising should now be clear, in the Schema from Section 4.1. We have an initial theory, T, and its set of isomorphic models, M_i $(i \in I)$. The theory and its models need not be well-defined for arbitrary values of the parameters; they may have limited validity, and also the dualities may hold only approximately (in an idealisation within the successor theory, in which certain interactions, or certain complicating factors are neglected). The successor theory, $T_{\rm S}$, is then able to reproduce T and its models (or something very close to them) as special cases, for particular approximations. $T_{\rm S}$ does not exhibit exact duality, and the models are not exact representations of $T_{\rm S}$. Also, $T_{\rm S}$ often reinterprets the specific structure $\bar{M}_i$ of the models physically: $T_{\rm S}$ then *changes the definition of the models*.

We see that there are two basic ways to make the theoretical and the heuristic functions compatible, i.e. we can:[15]

(i) Extend the models beyond their original domain of application, even if they are no longer exactly dual, and find a successor theory, $T_{\rm S}$, of which those models, perhaps modified, are now exact (but not necessarily dual) representations; or

(ii) We can simply find a new theory, $T_{\rm S}$, of which the original models are approximate representations.

The analogy with symmetries is that the physical contents of T and $T_{\rm S}$ are different. In T, one was entitled (though not invariably obliged) to interpret dualities as mere redundancies. This is no longer possible in $T_{\rm S}$, because the duality is now seen to be a consequence of an approximation to a certain physical situation: in other words, the physical contents of T and $T_{\rm S}$ are different, and part of the physical content of $T_{\rm S}$ now implies the approximate duality.

5.4 Heuristics: Discussion and Further Work

In this Chapter, I began by making a distinction between a theoretical and a heuristic function of duality and of theoretical equivalence. Then, using the Schema from previous Chapters, I described these functions. The aim of the theoretical function is to discover, or to (re-)construct, isomorphic or dual models, together with a common core theory.

The heuristic function aims to discover a *successor theory*, i.e. a theory whose content goes beyond the content of the original models, and of which the dual models are approximate instantiations. Dualities of course often come with indications of the range of parameters for which the successor theory should differ significantly from the given dual models (e.g. 'at strong coupling', for a specific coupling in the theory), and the regime in which it should reproduce the dual models (e.g. 'at weak coupling') or something close to them.

[15] This is analogous to the resolution of the tension between duality and emergence, in De Haro (2015) and Dieks et al. (2014).

The theoretical function also aims at excising the specific structure, i.e. finding a theory T that contains only the common core of the dual models. On the other hand, the examples suggest that the heuristic function *often makes use of the specific structure*, and reinterprets it as physical structure in the successor theory, $T_{\rm S}$.

The schema for duality also illustrates how the two functions of duality can be compatible. We distinguish two theories, T from $T_{\rm S}$, which stand in different relationships to both duality and to the models. The theoretical function aims to construct a theory, T, of which the dual models are exact representations. The heuristic function aims to construct the successor theory, $T_{\rm S}$: where the models are *not* instantiations, or representations, of the latter theory, but rather approximations to it. In particular, if both the reconstructed theory T and the successor theory $T_{\rm S}$ exist, then one expects that T can be obtained from $T_{\rm S}$ by making the appropriate approximations.

De Haro (2019a) contains detailed examples of the heuristic function of symmetries and dualities in quantum gravity. It also compares to other philosophical work on the heuristics of dualities, in particular the 'duality as gauge symmetry account', and to Rickles (2011, 2013, 2017). Van Dongen et al. (2020) discusses, in the context of black holes, the heuristics of emergence, correspondence, and the M theory programme. The more general topic of "the usefulness of philosophy for physics" (and vice-versa) is the topic of De Haro (2019c).

Chapter 6

Interpreting and Understanding Theories without a Spacetime

> *... a precipice not of absurdity but of a new understanding of old issues*
> (John Earman, 2002).

This Chapter has two aims: First, to briefly expound the close connection between interpretation (in the sense of classical referential semantics; more specifically, as in Chapter 1, as a partial map from a language to the world) and understanding, which has not received prior attention in the literature on scientific understanding. While the close connection between understanding and interpretation (in the general sense) has been acknowledged in the social sciences and humanities, it has been largely ignored in debates regarding understanding in the natural sciences. But we will see that interpretation is in fact essential in promoting understanding in modern theories of high-energy physics, so that it is worth exploring the connection, using the tools already developed. Second, the aim is to analyse the ways in which interpretations are constructed, and so the ways in which understanding can be had, in physical theories that do not contain a spacetime.

Current theoretical research suggests that space and time may not have the fundamental status that has traditionally been bestowed upon them. On the contrary, in recently proposed theories the fundamental notions of space and time seem to be absent; they are supposed to be replaced by alternative, non-spatiotemporal structures. An important question which arises, if such theories without spacetime are considered a live option, is how they should be interpreted. This is especially pressing for scientific realists, who regard theories as (ideally) true descriptions of reality and who will therefore ask: How can the real world possibly be the way a theory without spacetime says it is? Do space and time not exist at the fundamental level of reality? But the question is also urgent for those who do not regard themselves as realists. Constructive empiricists, for example, will need to know how theories without spacetime can yield empirically adequate models of the observable physical world, and moreover, what a literal construal of the theory would look like.

Understanding of phenomena, for example, has traditionally been regarded as requiring theories that are formulated in a spacetime framework, because this is a necessary condition for visualisation, which is regarded as an important—even indispensable—tool for understanding. (I can hardly do justice, in this short space, to the topic of visualisation in science and in philosophy: starting with Descartes and Kant, and on to quantum mechanics—where *Anschaulichkeit* was an important point of debate—and present-day physics.[1]).

So, how can physical theories in which space and time are fundamentally absent provide scientific understanding, if at all?[2] Again, the answer to this question depends on how theories without spacetime can or should be interpreted.

Section 6.1 reviews de Regt's contextual theory of understanding, and in its light discusses the connection between interpretation and understanding. Section 6.2 presents three tools for interpreting theories without spacetime. In the subsequent two sections, these tools are used to interpret a variety of theories. Thus section 6.3 focuses on theories for which visualisation via 'effective' spacetimes is possible, while section 6.4 deals with theories with interpretations that either completely resist visualisation, or do not require it for the problem at hand, so that they are interpreted in a non-visual manner.

6.1 The Contextual Theory of Understanding

In this section, I review my preferred account of understanding, viz. the contextual theory of understanding (from de Regt (2017)), and discuss the relation between interpretation and understanding further.

De Regt's (2017) contextual theory of understanding is highly suitable for our question, because it is attuned to scientific practice, especially the practice of the physical sciences, and because it acknowledges that scientific understanding of phenomena is related to the intelligibility of theories, where the latter is a contextual matter in the sense that it depends on the skills of scientists and the available tools for understanding. On this theory of understanding, I propose (and have argued in De Haro and De Regt (2018a)) that theories without spacetime *can* be intelligible, and accordingly there is no principled obstacle for achieving scientific understanding with such theories.[3]

The contextual theory of scientific understanding is based on the idea that scientists achieve understanding of a phenomenon P if they construct an appropriate model of P on the basis of a theory, T. More specifically, it contains the following criterion for scientific understanding (Criterion for Understanding Phenomena):

CUP: A phenomenon, P, is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory, T, and conforms to the basic epistemic

[1] For discussions of these topics, see de Regt and Dieks (2005), Beller (1999), Mössner (2018), De Haro and De Regt (2018a).

[2] A related problem is the problem of 'empirical incoherence'. For a discussion, see the special journal issue edited by Huggett and Wüthrich (2013).

[3] De Regt's (2017) theory of understanding of course does not use my account of scientific theories from Chapter 1. However, it will be helpful, in this Chapter, to recast his notion of 'understanding' in a form applicable to theories according to this account. This should also cast light on previous topics discussed in this thesis, and clarify the relation between understanding and interpretation.

values of empirical adequacy and internal consistency.[4]

The key term in this criterion is 'intelligible': understanding phenomena requires an intelligible theory, where intelligibility is defined as:

Intelligibility: the value that scientists attribute to the cluster of qualities of a theory that facilitate the use of the theory.[5]

A theoretical quality that is highly valued by scientists, past and present, is visualisability. Many scientists prefer visualisable theories because these are more tractable and easier to work with—in other words, visualisation is widely used as a tool for understanding (see de Regt (2014)). But the contextual theory of understanding does not entail that visualisability is a necessary condition for the intelligibility of scientific theories. It allows for alternative ways to render theories intelligible. Accordingly, although visualisation is an important tool for understanding, it is not indispensable. The fact that visualisation has proven to be a very effective tool in the past is not surprising, because visual skills are so widespread and well-developed among humans. This explains why visualisation is a dominant strategy for achieving understanding, but it does not follow that it is a necessary condition for it. Hence, visualisation is an often emphasised tool for understanding. One aim of the present Chapter is to present alternative tools for interpreting theories without spacetime in such a way that they can still be rendered intelligible.

Why do scientific theories need to be intelligible to the scientists who use them? This is because scientists understand phenomena by constructing *models* of them, and this involves pragmatic judgments and decisions, since models do not follow straightforwardly from theories (here, unlike in Chapter 4, I use the word 'model' in the usual sense in philosophy of science, i.e. as an approximated description of specific phenomena, usually in the context of a background theory). An objective test for assessing the intelligibility of a theory (for a scientist in a particular context) is provided by the following Criterion for the Intelligibility of Theories:

CIT$_1$: A scientific theory T is intelligible for scientists (in context C) if they can recognise qualitatively characteristic consequences of T without performing exact calculations.[6]

The basic idea behind CIT$_1$ is that the scientists have an 'insight' into the workings of the theory and are accordingly able to use the theory for the construction of models of the phenomena, which satisfy the basic values of empirical adequacy and internal consistency. The resulting models provide the explanations that produce understanding of the phenomena in question. The conceptual tools—e.g. visualisation—that scientists may use to recognise qualitative consequences of T are also the tools that will facilitate model construction.

The intelligible theory, T, is here to be understood as an interpreted theory, since a

[4] De Regt (2017: p. 92). This is a revised formulation of CUP as presented in De Regt and Dieks (2005: p. 150).

[5] De Regt (2017: p. 40).

[6] De Regt (2017: p. 102). The subscript in 'CIT$_1$' indicates that this is one among other possible criteria (tests) for intelligibility.

‘bare’ mathematical theory without an interpretation does not yield any predictions, let alone explanation or understanding. Only within an interpreted theory can one construct models of phenomena, in order to explain and predict. Recall, from section 1.2.3 (b), the close connection between our conception in terms of partial maps, i, of interpretation and the idea of models, as mediators between theory and phenomena. While not every interpretation, i, necessarily involves models in this sense, many of them do (as my examples in sections 6.3 and 6.4 will illustrate). More generally, interpretations proceed via the development of tools that allow scientists to recognise qualitatively characteristic consequences of the theory, T, without performing exact calculations (i.e. a case of CIT_1). I will present these tools, for the theories without a spacetime described here, in section 6.2. The idea is that, once an interpretation, $i : T \to D$, is constructed, scientists can use it to make qualitative and quantitative predictions: but, without an interpretation, a bare theory T makes no predictions. Thus, *interpretation is a precondition for intelligibility*, and different interpretations of T may lead to different degrees of intelligibility. The task of interpreting a theory should therefore not be taken as a simple and straightforward matter, especially because it should not be assumed that there is a unique interpretation for every scientific theory. The development of an interpretation of a theory is the first step in rendering the theory intelligible and in achieving understanding of phenomena (present or, indeed future) with the theory. The choice for a particular interpretation will be guided by preferences for specific tools for understanding. Indeed, an interpretation is itself built using specific interpretative tools, which in turn are tools for understanding, because interpretation leads to understanding.

