Symmetries in physics, metaphysics, and logic

Neil Dewar

University College, University of Oxford

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Hilary Term, 2016

For my family

Abstract

This thesis examines the idea that when a physical theory contains symmetries, the theory should be interpreted in such a way that symmetry-related models represent the same physical state of affairs. It argues that we can best do so by drawing on analogies to ideas in philosophy of logic and language: specifically, by thinking of symmetries as a means of translating a theory into itself. It consists of six chapters, together with an introduction and conclusion.

In Chapter 1, I set up the main ideas needed to more precisely frame the question at hand: namely, the notions of symmetry, interpretation, and possibility. I make some remarks about how I take these to be connected.

In Chapter 2, I argue that isomorphic models should be interpreted as equivalent. After giving some motivations for doing so, I consider the main obstruction: how to provide an account of *de re* modality. I review how counterpart theory may be used to overcome this obstruction, and clarify how counterpart theory relates to other positions in the debate over modality *de re*.

In Chapter 3, I show that the metaphysical debate over quidditism can be made precise by drawing on notions of translation from model theory, and argue in favour of an anti-quidditist attitude towards interpreting theories. I then consider the special case of translating a theory into itself: how such a theory should be interpreted, and what reformulations of the theory such an interpretation suggests.

In Chapter 4, I turn my attention to physics. I define the notion of an internal symmetry for a theory, and argue that they may be regarded as translations from a theory into itself (in the sense of Chapter 3); and, hence, that symmetry-related models should be interpreted as equivalent. Drawing on the analogy further, I look at how the theory may be reformulated to take this interpretation into account.

In Chapter 5, I look at external symmetries. I argue, drawing on ideas from Chapters 2 and 3, that models related by external symmetries should also be interpreted as equivalent. I discuss how implementing this interpretational lesson bears on finding the spacetime structure appropriate to a theory.

In Chapter 6, I consider a specific external symmetry: the accelerative symmetry of Newtonian gravitation. I show that one can reformulate the theory to take this into account, setting gravitation on a spacetime structure that has absolute rotation but no absolute acceleration.

Some elements of this thesis have been published in *Studies in History and Philosophy of Modern Physics*. This thesis is approximately 60,000 words long.

Acknowledgments

I am grateful to the Arts and Humanities Research Council for four years of support, first as a B.Phil. and then as a D.Phil. student; I would also like to thank Princeton University, and the Trustees of the Henry Fund, for a Procter Visiting Fellowship. My thanks also go to University College, the Oxford Philosophy Faculty, and the Princeton Department of Philosophy for providing stimulating and welcoming environments in which to work.

My greatest intellectual debt is to David Wallace. His impact on what is here (and on what is not) is too large to be properly expressed, but I hope that his influence is visible throughout what follows. Oliver Pooley's supervision, for the later stages of this project, has also been extremely helpful; the arguments here have benefited enormously from his trenchant comments. Finally, I have also been lucky enough to receive supervision from Shamik Dasgupta and Hans Halvorson, during three semesters as a visiting student at Princeton. They have also greatly shaped this thesis (and my own philosophical outlook), and I am indebted to them both.

Many other philosophers at Oxford and Princeton deserve thanks, for teaching, discussion, and general support and friendship: Thomas Barrett, Harvey Brown, Rachel Fraser, Dan Isaacson, Alex Kaiserman, Boris Kment, Niels Martens, Michaela McSweeney, Alex Meehan, Joseph Melia, Tushar Menon, Thomas Møller-Nielsen, Beau Mount, Michael Price, Carina Prunkl, Simon Saunders, David Schroeren, Teru Thomas, Chris Timpson, Dimitris Tsementzis, Andy Yu, and—especially—Katie Robertson.

I also owe significant debts outside Oxford and Princeton. Special thanks to the Munich Center for Mathematical Philosophy for their hospitality in the summer of 2014, and especially to Karim Thébault, Sebastian Lutz and Erik Curiel for many discussions. Otherwise, for correspondence and discussion of the ideas developed here (and many others besides), I thank Harjit Bhogal, Adam Caulton, John Dougherty, Josh Eisenthal, Alison Fernandes, James Fraser, Larry Lee, Chip Sebens, and Jim Weatherall.

Beyond philosophy, that the below exists at all is thanks to the generous and unstinting friendship of many many people—but especially Steph Bell, Hugh Burns, Jen Coyne, Hasan Dindjer, Erin Fitzgerald, Tom Hosking, Will Jones, Arthur Learoyd, Jonathan Leader Maynard, and Emma Webber. Finally, very special thanks to my parents and my sisters. The love and support they have provided over many years cannot remotely be articulated here; with gratitude and love, I dedicate this thesis to them.

Contents

Introduction

This thesis concerns the symmetries of physical theories. It's widely recognised that the symmetries of a theory are a key guide to its physical content. In particular, it is often suggested that the truly *physical* content of a theory is to be identified with its *invariant* content: if some feature of a theory's formalism changes non-trivially under the application of a symmetry transformation, then the theory should not be interpreted so as to give that feature ontological "weight". The purpose of this thesis is to clarify, defend, and apply that suggestion.[1