Above, we discussed the construction of *models* as one way in which scientists understand phenomena, by linking them to the theory: models mediate between theory and phenomena,[7] thus rendering theories intelligible. But also: models explain the phenomena, thus providing understanding of them. As such, models of phenomena (e.g. the Bohr model of the atom and its interactions with photons) function as interpretations of theories, as claimed in section 1.2.3 (b).

6.2 Three Interpretative Tools

In this section, I identify three specific tools which are used by physicists to interpret theories, and in particular the theories without spacetime that concern us.[8] And the leading idea, in this Section and the next, will be that these three tools contribute to the physicists’ understanding of the theory through the development of an interpretation. It is the interpretation that is guiding us in understanding the theory, and in constructing models to explain particular phenomena. An interpretation is often, and traditionally, regarded as having a visual component (which is why visualisation is often emphasised). But we will see that an interpretation *need not* be visual.

[7] See Morgan and Morrison (1999).

[8] I use the term ‘tool’ deliberately, to emphasise that my approach to interpretation and understanding is a pragmatic one. While the tools that I discuss may traditionally be regarded as ‘methods’, I want to highlight how they function as conceptual tools that can be used to achieve desired ends, namely interpretations that render theories intelligible.

The three tools aim at (i) developing interpretations of theories, (ii) constructing suitably explanatory models of phenomena, and (iii) recognising qualitative characteristic consequences of theories without performing exact calculations.

All the tools that we discuss in this Chapter are “accordion concepts”, realised by very different sorts of procedures that only bear some family resemblances with one another. Thus I will not attempt to provide, and we will not need, analytic definitions of the tools. It will suffice, for our purposes, to characterise the tools, and to point out the resemblances between the examples.

The first two tools relate *theories*: the first tool relates two theories through an approximative relation, or linkage; the second tool relates them in a looser way (through a relation of similarity, or analogy). On the other hand, the third tool compares terms within a *single theory* (again, through similarity or analogy). I will also give (in Section 6.2.2) a variant of the first tool, in which the interpretive arrows are reverted.

6.2.1 (Approximation): Use of approximative relations between theories

It is common practice in physics to study theories in specific regimes of parameters, through various approximations or by taking limits (see Section 3.1.2). For example, in quantum field theory one studies the semi-classical limit (of a small value of the coupling constant), in which the theory reproduces the classical results, up to small corrections. In statistical mechanics, one studies the (unphysical) limit in which the number of particles and the volume of the system go to infinity, thus reproducing thermodynamics.[9]

So, (Approximation) is here used in the same sense as in Chapter 3. It includes a diverse spectrum of such, not mutually exclusive, procedures and linkages between theories: namely, the three ways, (i)-(iii) of Section 3.1.2, of obtaining a formal linkage (mathematical limits, comparing physical situations, mathematical aproximations). I will also use the notation $T_{\rm b}$ (for ‘better, bottom or basic’) and $T_{\rm t}$ (for ‘tainted, tangible or top’ theory). Alternatively, rather than considering the theories themselves, one considers specific phenomena (and thus specific *models* of the theories) and makes these approximations—the subsequent relation then still being one of idealisation or approximation. This linkage relation can be regarded as a surjective[10] and non-injective map between the theories, i.e. a map: $\text{Approx} : T_{\rm b} \to T_{\rm t}$.

The important point for scientific understanding is that the application of approximations, or linkage relations, discussed in Chapter 3, can also be used to induce a (partial) interpretation of a theory that had none, as I now argue. And so, this is how (Approximation) comes to be a tool for interpretation.

Assume that $T_{\rm b}$ is a bare theory whose interpretation one is developing, and assume that its approximating theory, $T_{\rm t}$ (say, general relativity), already has an interpretation

[9] This is regardless of whether the limit is taken to be an idealisation or a non-idealising approximation (for a discussion and references, see (i) in Section 3.1.2 and Section 3.1.3).

[10] Surjectivity is what is meant by the phrase ‘$T_{\rm t}$ reduces to $T_{\rm b}$ in the approximation’: $T_{\rm t}$ is what is left of $T_{\rm b}$ after the approximation is made, i.e. each element of $T_{\rm t}$ corresponds (“comes from”) at least one element of $T_{\rm b}$.

$$\begin{tikzcd}
T_{\mathrm{t}} \arrow[r, "i_{\mathrm{t}}"] & D_{\mathrm{t}} \arrow[d, dashed, "\text{heuristics}"] \\
T_{\mathrm{b}} \arrow[u, "\text{Approx}"] \arrow[ur, "i'_{\mathrm{b}}"] \arrow[r, "i_{\mathrm{b}}"] & D_{\mathrm{b}}
\end{tikzcd}$$

Figure 6.1: (Approximation): an inherited interpretation, $i'_{\mathrm{b}} := i_{\mathrm{t}} \circ \text{Approx}$, is obtained, and a more comprehensive interpretation, i_{b}, is constructed heuristically. Solid lines: approximation map (blue) and interpretation maps (orange). Dahsed line: heuristics.

i_{t}, i.e. it can be mapped to the world (denoted, in Figure 6.1, as D_{t}). Furthermore, we will assume that the interpretation is "sufficiently detailed" that each element in the domain of application of the theory, D_{t}, is an image of some element of the theory (an assumption already introduced in Section 3.1.1). So, we have two maps: $\text{Approx} : T_{\mathrm{b}} \to T_{\mathrm{t}}$ is the approximative relation, or linkage, between the theories. And $i_{\mathrm{t}} : T_{\mathrm{t}} \to D_{\mathrm{t}}$ is the interpretation map from the top theory to (a domain in) the world. Both maps are surjective. Then these relations induce a *new interpretation map* for the basic theory, T_{b} (for some of its terms at least), and I will dub this map the *inherited interpretation*, i'_{b}. Namely, it results from the successive application of the two surjective maps, the approximation and the top theory's interpretation: $i'_{\mathrm{b}} := i_{\mathrm{t}} \circ \text{Approx}$, and it is again surjective. The inherited interpretation is an interpretation because it maps the bottom theory to a domain of the world D_{t}, viz. $i'_{\mathrm{b}} : T_{\mathrm{b}} \to D_{\mathrm{t}}$ (and, because Approx is surjective, it is as empirically adequate as i_{t}).

An inherited interpretation, i'_{b}, of the basic theory is of course not a full-fledged interpretation of all the terms in the theory, but only a *starting point* for a more comprehensive interpretation, denoted as i_{b} in Figure 6.1. To see that more than an approximation is needed to interpret the full theory T_{b}, consider the following. First, denote as D_{b} the domain of the world that T_{b} could describe through an appropriate interpretation map, i.e. i_{b}. Since this basic theory is more detailed, D_{b} may be larger than T_{t}'s domain, D_{t} (i.e. the range of the inherited interpretation, i'_{b}): e.g. because it applies to more cases and solves more problems (recall that Approx is surjective). In other words, the inherited interpretation, i'_{b}, describes a smaller or more restricted domain than the basic theory, T_{b}, could describe. And so, what we are really after is an interpretation map to the domain D_{b}, i.e. $i_{\mathrm{b}} : T_{\mathrm{b}} \to D_{\mathrm{b}}$.[11] Nevertheless, the inherited interpretation, i'_{b}, may be a good

[11] There is no claim here that strictly $D_{\mathrm{t}} \subset D_{\mathrm{b}}$ (and we do not need such a strong condition). For Kuhn's incommensurability will, in the cases where it applies, imply that the full interpretation i_{b} *does not describe some of the elements of D_t*, which the inherited interpretation i'_{b} *did* describe. Rather, the claim I are making is that the inherited interpretation, i'_{b}, can be used as a tool for developing the full interpretation, i_{t}, because there is a *non-zero overlap*, i.e. $D_{\mathrm{t}} \cap D_{\mathrm{b}} \neq \emptyset$. This is only the claim that inconmensurability, though real, is not total: there are some elements in the world which were described

starting point for a full interpretation, $i_{\rm b}$. This is how (Approximation) can be used to interpret a theory, from another already interpreted theory.

This is of course how, for the most part, quantum mechanics is interpreted. For large enough systems, quantum mechanical uncertainties disappear, and abstract quantities such as the square of the wave-function can receive a straightforward interpretation (as a probability for some process to occur). Formally, this can be done by taking a limit, $\hbar \to 0$ (see the discussion in Section 2.2.2). For example, one can interpret a Gaussian wave-function as describing a 'delocalised particle', by considering the fact that in the limit $\hbar \to 0$, the wave-function becomes better and better localised, approaching a Dirac delta function in the position variable (what we call a classical 'point particle'). The interpretation of the Gaussian wave-function thus inherits its interpretation, $i'_{\rm b}$, as *de*localised *particle* from the case in which the particle is perfectly *localised* (and the point particle interpretation is $i_{\rm t}$). Ultimately, we may wish to do away completely with the concept of 'particle' in quantum mechanics, and so develop an interpretation $i_{\rm b}$ that does not refer to them: and such interpretations are indeed suggested by, or can be guessed from, the fact that $i_{\rm t}$ interpreted the delta function as a point particle and $i'_{\rm b}$ interpreted the Gaussian wave-function as a delocalised particle. If no such limit were available at all, so that we did not have the sequence from $i_{\rm t}$ to $i'_{\rm b}$, it would be much harder to try and work out $i_{\rm b}$.

It may appear from the above that the more approximations, with their inherited interpretations, one can take, the easier it is to construct a full interpretation $i_{\rm b}$, mapping all the physical elements of the theory to the world. But this need not be so. Think, for example, how in quantum mechanics there is a particle interpretation, which holds in a certain limit and for a certain set of examples; and a wave interpretation, which holds in other limits for other sets of examples. It is not easy to come up with an interpretation $i_{\rm b}$ that reconciles these two. So, different approximations may lead to mutually incompatible interpretations.

There is of course no necessary connection between (Approximation) and spacetime or visualisation. For example, in statistics, one may take the limit in which the number of e.g. individuals in a population is infinite, without the limit's generating any spacetime. Thus (Approximation) is a very general linkage relation, applicable to a wide class of theories, with or without a spacetime. But, since we are here interested in the question of spacetime visualisation, we will narrow down the scope of (Approximation) (and its cousin to be discussed next, (iApproximation)), and consider only examples that *do* generate spaces with geometrical structures.

Notice that *all* theories of quantum gravity ultimately aim at reproducing general relativity in an appropriate (Approximation). And so, one can strictly speaking not say that these theories contain no spacetimes at all. Rather, we need to distinguish between theories whose interpretation depends on their having a spacetime, and those whose interpretation is built using other tools: the latter are the theories without a spacetime. As we will see in Section 6.4, the interpretive arrows, for the latter kinds of theories, can be reversed—so that new light is cast on old spacetime theories by the new

by the top theory's interpretation $i_{\rm t}$, and which are still described by $i_{\rm b}$. And of course, $i_{\rm t}$ and $i_{\rm b}$, as interpretations, may well be inconsistent with each other (unlike $i_{\rm t}$ and $i'_{\rm b}$, which are consistent with each other).

$$\begin{tikzcd}
T_{\mathrm{t}} \arrow[r, "i_{\mathrm{t}}"] \arrow[dr, "i'_{\mathrm{t}}"] & D_{\mathrm{t}} \\
T_{\mathrm{b}} \arrow[u, "\text{Approx}"] \arrow[r, "i_{\mathrm{b}}"] & D_{\mathrm{b}} \arrow[u, dashed, "\text{reinterpretation}"']
\end{tikzcd}$$

Figure 6.2: (iApproximation): the top theory is reinterpreted using $i'_{\mathrm{t}} := i_{\mathrm{b}} \circ \mathrm{Approx}^{-1}$. Solid lines: approximation map (blue) and interpretation maps (orange). Dahsed line: heuristic reinterpretation.

theories that replace them. T_{t} does not help interpret T_{b}, but the other way around. We will call this an 'inverse approximative relation':

6.2.2 (iApproximation): Inverse approximative relations

Inverse approximative relations are cases of approximations in which the bottom theory T_{b} is used to interpret (usually: to *re*interpret) the *top* theory, T_{t}. The linkage relation $\mathrm{Approx} : T_{\mathrm{b}} \to T_{\mathrm{t}}$ is the same as before, but in constructing the interpretation we now use one of its inverses, $\mathrm{Approx}^{-1} : T_{\mathrm{t}} \to T_{\mathrm{b}}$. I say 'one of its inverses' because an inverse of the linkage relation between two theories need not be unique (indeed, we cannot expect it to be unique: but one will exist, because the approximative map is surjective).

Now imagine that we are in the possession of some interpretation of T_{b} not obtained through (Approximation), but through the other tools to be introduced in this Section, (Similar) or (Internal). Denote this interpretation as i_{b}, i.e. $i_{\mathrm{b}} : T_{\mathrm{b}} \to D_{\mathrm{b}}$ (see Figure 6.2). We obtain an inherited interpretation, i'_{t}, of the top theory through an inverse approximative relation, defined by the successive application of the inverse of the approximative map and the interpretation map: $i'_{\mathrm{t}} := i_{\mathrm{b}} \circ \mathrm{Approx}^{-1} : T_{\mathrm{t}} \to D_{\mathrm{b}}$.

The theories that we will consider in Section 6.4 replace spacetime by a more fundamental discrete structure underlying it. (iApproximation) reinterprets the spacetime of general relativity in light of the interpretation of the more fundamental discrete structure which replaces spacetime.

An analogy can be helpful here. The discovery that matter (described by theory T_{t}) is made of atoms, which can be classified by their atomic number and electron configurations (described by theory T_{b}), obviously compels one to rethink whatever interpretation of matter and of its interactions one held before making that discovery. For example, one is now able to interpret chemical reactions as the rearranging of the electrons in the orbitals of the atoms involved, to break old bonds and form new ones (so that i'_{t} says "reactions rearrange the atomic bonds"). In a similar way, the discrete structures underlying spacetime should shed new light on old aspects of spacetime, and so help develop new interpretations. For example, some physicists may interpret the fact that the underlying structure of spacetime is discrete, and that in some cases a continuum limit cannot be taken (because the mathematical limit does not exist), as saying that continuous spacetime is a mere appearance: namely, something that looks continuous to our senses but which ought to be thought of as discrete.

The difference between (Approximation) and (iApproximation), then, is in the direction of the interpretative (diagonal) arrows, and of the heurstic arrows, in Figures 6.1 and 6.2. Both emphasise the fact that interpretations are *dynamic*: as I mentioned in Section 1.2.3, interpretations are not unique and can change according to aims and—as we see here—in the face of new scientific theories.

How does (iApproximation) relate to visualisation? Even though a space with geometric structures is formally developed in the (iApproximation), this structured space is now to be reinterpreted starting from the discrete theory. So, the visualisation of the space is no longer such an important tool in (iApproximation): rather, the spacetime visualisation will now be *reinterpreted* in the light of the basic theory, which (in the cases of interest for our aims) cannot be so visualised.

Notice that, in practice, (Approximation) and (iApproximation) will often appear combined, in a hermeneutic circle, which can be described as follows. One typically starts with a tentative interpretation, i_{b}^0, of the basic theory, T_{b}. This interpretation could have been obtained either directly through the tools (Similar) and (Internal), or via (Approximation), from the spacetime interpretation of the top theory, T_{t}—general relativity, say. One then improves the interpretation, i_{b}^0, of the basic theory using (Similar) and (Internal). These tools lead to new interpretative aspects which are added to the tentative interpretation, i_{b}^0. At the same time, one may try to strip this interpretation, as much as possible, from any of the original spacetime elements that it may still have. The result of this is a new interpretation, i_{b}^1, of the basic theory. In turn, this interpretation may have consequences for, and thus call for a revision of, one's original interpretation of the top theory, T_{t}: specifically, it may call for a revision of the interpretation of spacetime in general relativity. Thus one constructs an inherited interpretation of T_{t}, viz. $i_{\mathrm{t}}^1 := i_{\mathrm{b}}^1 \circ \mathrm{Approx}^{-1}$, which should give an improved, or more fundamental, interpretation of the spacetime in general relativity (or any other spacetime theory that one started with). This process, which includes a combination of Figures 6.1 and 6.2, can continue until one is left with an interpretation of the basic theory that makes no direct reference to spacetime.

An interpretation that is obtained through (Similar) and (Internal), but without the use of (Approximation), may well be rather minimal. And then it may well be desirable to get a fuller interpretation through the use of (Approximation), e.g. to compare with the spacetime of general relativity. But, in such a case, we still insist that the two interpretative steps are conceptually distinct: and that even before one uses visualisation, one already can have a (minimal) interpretation of the theory. This illustrates the point that spacetime is not strictly needed for interpretation.

Strictly speaking, in the applications I am concerned with, (iApproximation) is not a tool for interpreting theories without a spacetime: but a tool for reinterpreting general relativity (or other spacetime theories) from the basic theories without a spacetime underlying them. And so, if I am correct, it predicts that a good theory of quantum gravity would lead to a reinterpretation of general relativity's spacetime interpretation.

6.2.3 (Similar): Similarities in the use of concepts between theories

The first tool just discussed, (Approximation) (and its inverse relation (iApproximation)), involved a relation of linkage between theories, i.e. a one-way entailment. The second tool, (Similar), involves a relation of *similarity between different theories*. Similarities between theories can indeed suggest ways of interpreting the concepts or terms in a theory. Similarities will usually relate individual terms between theories, and how these are related to the overall theoretical structure of a theory.[12]

Notice that similarity between different theories comprises a spectrum of likeness, ranging from (almost) identity to analogy.[13] Thus (Similar) will admit the same varieties.

The similarities between two interpreted theories can be either in the bare theory, i.e. in the formalism, or in the interpretation. Thus there are two variants of (Similar):

(i) *Formal similarity:* a similarity between parts of the mathematical formalism of two theories can be used to draw consequences about the interpretations of terms in the bare theory (mathematical formulas, concepts, etc.). That is, a given formal similarity between two theories is used to construct a "matching" interpretation for one of the two theories.

Formal similarity was already considered in Section 2.2.2 as a specific kind of correspondence between theories (see Post (1971) and Radder (1991)).

An example of formal similarity is that of conservation laws in Maxwell's theory, general relativity, and quantum mechanics (and naturally, the similarity is not *only* formal, but also interpretative!). The electrical charge and current in Maxwell's theory play the role of sources for the electric field, and they are associated with a conservation law that follows as an identity from Maxwell's equations. This general idea (of being a source for a field, and of being associated with a conservation law which expresses a certain identity that follows from the equations of motion) can then be applied to other theories, like general relativity, where the stress-energy tensor plays a very similar role: of a source, which satisfies a conservation law, which follows from Einstein's equation. The expression of this formal similarity is of course Noether's two theorems, which apply to a wide class of theories (for the case of general relativity, see Brading and Brown (2003)).

(ii) *Interpretative similarity:* a similarity between some aspects of the interpretations of two theories can be used to construct a fuller, or more complete, interpretation of one of the theories, matching the interpretation of the other theory.

A specific example of this kind of similarity, namely similarity between concepts, was already discussed in Section 2.2.2. As already remarked there, formal and interpretative

[12] All the tools discussed in this Chapter can be used for theories, for theory parts, and for models. In what follows, what I say about theories holds for these other cases as well: but I will simply use the word 'theories'.

[13] The 'almost' here is because when we compare *different theories*, as we do in (Similar), there will always be some disanalogies, somewhere in the interpretation of the theoretical structure. There will never be a rule that automatically tells us, from the interpretation of an old theory, how a new theory ought to be interpreted. One should always look out for surprises!

similarity usually go hand-in-hand, because interpretative similarity can help unravel a formal similarity, so that technical-mathematical progress can be made.[14]

6.2.4 (Internal): Internal criteria

The third tool involves a *single theory of model.* More precisely, it is about *internal criteria* which can be used to interpret the concepts or terms in the bare theory. As in the case of (Similar), these criteria can vary: from structural similarity, to analogy. Also, as in the case of (Similar), the internal criteria can be of two types: formal or interpretative similarity. Formal or mathematical similarity between two parts of the theory is used to draw consequences about the interpretations of certain expressions. Interpretative similarity is used to interpret uninterpreted terms in the theory, or to further develop an existing interpretation.

For example, in quantum mechanics, the fact that certain quantities (such as the standard deviation of a distinguished operator, divided by its time change) are formally conjugate to the energy, in the sense that they satisfy Heisenberg's uncertainty principle together with the energy, has led some to take these quantities as measures of uncertainty in time.[15] The reasoning here is from structural similarity within the theory of quantum mechanics itself: if position and momentum satisfy certain structural relations, and energy and some other quantity satisfy the same structural relations, then one is entitled to take this other quantity as a measure for the uncertainty of time (which is the expected quantity to satisfy those structural relations).

(Internal) will often be applied *after* external methods like (Approximation) or (Similar) have been applied. Once some terms in a theory receive a tentative interpretation through external methods, internal comparison can lead to the interpretation of other, still uninterpreted, terms in the theoretical structure. For example, this is how non-observational quantities (quantities that are not observable according to current methods, or quantities that will perhaps always remain out of reach) can receive an interpretation: formally, from their relation with other quantities which already have an interpretation. This is the case with for example dark matter and dark energy, which currently cannot be observed directly (and perhaps they never will be) but which were postulated in order to explain other phenomena. Comparing the rotation curves of galaxies to the predictions of Newton's theory, the missing dark matter, if it exists, can be interpreted as interacting gravitationally, and its mass can be calculated.

Internal criteria are interpretative criteria for dualities, as I also argued in Chapters 4 and 5. Equivalent theories can be seen as 'models (in the sense of instantiations) of the same theory', and an interpretation of a theory can be worked out that starts from the common core between the models. Though they involve different models, such interpretations are cases of (Internal) not (Similar), because the criteria of interpretation

[14] For the example of the billiard ball model for the kinetic theory of gasses, see Hesse (1966: pp. 4-9). For an example in quantum mechanics, see De Haro and De Regt (2018: p. 648).

[15] Of course, a theorem by Pauli proves that there is no self-adjoint operator representing time, if the Hamiltonian is bounded from below. But the intended interpretation is only that a specific quantity can be taken to represent uncertainty in time, not that time is described by a self-adjoint operator, or that the Hamiltonian is bounded from below. See Pauli (1958), Busch (2008).

are internal to the theory.

6.3 Visualisation via Effective Spacetimes

In this Section, I discuss theories in which (Approximation) is used as a tool to develop a *visual* interpretation, i.e. an interpretation that employs space (and perhaps time). Thus, in the use of (Approximation) we will consider here, either a new spacetime appears in the theory, or an already existing spacetime is transformed in some way that is useful for interpreting the theory, thus rendering it intelligible (in the sense of Section 6.1).

I will first introduce two distinctions that broadly characterise the effective spacetimes that we will be dealing with:

(a) *Auxiliary, visualisable spacetimes, vs. physically real spacetimes.* This distinction is prompted by the idea that 'visualisable' does not entail 'physically real', and so one should distinguish between these two types of spacetimes, both of which are effective.

The general question, of how visualisation in spacetime is used to construct an interpretation, is logically independent from the question whether physicists consider a spacetime to be physically real and-or observable: hence our term, 'effective' spacetimes. Visualisation helps rendering a theory intelligible, even if the space is not considered to be real. For example, in the method of image charges in electrostatics, a boundary condition (such as the vanishing of the electric potential on a plane bounding the problem) is replaced by a (set of) image charge(s). This is a visual method (an interpretation) because it allows us to picture the otherwise rather abstract boundary condition imposed on the potential (and, more importantly, it allows to do calculations in an intuitive way, which would otherwise be rather more cumbersome). But of course, the imaginary charges that are visualised on the other side of the plane are not physically real. Rather, we compare our world with another *possible world* which is much easier to visualise, and in which for each charge there is a mirror charge of opposite sign. In the cases we discuss in this Section, it is *spacetime itself* that may not be physically real, but auxiliary—like the imaginary charges.

Such auxiliary, or effective, spacetimes are allowed in my conception of interpretation as given by maps i: for, as we remarked in Section 1.2.3, interpretation maps not only map to the real world, but often also to other possible worlds—as in the case of the imaginary charge, which is not real, yet its use hinges on its interpretation and properties as a 'charge'.

(b) *Physical vs. merely mathematical spaces.* But despite the above remark, that visualisation is logically independent of the question of physical reality: a second distinction needs to be made. Namely, whether a space (or spacetime) is interpreted as *physical* (in the actual world, or in some possible world) or as merely *mathematical*, is an important question, which we will address. So, when I refer to 'theories without spacetime', I have in mind the former meaning of spacetime. This is because it is very hard (and it would be contrived) to formulate a theory that does not make any use at all of the mathematical concept of space, i.e. of a set with some added structure.[16] For example, the numeri-

[16] For an overview of Quine and Putnam's indispensability argument, see Colyvan (2019). For a nomi-

cal parameters of physical theories all take values in fields, in the algebraic sense, which are examples of spaces. As soon as we have a quantity taking values in $\mathbb{R}$, we have a space. It would surely be unreasonable to demand, on the grounds that we must have a theory without spacetime, that $\mathbb{R}$ may not appear anywhere in our theories—at the very least, such a revisionist project is not the kind of project that researchers have in mind when they talk about theories 'without a spacetime'. So, in what follows, our concept of space will be physical not merely mathematical, and when I talk about 'visualisation' I will mean visualisation using a space with physical properties (such as particles, fields or forces defined on it). And I submit that visualisation using physical spaces (whether in our world, or in another possible world) contributes to the intelligibility of the theory more than visualisation via merely mathematical spaces, because in the former case physicists can use physical arguments and intuitions, in addition to mathematical ones.

Agreed, the distinction between merely mathematical and physical spaces can sometimes be blurry, if the mathematical spaces come with sufficient structure (as, say, normed vector spaces, or even inner product spaces). For such a mathematical space may, if appearing within a physical theory, easily inherit a physical interpretation (in the way described in Section 6.2). But this need not always be so: for example, the geometric interpretation of phase space in classical mechanics, as a symplectic manifold, is not physical: for the phase space is itself not deemed to be physically real in a geometric sense, under the standard interpretation of classical mechanics. And yet it provides a visualisation which helps physicists to apply concepts such as the existence of forms on this space, the independence of the choice of coordinates, etc., which helps them better grasp certain aspects of the theory (such as e.g. canonical transformations). But in this Chapter we will not be concerned with such cases of mathematical but not physical spaces, since what is at stake in theories of quantum gravity is of course spacetime itself, and how theories without a physical spacetime can be interpreted.

Summarising this discussion, (a)-(b): the topic of visualisation in scientific understanding should in general also consider spaces which are mathematical and not physical. But for our specific question, of how theories without spacetime can be interpreted, only the question of whether there is a *physical* space is relevant, i.e. (b)—but such a space need not be physically real, i.e. can be effective à la (a).

My main example in this Section is a model that does not start off with a spacetime, but which, in the approximation taken, gives rise to a curved two-dimensional surface with holes, viz. a Riemann surface. For an introduction to random matrix models aimed at philosophers, and a discussion of the details, see De Haro (2018b).

The idea is as follows: the quantities of the theory are non-abelian *matrices* (like in Yang-Mills theory), but without any space or time dependence, and consequently with no spacetime dynamics—there is no kinetic term in the Lagrangian: only a potential term. In this way, one gets what is called a 'random matrix model', i.e. a model in which the fields are just $N \times N$ matrices, with their eigenvalues distributed according to an appropriate probability density (given by the Lagrangian) but with no spacetime dependence.

nalism that does not quantify over mathematical entities, see Field (2016).

Because of symmetry, the physical quantities do not depend on all N^2 components of the matrix, but only on its N eigenvalues. And so, the dynamics of this model is essentially the dynamics of the eigenvalues, subject to a potential. In principle, these eigenvalues can take any complex values, and so they correspond to N points on the complex plane. However, in the saddle-point approximation, the potential imposes a "zero-force condition", and the eigenvalues tend to fill up parts of the complex plane (physicists speak of *condensation of the eigenvalues*), forming branch cuts. One can show that, in the relevant approximation, this 'plane with branch cuts' is best described as a Riemann surface with non-trivial cycles: some cycles are around the branch cuts, and others connect different branch cuts.

In other words: one started with a theory in which there were only matrices and no spacetime. Then, in the appropriate approximation (namely, taking N to be large), a non-trivial geometry emerges out of the eigenvalue condensation. It is this geometry that is extremely useful for the intelligibility the dynamics of the random matrix model. The physical quantities turn out to have a topological expansion in terms of the increasing genus of Riemann surfaces.

The (Approximation) allows visualisation, because it fabricates a two-dimensional Euclidean space (there is no time) in which the theory is easy to visualise, and therefore easier to interpret. The random matrix model itself does not have an initial spacetime: it is an algebraic structure. Thus, the Riemann surface that one ends up with is a *novel geometric structure*, within a theory which had no initial spacetime.

6.4 No Spacetime and No Visualisation

Section 6.3 focussed on interpretations that are constructed by developing an effective spacetime in a specific (Approximation). This effective spacetime then allows physicists to visually interpret the theory. In this Section, we move on to cases in which visualisation and effective spacetimes are not needed, and not used as the primary interpretative tools—even if spaces might be present somewhere else in the theoretical structure. Rather, other tools, (Similar) and (Internal) from Section 6.2, are used, which do not require a spacetime interpretation.

6.4.1 Merely mathematical spaces and theories without spacetime

In this subsection, I collect my remarks (a)-(b), from the preamble of Section 6.3, to work towards a simple conception of a 'theory without spacetime'—which I will give at the end of the subsection; as well as the related notion of a 'merely mathematical or pre-geometric space'.

Recall my two distinctions, from the preamble of Section 6.3, which bear on the kinds of spacetimes used for interpretation:

(a) an *auxiliary spacetime* with visualisation vs. a *physically real* spacetime;

(b) *physical space(time)* vs. *mathematical or pre-geometric space.*

For the project of this Section, we do not want to allow interpretations that entail spaces that are *physical*, either in the sense of ‘physically real’ (e.g. the space of general relativity), or in a fictitious, but still physical sense (e.g. phase space). Therefore, in this Section I focus on examples that do not contain any of the spaces of the disjunction (a), i.e. not any auxiliary spacetime à la (a) first part and not any physically real spacetimes à la (a) second part, and contain only the *mathematical spaces* of the disjunction (b). I will dub such spaces ‘merely mathematical’, and give a more explicit conception of them below.

But I first make two warnings about merely mathematical spaces, in relation to the tools (Similar)-(Internal) which I will illustrate in the next two subsections. The first warning is about the tools to be used in theories without spacetime; the second is about visualisation of merely mathematical spaces:—

(1) *(Similar) and (Internal) can be had in theories without spacetime, despite the presence of a spacetime somewhere in the theory.* This is because my claim, that we are dealing with ‘theories without spacetime’, does not entail the absolute non-existence of any sort of physical space anywhere in the theories and models I will present. Ultimately, *every* realistic theory of quantum gravity attempts to recover the pseudo-Riemannian spacetime of general relativity, typically in an (Approximation) or an (iApproximation): and so, every theory without spacetime, attempting to describe the world, must at one point or another recover a physical space. But remember what my no-spacetime claim refers to: not to the *presence*, but to the *use* of spacetime as a tool for interpretation. More specifically, the claim is that these theories do not use (Approximation) when they develop their approximations; and that these theories do illustrate the use of (Similar) and (Internal) without reference to spacetime or visualisation. They also sometimes illustrate (iApproximation), with its inverse relation between bottom theory and spacetime visualisation. So, the title of this Section—‘no spacetime and no visualisation’—should be read as relative to the tools which the interpretation of the theory actually uses.

(2) *Merely mathematical or pre-geometric spaces can be free from spacetime visualisation.* I should say a bit more about our distinctions (a)-(b) recalled above, between physical and merely mathematical spaces, even if I do not attempt at a systematic treatment of this question here (nor will I need it: for a discussion of a concrete case, cf. De Haro (2018b: §4.2)). Indeed, between a finite set of numbers, a set of points with the cardinality of the continuum, and a manifold endowed with a metric, there is an almost continuous spectrum of possibilities for what we can call ‘merely mathematical’ vs. ‘physical spacetime structure’. Whether a space qualifies as ‘merely mathematical or pre-geometric’ or as ‘physical’ depends on how strongly it is physically interpreted. Accordingly, I will call these two cases ‘merely mathematical or pre-geometric’ vs. ‘physical’ spaces. And of course, even in the case of a set consisting of three elements, one can use visualisation, viz. three encircled dots. But it would be foolish to mistake such a visualisation for a counterexample to the claim that we are dealing with a theory without spacetime. For, first: such a visualisation is not needed. And second: the picture of the encircled dots is

rather more a mnemonic, or a visual way to fix ideas, than providing a spacetime interpretation of a set consisting of three elements. It is not *spacetime visualisation.*

With these two comments in mind, and motivated by my earlier notion of visualisation, I now settle for the following notion of a *merely mathematical or pre-geometric space*, viz. as either: (A) a discrete (finite or infinite) space, or (B) a continuous space, but with no rich geometric structures on it (e.g. only topological structure).[17]

Theories without spacetime will consequently, for our purposes in this Section, be theories that use merely mathematical or pre-geometric spaces, and no physical spaces, for interpretation: indeed, physical spaces are not used in the specific interpretations that I will review from the literature.

How will merely mathematical spaces figure in theories without spacetime, and how do they relate to the physical space which emerges in (Approximation)? Remember that the merely mathematical or pre-geometric spaces are either discrete, or continuous but without geometric structure. In the three cases we will study, such spaces will constitute part of the basic structure of the theory (e.g. they go into the fundamental degrees of freedom which are then quantised) but they are pre-geometric, in that they have no straightforward geometric interpretation. These spaces will underlie the physical, pseudo-Riemannian space of general relativity, which will be recovered with (iApproximation). But crucially, as I stressed in point (1) of Section 6.4.1, physicists who work on the project of interpreting a theory without a spacetime, i.e. of giving a deeper interpretation to a theory from which a spacetime theory may later be derived as an approximation, do *not* use (Approximation) to develop key parts of that deeper interpretation. They sometimes use it as part of a hermeneutic circle as in Section 6.2.2, but—crucially—the basic theory is purified of its geometric connotations. Thus, the correct relation between the merely mathematical space and the physical space is that the former underlies—in fact, replaces—the latter; and that the latter is derived from and emerges from the former, in an (iApproximation). Physical space is derived, explained and, ultimately, interpreted, from merely mathematical space plus an (iApproximation): and not the other way around.

6.4.2 Spin Foams

In this Section I will give an example from spin foams, an approach which arose from one of the quantisation programmes of gravity: namely, loop quantum gravity. Due to space constraints, I cannot give details here about even the most elementary aspects of

[17] The phrases 'merely mathematical' and 'pre-geometric' should here not be interpreted as normative statements recommending ways to *construct* theories without a spacetime, nor as suggesting that the structures referred to are unphysical. Rather, the terminology aims to clarify how physicists often *interpret* spacetime theories. The statement is that merely mathematical or pre-geometric spaces are not fully *spacetime* structures, in the innocuous sense of their supporting the usual continuous structures that we associate with spacetime theories in physics: light-cones, metric and-or other fields, etc. This aspect is brought out by several of the *authors themselves.* For example, Oriti (2016) calls such discrete structures 'pre-geometric, non-spatiotemporal structures', thus clearly implying that they are not fully spatio-temporal structures. Rovelli and Vidotto (2015) use the stronger phrases 'the end of spacetime', 'physics without time', etc., again stressing that these structures are not what physicist usually conceive of by 'spacetime' in physics.

the theory. I refer the interested reader to Rovelli and Vidotto (2015).

I will here illlustrate two aspects about loop gravity: (a) Its being a theory without spacetime, under the conception of 'theory without spacetime' introduced in Section 6.4.1. (b) How to develop an interpretation of the theory, physicists use (Similar) and (Internal).[18]

The first point—loop gravity's being a theory without spacetime—is already illustrated, for example, in the titles of some of the Sections of Rovelli and Vidotto (2015): 'The end of space and time' (§1.2), 'Fuzziness: disappearance of classical space and time' (§1.4.2), 'Physics without time' (§2). Also explicitly in their wording (which actually also emphasises the topic of understanding):

> The description of spacetime as a (pseudo-) Riemannian manifold cannot survive quantum gravity. We have to learn a new language for describing the world: a language which is neither that of standard field theory on flat spacetime, nor that of Riemannian geometry. We have to understand what quantum space and what quantum time are. This is the difficult side of quantum gravity, but also the source of its beauty. (p. 19).

This verbal description is illustrated technically in the Section immediately following (§1.3), about quantised geometry. And here, as in our remark (1) in Section 6.4.1, we should bear in mind the fact that the programme is one of replacing the classical geometry of general relativity by a discrete structure—a merely mathematical space—which underlies physical space.

The merely mathematical or pre-geometric space that lies at the basis of the development of loop gravity is a tetrahedron. A tetrahedron can be parametrised by a set of four vectors, $\{\mathbf{L}_a\}$, or $\{L^i_a\}$, one for each face of the tetrahedron (where the index $a = 1, \ldots, 4$ labels each of the faces, and $i = 1, 2, 3$ are spatial indices). Since this is a quantum theory, the vectors L^i_a are quantised: they are taken to satisfy the angular momentum algebra of SU(2). The areas and volumes constructed from these basic objects thus turn out to be quantised.

Notice that spacetime is to be explained and derived from these objects, and not the other way around. In fact, it seems that the limit of an infinite number of quanta does not even exist. In any case, the geometric intuition is of little help here: for these are discrete quanta of space, fuzzy objects whose values cannot be determined simultaneously. And so, (Approximation) is here not the primary aid.

Another way to say this is as follows. The classical picture one starts with is one of a piece of space with a tetrahedron. This picture is only used to identify the variables to quantise (since the theory is supposed to reproduce general relativity in the right (Approximation)). But the interpretation of the quantum theory does not start from the classical interpretation; it starts somewhere else, namely from the quantum theory itself,[19] as the rest of the Section makes clear.

[18] Remember my remark (1) in Section 6.4.1, that even if a spacetime must be obtained in an (Approximation), this is no objection to the development of already a simple interpretation of the theory. This is because, as I said in the last paragraph of Section 6.4.1, the explanatory relation between merely mathematical spaces and physical spaces is reversed in these theories: it is the former that explain the latter.

[19] This is then another case of a hermeneutic circle.

A useful tool to be used here is (Similar): Rovelli and Vidotto (2015) exploit the analogy with quantum mechanics. The algebra of the L^i_a's is a slight generalisation of the usual SU(2) angular momentum algebra, familiar from elementary quantum mechanics. Indeed, most of the exercises given in their first chapter as practice for the reader are about SU(2), spin, the Pauli matrices, and the angular momentum variables. They are *not* about spacetime or general relativity. It is evident that the authors are drawing on the reader's knowledge of quantum mechanics to become familiar with this case, rather than drawing on a visualisation in spacetime.

(Internal) is also used, for example, in discovering the role of Newton's constant in this theory. Once the algebra of the theory has been related to the angular momentum algebra of quantum mechanics, an interpretation can be developed for the analogue of Planck's constant that appears in the algebra as a 'quantum of area', since the area (and volume) operators turn out to be quantised in units of a fundamental length scale. At this particular point, a comparison with general relativity *is* being used, in order to interpret the eigenvalues of these operators as 'areas' and 'volumes'.[20] Nevertheless, it seems that we have here a case of the hermeneutic circle mentioned in Section 6.2: an initial interpretation as 'area' or 'volume' is corrected, in (iApproximation), by the appearance of an underlying discrete algebra that replaces what we usually call areas and volumes. Areas and volumes are no longer to be interpreted as classical geometric quantities, but rather as classical limits of operators with a discrete spectrum, and in terms of states that can be in superpositions of eigenvalues.

(Similar) is also used to develop interpretations that borrow from theories other than elementary quantum mechanics: thus quantum field theory techniques are used in their Section 1.4. The interpretation emphasises the phase transitions that might occur, and that these are independent of the microscopic cut-off.

> The system is characterised by a physical and *finite* cut-off scale—the atomic scale—and there are no modes of the [iron] bar beyond this scale. The bar can be described as a system with a large but *finite* number of degrees of freedom. The short-distance cut off in the modes is not a mathematical trick for removing infinities, nor a way for hiding unknown physics: it is a genuine physical feature of the system. Quantum gravity is similar: the Planck-scale cut-off is a genuine physical feature of the system formed by quantum spacetime (p. 31).

An alternative, more modern, way to construct these pre-geometric structures are the so-called *spin foams*. The construction is similar to that of loop gravity, but now the dual links of a triangle on the plane (rather than a tetrahedron) are quantised. Again, at this point this is a merely mathematical space that is going to be quantised. The Hilbert space is then L^2 on the SU(2) associated with the rotations about each link, up to gauge transformations (rotations about the nodes). The transition amplitudes constructed from concatenating multiple copies of these objects in an approate way (using the Feynman path integral), form a spin foam.

[20] For this reason, loop gravity seems more tied with a classical spacetime interpretation than does group field theory.

Oriti (2013: §III.A) emphasises the fact that the Hilbert space built as above does not contain *geometric* data, but rather 'discrete combinatorial structures (the graphs), labelled by algebraic data only.'

This point is also emphasised by Markopoulou: 'These approaches start with an underlying microscopic theory of quantum systems in which no reference to a spatiotemporal geometry is to be found.' (p. 129). There is also the fact that mathematical spaces can also have other, non-spatiotemporal interpretations, obtained via (Similar)-ity. For example, there is a formal similarity with circuit models of quantum computation which can, in some sense, be used:

> While [the graph] has the same properties as a causal set, i.e. the discrete analog of a Lorentzian spacetime, it does not have to be one. For example, in the circuit model of quantum computation, a circuit, that is, a collection of gates and wires also has the same properties as [the graph] and simply represents a sequence of information transfer which may or may not be connected to spatiotemporal motions (Oriti (2013: p. 136)).

The tools available here are again (Similar) and (Internal). And as before, mathematical structures and techniques, familiar from quantum mechanics and quantum field theory, are used effectively to develop intuition and perform calculations. Hence the (Similar)-ity with quantum mechanics and quantum field theory are used. Once these structures are understood, (Internal) is used to shed light on other parts of the structure.

Among the mathematical techniques used from quantum mechanics and quantum field theory, are: harmonic analysis on the group SU(2), the Haar measure on the group, the Clebsch-Gordan coefficients, Wigner's 3j- and 6j-symbols, the representation theory of groups, spinors. To get the classical (Approximation), coherent states are obtained.

For example, A. Perez, 'The spin foam representation of loop quantum gravity' explicitly illustrates the analogies, which amount to the use of (Similar):

> A sum over gauge-histories in a way which is technically analogous to a standard path integral in quantum mechanics. The physical interpretation is however quite different... The spin foam representation arises naturally as the path integral representation of the field theoretical analog of P [a projector operator] in the context of loop quantum gravity (p. 275).

6.5 Scientific Understanding: Summary and Further Work

Theories in which there is no spacetime pose an obvious question: how should such theories be interpreted? My preferred account of interpretation, in terms of suitable maps, was argued to be suitable to answer this question. An interpretation was argued to be a precondition for the intelligibility of the theory. Intelligibility is, in turn, a necessary condition for achieving understanding of the phenomena, according to De Regt's contextual theory of understanding. Depending on the context, different conceptual tools may be used for constructing interpretations that render a theory intelligible.

I classified three conceptual tools which physicists have at their disposal to interpret theories, with or without a spacetime. (Approximation) relates two theories through an approximative scheme, and thus develops an interpretation. (Similar) draws on similarities (formal, or conceptual) with other already known theories. (Internal) derives its interpretation from the theory itself.

My main examples illustrated how these tools are used in actual practice. Spin foams and group field theory develop interpretations that do not need an appeal to a spacetime (or, in some cases, have no spacetime at all). Nevertheless, one has to admit that all these approaches retain traces of spacetime *connotations* in their interpretations: which illustrates the fact that visualisation, though not necessary, is an often emphasised tool. I characterised the development of such interpretations as a 'hermeneutic circle', in which interpretations are stripped, where possible and required, of those connotations.

The general relation between visualisation and understanding, and the question of whether theories without a spacetime can provide scientific understanding, is discussed in De Haro and De Regt (2018a). This paper also reframes a well-known debate between John Earman and Tim Maudlin about the status of time and change, in general relativity, in terms of scientific understanding—thus further underlining the importance of the subject. We further the notion of 'understanding' in quantum gravity in De Haro, van Dongen et al. (2020) and van Dongen et al. (2020).

De Haro and De Regt (2018) gives more examples of scientific understanding for theories without a spacetime: planar diagrams in quantum field theory and in random matrix models (pp. 652-657), causal sets (pp. 659-663), and group field theory (pp. 667-668).

Bibliography

Aharony, O., Gubser, S. S., Maldacena, J. M., Ooguri, H. and Oz, Y. (1999). 'Large N field theories, string theory and gravity'. *Physics Reports*, 323, 2000, p. 183.

Anderson, P. W. (1972). 'More is Different'. *Science*, 177 (4047), pp. 393-396.

Anderson, P. W. (1989). 'Theoretical Paradigms for the Sciences of Complexity'. In *A Career in Theoretical Physics*, Anderson, P. W., 2005, Wold Scientific, 2nd edition.

Bain, J. and Norton, J. D. (2001). 'What Should Philosophers of Science Learn from the History of the Electron?' In: *Histories of the Electron. The Birth of Microphysics,* Buchwald, J. Z. and Warwick, A. (Eds.). Cambridge, MA: MIT Press.

Banks, T., Fischler, W., Shenker, S. H. and Susskind, L. (1996). 'M theory as a matrix model: A Conjecture', *Physical Review* D, 55, 1997, p. 5112. doi:10.1103/PhysRevD.55.5112 [hep-th/9610043].

Barrett, T. W. (2018). 'Equivalent and Inequivalent Formulations of Classical Mechanics'. PhilSci: http://philsci-archive.pitt.edu/13092.

Barrett, T. W. and Halvorson, H. (2016). 'Glymour and Quine on theoretical equivalence'. *Journal of Philosophical Logic*, 45 (5), pp. 467-483.

Batterman, R. (2002). *The Devil in the Details.* Oxford University Press.

Bedau, M. A. (1997). 'Weak Emergence'. *Philosophical Perspectives*, 11, pp. 375-399.

Bedau, M. A. and Humphreys, P. (2008). 'Emergence: Contemporary Readings in Philosophy and Science.' Cambridge, MA: The MIT Press.

Beller, M. (1999). *Quantum Dialogue. The Making of a Revolution.* Chicago and London: The University of Chicago Press.

Belot, G. (1995). 'Determinism and Ontology'. *International Studies in the Philosophy of Science*, 9 (1), pp. 85-101.

Black, R. (2000). 'Against Quidditism'. *Australasian Journal of Philosophy,* 78:1, pp. 87-104.

Bokulich, A. (2008). *Reexamining the Quantum-Classical Relation. Beyond Reductionism and Pluralism.* Cambridge: Cambridge University Press.

Brading, K. and Brown, H. R. (2003). 'Symmetries and Noether's Theorems'. In: *Symmetries in Physics. Philosophical Reflections,* Brading, K. and Castellani, E. (Eds.), pp. 89-109.

Broad, C. D. (1925). *The Mind and its Place in Nature.* London: Kegan Paul, Trench, Trubner.

Brown, H. R. (1993). 'Correspondence, Invariance and Heuristics in the Emergence of Special Relativity'. In: *Correspondence, Invariance and Heuristics,* S. French and H. Kamminga (Eds.). Dordrecht: Springer.

Butterfield, J. (1988). 'The Shaky Game: Einstein, Realism and the Quantum Theory, by Arthur Fine'. *Mind,* New Series, 97 (386), pp. 291-295.

Butterfield, J. (2011). 'Emergence, reduction and supervenience: a varied landscape', *Foundations of Physics*, 41 (6), pp. 920-959.

Butterfield, J. (2011a). 'Less is different: emergence and reduction reconciled'. *Foundations of Physics*, 41 (6), pp. 1065-1135.

Butterfield, J. (2020). 'On Dualities and Equivalences Between Physical Theories'. Forthcoming in *Space and Time after Quantum Gravity,* Huggett, N. and Wüthrich, C. (Eds.).

Busch, P. (2008). *Time in Quantum Mechanics,* pp. 73-105. Berlin, Heidelberg: Springer.

Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Castellani, E. and De Haro, S. (2020). 'Duality, Fundamentality, and Emergence'. Forthcoming in *The Foundation of Reality: Fundamentality, Space and Time*, Glick, D., Darby, G., Marmodoro, A. (Eds.), Oxford University Press.

Castellani, E. and Rickles, D. (2017). 'Introduction to special issue on dualities'. *Studies in History and Philosophy of Modern Physics*, 59: pp. 1-5.

Carnap, R. (1947). *Meaning and Necessity*, Chicago: University of Chicago Press.

Chakravartty, A. (2007). *A Metaphysics for Scientific Realism. Knowing the Unobservable.* Cambridge: Cambridge University Press.

Chakravartty, A. (2017). 'Scientific Realism'. *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/scientific-realism.

Chalmers, D. J. (2006). 'Strong and weak emergence'. *The reemergence of emergence*, pp. 244-256.

Chang, H. (2012). *Is Water H_2O?* Springer Dordrecht Heidelberg New York London.

Chang, H. (2018). 'Is Pluralism Compatible with Scientific Realism?' In: Saatsi (2018), pp. 176-186.

Coffey, K. (2014). 'Theoretical Equivalence as Interpretative Equivalence'. *The British Journal for the Philosophy of Science*, 65, pp. 821-844.

Colyvan, M. (2019). *Indispensability Arguments in the Philosophy of Mathematics.* Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/entries/mathphil-indis.

Corfield, D. (2017). "Duality as a Category-Theoretic Concept". *Studies in History and Philosophy of Modern Physics*, 59, pp. 55-61.

Corkum, P. (2008). 'Aristotle on Ontological Dependence'. *Phronesis*, 53, pp. 65-92.

Crowther, K. (2016). *Effective Spacetime.* Springer International Publishing Switzerland.

Curd, M. and Cover, J. A. (1998). *Philosophy of Science. The Central Issues.* New York and London: W. W. Norton & Company.

Davidson, D. (1967). 'Truth and Meaning'. *Synthese,* 17 (3), pp. 304-323.

Dawid, R. (2006). 'Under-determination and Theory Succession from the Perspective of String Theory'. *Philosophy of Science,* 73 (3), pp. 298-322.

Dawid, R. (2013). *String Theory and the Scientific Method.* Cambridge: Cambridge University Press.

De Haro, S. (2017). 'Dualities and Emergent Gravity: Gauge/gravity duality'. *Studies in History and Philosophy of Modern Physics,* 59, pp. 109-125.

De Haro, S. (2017a). 'The Invisibility of Diffeomorphisms'. *Foundations of Physics,* 47

(11), 2017, p. 1464.

De Haro, S. (2019). ‘Towards a Theory of Emergence for the Physical Sciences’. *European Journal for Philosophy of Science,* 9, 38, pp. 1-52.

De Haro, S. (2019a). ‘The Heuristic Function of Duality’. *Synthese,* 196 (12), pp. 5169-5203. https://doi.org/10.1007/s11229-018-1708-9

De Haro, S. (2019b). ‘Theoretical Equivalence and Duality’. *Synthese,* pp. 1-39, special issue on ‘Symmetries and Asymmetries in Physics’. Editors: M. Frisch, R. Dardashti, G. Valente.

De Haro, S. (2019c). ‘Science and Philosophy: A Love-Hate Relationship’. *Foundations of Science,* 25, 2020, pp. 297-314.

De Haro, S. (2020). ‘Spacetime and Physical Equivalence’. Forthcoming in *Space and Time after Quantum Gravity*, Cambridge University Press. Huggett, N. and Wüthrich, C. (Eds.), http://philsci-archive.pitt.edu/13243.

De Haro, S. (2020a). ‘On Empirical Equivalence and Duality’. To appear in *100 Years of Gauge Theory. Past, Present and Future Perspectives,* S. De Bianchi and C. Kiefer (Eds.). Springer.

De Haro, S. (2020b). ‘The Empirical Under-determination Argument Against Scientific Realism for Dual Theories’. *Erkenntnis,* accepted with minor corrections.

De Haro, S. (2020c). ‘An Extensional Scientific Realism’. Submitted.

De Haro, S. (2020d). ‘On the Emergence of Masslessness’. Resubmission invited at, *Studies in History and Philosophy of Modern Physics.*

De Haro, S. (2020e). ‘The Emergence of Space in Random Matrix Models’. In preparation.

De Haro, S. and Butterfield, J.N. (2018). ‘A Schema for Duality, Illustrated by Bosonization’. In: Kouneiher, J. (Ed.), *Foundations of Mathematics and Physics one century after Hilbert,* pp. 305-376. Springer.

De Haro, S. and Butterfield, J. N. (2019). ‘On Symmetry and Duality’. Forthcoming in *Synthese,* in a special issue on ‘Symmetries and Asymmetries in Physics’ edited by M. Frisch, R. Dardashti, and G. Valente. doi:10.1007/s11229-019-02258-x

De Haro, S. and De Regt, H. W. (2018). ‘Interpreting theories without a Spacetime’. *European Journal for Philosophy of Science*, 8 (3), pp. 631-670.

De Haro, S. and De Regt, H. W. (2018a). ‘A Precipice Below Which Lies Absurdity? Theories without a spacetime and scientific understanding’. *Synthese,* pp. 1-29.

De Haro, S., Mayerson, D., Butterfield, J.N. (2016). ‘Conceptual Aspects of Gauge/Gravity Duality’. *Foundations of Physics,* 46 (11), pp. 1381-1425.

De Haro, S., Teh, N., Butterfield, J.N. (2016). ‘On the Relation between Dualities and Gauge Symmetries’. *Philosophy of Science*, 83 (5), pp. 1059-1069.

De Haro, S., Teh, N., Butterfield, J.N. (2017). ‘Comparing dualities and gauge symmetries’. *Studies in History and Philosophy of Modern Physics*, 59, pp. 68-80.

De Haro, S., van Dongen, J., Visser, M., Butterfield, J.N. (2020). ‘Conceptual Analysis of Black Hole Entropy in String Theory’. *Studies in History and Philosophy of Modern Physics,* published online. doi.org/10.1016/j.shpsb.2019.11.001.

De Regt, H. W. (2014). “Visualization as a tool for understanding”, *Perspectives on Science*, 22, pp. 377-396.

De Regt, H. W. (2017). ‘Understanding Scientific Understanding’. Oxford: OUP.

De Regt, H. W., and Dieks, D. (2005). “A contextual approach to scientific understanding”. *Synthese*, 144 (1), 137-170.

Dewar, N. (2017). ‘Interpretation and Equivalence; or, Equivalence and Interpretation’. Forthcoming in: E. Curiel and S. Lutz (Eds.), *The Semantics of Theories.*

Dieks, D., Dongen, J. van, Haro, S. de (2015). ‘Emergence in Holographic Scenarios for Gravity’. *Studies in History and Philosophy of Modern Physics,* 52, pp. 203-216.

Dijkgraaf, R. (1997). ‘Les Houches lectures on fields, strings and duality’. In *Les Houches 1995, Quantum symmetries*, pp. 3-147. [hep-th/9703136].

Dirac, P. A. M. (2001) [1964]. *Lectures on Quantum Mechanics.* Mineola: Dover.

Dizadji-Bahmani, F., Frigg, R., Hartmann, S. (2010). ‘Who’s afraid of Nagelian reduction?’ *Erkenntnis*, 73 (3), pp. 393-412.

Earman, J. (2002). ‘Thoroughly Modern McTaggart. Or what McTaggart would have said if he had learned General Relativity Theory’. *Philosophers Imprint 2,* pp. 1-28.

Egg, M. (2018). ‘Entity Realism’. In: Saatsi (2018), pp. 120-132.

Ellis, G., and Silk, J. (2014). 'Scientific Method: Defend the Integrity of Physics'. *Nature News,* 516 (7531), p. 321.

Feyerabend, P. K. (1963). 'How to Be a Good Empiricist—A Plea for Tolerance in Matters Epistemological'. In: Baumrin, B. (Ed.), *Philosophy of Science, The Delaware Seminar,* vol. 2, New York: Interscience Publishers, pp. 3-39. Reprinted in: M. Curd, and J. A. Cover, *Philosophy of Science,* New York London: W. W. Norton, pp. 922-949.

Field, H. (2016). *Science Without Numbers,* Second Edition. Oxford: Oxford University Press.

Fine, A. (1984). 'The Natural Ontological Attitude'. In: *Scientific Realism*, J. Leplin (Ed.), pp. 83-107. Berkeley: University of California Press.

Fine, K. (2012). 'Guide to Ground'. In: Correia, F., Schnieder, B., *Metaphysical Grounding. Understanding the Structure of Reality.* Cambridge University Press.

Fletcher, S. C. (2016). 'Similarity, topology, and physical significance in relativity theory.' *The British Journal for the Philosophy of Science*, 67 (2), pp. 365-389.

Floridi, L. (2011). *The Philosophy of Information,* Oxford: Oxford University Press.

Franklin, A. and Knox, E. (2018). 'Emergence without limits: The case of phonons'. *Studies in History and Philosophy of Modern Physics,* 64, pp. 68-78.

Fraser, D. (2017). 'Formal and physical equivalence in two cases in contemporary quantum physics'. *Studies in History and Philosophy of Modern Physics*, 59, pp. 30-43.

Frege, G. (1892), 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und philosophische Kritik*, pp. 25-50; translated as 'On Sense and reference', in P.T. Geach and M. Black eds. (1960), *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Blackwell.

Frege, G. (1956). 'The Thought: A Logical Inquiry'. *Mind,* LXV, 259, pp. 289-311. Translated from German [1918].

French, S. (2018). 'Realism and Metaphysics'. In: Saatsi (2018), pp. 394-406.

Frigg, R. and Votsis, I. (2011). 'Everything you always wanted to know about structural realism but were afraid to ask'. *European Journal for Philosophy of Science,* 1, pp. 227-276.

Frisch, M. (2005). *Inconsistency, Asymmetry, and Non-Locality. A philosophical investigation of classical electrodynamics*. Oxford University Press.

Galison, P. (1997). *Image and Logic. A Material Culture of Microphysics.* Chicago and London: The University of Chicago Press.

Glymour, C. (1970). 'Theoretical Equivalence and Theoretical Realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1970, pp. 275-288.

Glymour, C. (1977). 'The epistemology of geometry'. *Noûs*, pp. 227-251.

Glymour, C. (2013). 'Theoretical Equivalence and the Semantic View of Theories'. *Philosophy of Science*, 80, pp. 286-297.

Guay, A., Sartenaer, O. (2016). 'A new look at emergence. Or when after is different'. *European Journal for Philosophy of Science*, 6 (2), pp. 297-322.

Hacking, I. (1983). *Representing and Intervening. Introductory Topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press.

Halvorson, H. (2012). 'What Scientific Theories Could Not Be'. *Philosophy of Science*, 79, pp. 183-206.

Halvorson, H. (2013). 'The Semantic View, If Plausible, Is Syntactic'. *Philosophy of Science*, 80, pp. 475-478.

Halvorson, H. and Tsementzis, D. (2015). 'Categories of Scientific Theories'. PhilSci: http://philsci-archive.pitt.edu/11923.

Han, D., Kim, Y. S. (1981). 'Little Group for Photons and Gauge Transformations'. *American Journal of Physics*, 49, p. 348.

Hardin, C. L. and Rosenberg, A. (1982). 'In Defense of Convergent Realism'. *Philosophy of Science,* 49 (4), pp. 604-615.

Hartmann, S. (2002). 'Essay review On Correspondence'. *Studies in History and Philosophy of Modern Physics,* 33, pp. 79-94.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar.* Malden and Oxford: Blackwell.

Hempel, C. (1966). *Philosophy of Natural Science.* New York: Prentice-Hall, New York.

Hendry, R.F. (2010). 'Ontological reduction and molecular structure'. *Studies in History and Philosophy of Modern Physics*, 41, pp. 183-191.

Hesse, M. (1966). *Models and Analogies in Science.* Notre Dame: University of Notre Dame Press.

Hey, S. P. (2016). 'Heuristics and Meta-heuristics in Scientific Judgment'. *The British Journal for the Philosophy of Science*, 67, pp. 471-495.

Hoefer, C. and Smeenk, C. (2016). "Philosophy of the physical sciences". In: Paul Humphreys (Ed.), *The Oxford Handbook of Philosophy of Science.* Oxford University Press. Pp. 115-136.

Houkes, W. and Vermaas, P.E. (2010). 'Technical Functions. On the Use and Design of Artefacts'. Dordrecht Heidelberg London New York: Springer.

Hudetz, L. (2018). 'Definable Categorical Equivalence'. PhilSci: http://philsci-archive.pitt.edu/14297.

Huggett, N. (2017). 'Target space $\neq$ space'. *Studies in History and Philosophy of Modern Physics*, 59, 81-88.

Huggett, N. and Wüthrich, C. (2013). 'Emergent spacetime and empirical (in)coherence'. *Studies in History and Philosophy of Modern Physics*, 44, pp. 276-285.

Humphreys, P. (2016). *Emergence. A Philosophical Account.* Oxford University Press.

Inönü, E. and Wigner, E. P. (1953). 'On the Contraction of Groups and their Reperesentations'. *Proceedings of the National Academy of Sciences*, 39 (6), pp. 510-524.

Joos, E. and Zeh, H. D. (1985). 'The Emergence of Classical Properties Through Interaction with the Environment'. *Condensed Matter,* 59, pp. 223-243.

Joos, E., Zeh, H. D., Kiefer, C., Giulini, D. J., Kupsch, J., Stamatescu, I. O. (2013). *Decoherence and the appearance of a classical world in quantum theory.* Berlin, Heidelberg: Springer-Verlag.

Kaplan, D. (1977). 'Demonstratives. An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals'. In: *Themes from Kaplan,* Almog, J., Perry, J., Wettstein, H. (Eds.), 1989, Oxford University Press.

Kim, Y.S. (2001). 'Internal Space-time Symmetries of Massive and Massless Particles and their Unification'. *Nuclear Physics* B, Proceedings Supplements, 102, pp. 369-376.

Kitcher, P. (1993). *The Advancement of Science.* New York and Oxford: Oxford

University Press.

Kramers, H. A., and Wannier, G. H. (1941). 'Statistics of the Two-Dimensional Ferromagnet. Part I'. *Physical Review,* 60 (3), 252.

Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F., Smalley, R. E.(1985). 'C_{60}: Buckminsterfullerene'. *Nature,* Letters, 318, pp. 162-163.

Kroto, H. W., Curl, R. F., Smalley, R. E.(1996). *Press release*, 1996 Nobel Prize in Chemistry, https://www.nobelprize.org/prizes/chemistry/1996/press-release.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* Chicago: The University of Chicago Press, Second Edition, 1970.

Kuipers, T. A. F. (1982). 'Approaching Descriptive and Theoretical Truth'. *Erkenntnis,* 18, pp. 343-378.

Kuipers, T. A. F. (1987). *What Is Closer-to-the-Truth?* Rodopi, Amsterdam.

Ladyman, J. (2014). 'Structural Realism'. *Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/entries/structural-realism.

Ladyman, J. and Ross, D. (2007). *Every Thing Must Go. Metaphysics Naturalized.* With Spurrett, D. and Collier, J. Oxford: Oxford University Press.

Landsman, N. P. (2007). 'Between Classical and Quantum'. In: *Philosophy of Physics. Part A,* Handbook of the Philosophy of Science, J. Butterfield and J. Earman (Eds.), Amsterdam: Elsevier.

Landsman, N. P. (2013). 'Spontaneous Symmetry Breaking in Quantum Systems: Emergence or Reduction?' *Studies in History and Philosophy of Modern Physics,* 44(4), pp. 379-394.

Laudan, L. (1977). *Progress and Its Problems.* Berkeley and Los Angeles: University of California Press.

Laudan, L. (1981). 'A Confutation of Convergent Realism'. *Philosophy of Science,* 48 (1), pp. 19-49.

Laudan, L. (1984). 'Realism without the Real'. *Philosophy of Science,* 51 (1), pp. 156-162.

Laudan, L. (1984). 'Science and Values. The aims of science and their role in scientific debate'. Berkeley Los Angeles London: University of California Press.

Laughlin, R. B. and Pines, D. (2000). 'The Theory of Everything'. *Proceedings of the National Academy of Sciences of the United States of America*, 97 (1), pp. 28-31.

Le Bihan, B. and Read, J. (2018). 'Duality and Ontology'. *Philosophy Compass,* 13 (12), e12555.

Lewis, D. K. (1970). 'How to Define Theoretical Terms'. *The Journal of Philosophy,* 67 (13), pp. 427-446.

Lewis, D. K. (1972). 'Psychophysical and Theoretical Identifications'. *Australasian Journal of Philosophy,* 50 (3), pp. 249-258.

Lewis, D. K. (1980). 'Index, Context, and Content'. Reprinted in: *Papers in Philosophical Logic,* 1998, pp. 21-44. Cambridge: Cambridge University Press.

Lewis, D. K. (1983). 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy*, 61 (4), pp. 343-377.

Lewis, D. K. (1984). 'Putnam's Paradox'. *Australasian Journal of Philosophy,* 62 (3), pp. 221-136.

Lutz, S. (2017). 'What Was the Syntax-Semantics Debate in the Philosophy of Science About?' *Philosophy and Phenomenological Research*, XCV (2), pp. 319-352.

Mainwood, P. (2006). *Is More Different? Emergent Properties in Physics.* PhD dissertation, University of Oxford. http://philsci-archive.pitt.edu/8339.

Markopoulou (2009), "New directions in background independent Quantum Gravity", in *Approaches to Quantum Gravity*, Oriti, D. (Ed.), pp. 129-166. CUP: Cambridge.

Martin, M. (1971). 'Referential Variance and Scientific Objectivity'. *The British Journal for the Philosophy of Science,* 22 (1), pp. 17-26.

Massimi, M. (2016). 'Four Kinds of Perspectival Truth'. *Philosophy and Phenomenological Research,* XCVI (2), pp. 342-359.

Massimi, M. (2018). 'Perspectivism'. In: Saatsi (2018), pp. 164-175.

Maxwell, J. C. (1954). *A Treatise on Electricity and Magnetism,* volume 2. Reprint of 1891 edition. London: Dover.

Matsubara, K. (2013). 'Realism, Underdetermination and String Theory Dualities. *Synthese,* 190, 471-489.

Messiah, A. (1961). 'Quantum Mechanics. Volume I'. Amsterdam: North-Holland.

Mill, J. S. (1882). *A System of Logic.* Eighth Edition, New York: Harper and Brothers.

Miller, D. (1974). 'Popper's Qualitative Theory of Verisimilitude'. *The British Journal for the Philosophy of Science,* 25, pp. 166-177.

Montague, R. (1970). 'Pragmatics and Intensional Logic'. *Synthese*, 22, pp. 68-94.

Morgan, M. S. and Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science.* Cambridge: Cambridge University Press.

Mössner, N. (2018). *Visual Representations in Science. Concept and Epistemology.* London and New York: Routledge.

Musgrave, A. (1985). 'Realism versus Constructive Empiricism'. In: *Images of Science,* P. M. Chruchland and C. A. Hooker (Eds.), Chicago: University of Chicago Press.

Musgrave, A. (1989). 'Noa's Ark—Fine for Realism'. *The Philosophical Quarterly,* 39 (157), pp. 383-398.

Myrvold, W. C. (2019). "—It would be possible to do a lengthy dialectical number on this;" PhilSci 16675, http://philsci-archive.pitt.edu/16675.

Nagel, E. (1949). 'The Meaning of Reduction'. In: *Philosophy of Science*, A. Danto and S. Morgenbesser (Eds.), pp. 288-312. Meridian Books: Cleveland and New York.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation.* New York: Harcourt.

Nagel, E. (1979). 'Issues in the Logic of Reductive Explanations'. In: *Teleology Revisited,* New York: Columbia University Press, pp. 95-113. Reprinted in: M. Curd, and J. A. Cover, *Philosophy of Science,* New York London: W. W. Norton, pp. 905-921.

Neuber, M. (2018). 'Realism and Logical Empiricism'. In: Saatsi (2018), pp. 7-19.

Newton, I. (1721). *Opticks: or, a Treatise of the Reflections, Refractions, Infections and Colours of Light.* Third Edition. London, William and John Innys.

Newton, I. (1999). The Principia. Mathematical Principles of Natural Philosophy. The Authoritative Translation by I. Bernard Cohen and Anne Whitman. Oakland: University of California Press.

Nickles, T. (1973). ‘Two Concepts of Intertheoretic Reduction’. *The Journal of Philosophy,* 70 (7), pp. 181-201.

Norton, J. D. (2012). ‘Approximation and Idealization: Why the Difference Matters’, *Philosophy of Science*, 79 (2), 207-232.

Norton, J. D. (2016). ‘The Impossible Process: Thermodynamic Reversibility’. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 55, pp. 44-61.

O’Connor, T., and Wong, H. Y. (2002). ‘Emergent Properties’. *Stanford Encyclopedia of Philosophy.* First published 24 September 2002; substantive revision on 3 June 2015.

Oriti, D. (Ed.), 2009, *Approaches to Quantum Gravity.* CUP: Cambridge.

Oriti, D. (2013). “Disappearance and emergence of space and time in quantum gravity”, *Studies in History and Philosophy of Modern Physics* 46, pp. 186-199.

Oriti, D. (2016). “Space and time are emergent, in quantum gravity. What is cosmology, then?” Talk given at the conference *Foundations of Physics*, LSE, London. 16/07/2016.

Pauli, W. (1958). ‘Die allgemeinine Prinzipien der Wellenmechanik’. In: *Handbuch der Physik,* vol. 5. Flügge, S. (Ed.). Berlin: Springer, Pt. 1. (Translated into English: Berlin: Springer-Verlag, 1980).

Perez, A. (2009). “The spin foam representation of loop quantum gravity”, in *Approaches to Quantum Gravity*, Oriti, D. (Ed.), pp. 272-289. CUP: Cambridge.

Perry, J. (1979). ‘The Problem of the Essential Indexical’. *Noûs,* 13 (1), pp. 3-21.

Pitts, J. B. (2011). ‘Permanent Underdetermination from Approximate Empirical Equivalence in Field Theory’. *The British Journal for the Philosophy of Science*, 62 (2), pp. 259-299.

Polchinski, J. (1998). ‘String theory. Volume 1: An introduction to the bosonic string’. Cambridge: CUP.

Polchinski, J. (2017). ‘Dualities of Fields and Strings’. *Studies in History and Philosophy of Modern Physics*, 59, 2017, pp. 6-20.

Pooley, O. (2017). ‘Background Independence, Diffeomorphism Invariance and the Meaning of Coordinates. In: *Towards a Theory of Spacetime Theories,* Lehmkuhl, D., Schiemann, G., Scholz, E. (Eds.). New York: Birkhäuser.

Popper, K. R. (1962). *Conjectures and Refutations. The Growth of Scientific Knowledge,* Basic Books, New York and London.

Psillos, S. (1999). *Scientific Realism. How Science Tracks Truth.* London and New York: Routledge.

Psillos, S. (2012). 'Causal Descriptivism and the Reference of Theoretical Terms'. In: *Perception, Realism, and the Problem of Reference,* A. Raftopoulos and P. Machamer (Eds.), pp. 212-238.

Psillos, S. (2018). 'The Realist Turn in the Philosophy of Science'. In: Saatsi (2018), pp. 20-34.

Putnam, H. (1978). *Meaning and the Moral Sciences.* Oxon and New York: Routledge.

Quine, W. V. O. (1951). 'Ontology and Ideology'. *Philosophical Studies. An International Journal for Philosophy in the Analytic Tradition*, 2 (1), pp. 11-15.

Quine, W. V. (1960). *Word and Object,* New Edition, 2013. Cambridge, MA and London: The MIT Press.

Quine, W. V. (1970). 'On the Reasons for Indeterminacy of Translation'. *The Journal of Philosophy,* 67 (6), pp. 178-183.

Quine, W. V. (1975). 'On empirically equivalent systems of the world'. *Erkenntnis*, 9 (3), pp. 313-328.

Radder, H. (1991). 'Heuristics and the Generalized Correspondence Principle'. *The British Journal for the Philosophy of Science*, 42, 195-226.

Radder, H. (2012) [1984]. *The Material Realization of Science,* Revised Edition (first edition: 1984). Springer Dordrecht Heidelberg New York London.

Read, J. (2016). 'The Interpretation of String-Theoretic Dualities'. *Foundations of Physics,* 46, pp. 209-235.

Read, J. and Møller-Nielsen, T. (2018). 'Motivating Dualities'. Forthcoming in *Synthese.*

Richardson, R. C. (2007). 'Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality'. *Notre Dame Philosophical Reviews*, 2007.12.14.

Rickles, D. (2011). ‘A Philosopher Looks at String Dualities’. *Studies in History and Philosophy of Modern Physics,* 42, pp. 54-67.

Rickles, D. (2013). ‘AdS/CFT duality and the emergence of spacetime’. *Studies in History and Philosophy of Modern Physics*, 44, pp. 312-320.

Rickles, D. (2017). ‘Dual Theories: ‘Same but Different’ or ‘Different but Same’?’ *Studies in History and Philosophy of Modern Physics,* 59, pp. 62-67.

Roberts, B. W. (2014). ‘Disregarding the ‘Hole Argument”. arXiv preprint arXiv:1412.5289. http://philsci-archive.pitt.edu/11687

Rosen, G. (1994). ‘What Is Constructive Empiricism?’ *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 74 (2), pp. 143-178.

Rosenstock, S., Barrett, T. W., Weatherall, J. O. (2015). ‘On Einstein Algebras and Relativistic Spacetimes’. *Studies in History and Philosophy of Modern Physics,* 52, pp. 309-316.

Rovelli, C. and Vidotto, F. (2015). “Covariant Loop Quantum Gravity. An Elementary Introduction to Quantum Gravity and Spinfoam Theory”, CUP: Cambridge.

Rueger, A. (2000). ‘Physical Emergence, Diachronic and Synchronic’. *Synthese*, 124 (3), pp. 297-322.

Ruetsche, L. (2011). *Interpreting Quantum Theories.* Oxford University Press.

Rynasiewicz, R. (2015). ‘The (?) correspondence principle’, pp. 175-199 in Aaserud, F. and Kragh, H. (eds.), *One hundred years of the Bohr atom*, Copenhagen: The Royal Danish Academy of Sciences and Letters.

Saatsi, J. (Ed.) (2018). *The Routledge Handbook of Scientific Realism.* Oxon and New York: Routledge.

Saatsi, J. (2018). ‘Realism and the Limits of Explanatory Reasoning’. In: Saatsi (2018), pp. 200-211.

Saunders, S. (1993). ‘To What Physics Corresponds’. In: *Correspondence, Invariance and Heuristics,* S. French and H. Kamminga (Eds.). Dordrecht: Springer.

Sankey, H. (2018). ‘Kuhn, Relativism, and Realism’. In: Saatsi (2018), pp. 72-83.

Schaffer, J. (2009). ‘On What Grounds What’. In: Chalmers, D. J., Manley, D., Wasserman, R., *Metametaphysics. New Essays on the Foundations of Ontology.*

Oxford: OUP.

Schaffner, K. F. (1967). ‘Approaches to Reduction’. *Philosophy of Science,* 34 (2), pp. 137-147.

Schaffner, K. F. (1972). *Nineteenth-Century Aether Theories.* Oxford and New York: Pergamon Press.

Schaffner, K. F. (2012). ‘Ernest Nagel and reduction’. *The Journal of Philosophy*, 109 (8/9), pp. 534-565.

Scheffler, I. (1967). *Science and Subjectivity.* Indianapolis New York Kansas City: The Bobbs-Merrill Company.

Schurz, G. and Weingartner, P. (1987). ‘Verisimilitude Defined by Relevant Consequence-Elements’. In: Kuipers (1987), pp. 47-77.

Sklar, L. (1974). *Space, Time, and Spacetime.* Berkeley and Los Angeles: University of California Press.

Sklar, L. (1975). ‘Methodological Conservatism’. *The Philosophical Review,* 84 (3), pp. 374-400.

Speaks, J. (2019). ‘Theories of Meaning’. *Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/entries/meaning.

Stanford, P. K. (2006). *Exceeding Our Grasp.* New York: Oxford University Press.

Stein, H. (1989). ‘Yes, but... Some Skeptical Remarks on Realism and Anti-Realism’. *Dialectica,* 43 (1/2), pp. 47-65.

Tarski, A. (1935). ‘The Concept of Truth in Formalized Languages’. In: *Logic, Semantics, Metamathematics,* pp. 152-278. Oxford: Clarendon Press, 1956.

Tarski, A. (1944). ‘The Semantic Conception of Truth: and the Foundations of Semantics’. *Philosophy and Phenomenological Research,* 4 (3), pp. 341-376.

Teh, N. J. and Tsementzis, D. (2017). “Theoretical equivalence in classical mechanics and its relationship to duality”, *Studies in History and Philosophy of Modern Physics*, 59, pp. 44-54. doi.org/10.1016/j.shpsb.2016.02.002

Tichý, P. (1974). ‘On Popper’s Definitions of Verisimilitude’. *The British Journal for the Philosophy of Science,* 25, pp. 155-160.

Toulmin, S. (1961). 'Foresight and Understanding. An enquiry into the aims of Science'. Indiana University Press.

van Dongen, J., De Haro, S., Visser, M., Butterfield, J.N. (2020). 'Emergence and Correspondence for String Theory Black Holes'. *Studies in History and Philosophy of Modern Physics,* published online. doi.org/10.1016/j.shpsb.2019.11.002.

van Fraassen, B. C. (1970). 'On the Extension of Beth's Semantics of Physical Theories'. *Philosophy of Science,* 37 (3), pp. 325-339.

van Fraassen, B. C. (1980). *The Scientific Image.* Oxford: Clarendon Press.

van Fraassen, B. C. (1989). *Laws and Symmetry.* Oxford: Clarendon Press.

van Fraassen, B. C. (1994). 'Gideon Rosen on Constructive Empiricism'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition,* 74 (2), pp. 179-192.

van Fraassen, B. C. (2014). 'One or Two Gentle Remarks about Hans Halvorson's Critique of the Semantic View'. *Philosophy of Science*, 81, pp. 276-283.

Vickers, P. (2013). *Understanding Inconsistent Science.* Oxford: Oxford University Press.

Wallace, D. (2012). *The Emergent Multiverse.* Oxford: Oxford University Press.

Weatherall, J. O. (2015). 'Categories and the Foundations of Classical Field Theories'. PhilSci: http://philsci-archive.pitt.edu/11587.

Weatherall, J. O. (2016). 'Are Newtonian gravitation and geometrized Newtonian gravitation theoretically equivalent?' *Erkenntnis*, 81 (5), pp. 1073-1091.

Weatherall, J. O. (2016a). 'Understanding Gauge'. *Philosophy of Science*, 83, pp. 1039-1049.

Weatherall, J. O. (2019). 'Equivalence and Duality in Electromagnetism'. arXiv:1906.09699.

Weyl, H. (1934). 'Mind and Nature'. *Hermann Weyl, Mind and Nature, Selected Writings on Philosophy, Mathematics, and Physics.* P. Pesic, 2009, pp. 95-96. Princeton and Oxford: Princeton University Press.

Whewell, W. (1876). Letter to Prof. J. D. Forbes (1960), in *William Whewell, D. D., Master of Trinity College, Cambridge: an account of his literary and scientific correspondence*, vol. 2, Todhunter, I. London: Macmillan 1876.

Whittaker, E. (1951). *A History of the Theories of Aether and Electricity,* London: Thomas Nelson and Sons.

Wigner, E. (1939). 'On Unitary Representations of the Inhomogeneous Lorentz Group'. Reprinted in: *Nuclear Physics B* (Proceedings Supplement), 6, 1989, pp. 9-64.

Wilson, M. (2006). *Wandering Significance. An Essay on Conceptual Behavior.* Oxford: Oxford University Press.

Wimsatt, W. C. (1997). 'Reductive Heuristics for Finding Emergence'. *Philosophy of Science*, 64, pp. S372-S384.

Wimsatt, W.C. (2007). *Re-Engineering Philosophy for Limited Beings. Piecewise Approximations to Reality.* Cambridge, Massachusetts, and London: Harvard University Press.

Witten, E. (1995). 'String Theory Dynamics in Various Dimensions'. *Nuclear Physics B,* 443 (1-2), pp. 85-126.

Wong, H. Y. (2010). 'The Secret Lives of Emergents'. In: A. Corradini and T. O'Connor, *Emergence in Science and Philosophy,* pp. 7-24. New York and London: Routledge.

Worrall, J. (1989). 'Structural Realism: The Best of Both Worlds?' *Dialectica,* 43 (1-2), pp. 99-124.

Wray, K. B. (2018). 'Success of Science as a Motivation for Realism'. In: Saatsi (2018), pp. 37-47.

Zurek, W. Z. (2002). 'Decoherence and the Transition from Quantum to Classical—*Revisited*'. *Los Alamos Science,* 27, pp. 2-25. arXiv preprint quant-ph/0306072.

Zwiebach, B. (2009). *A First Course in String Theory.* Cambridge: Cambridge University Press.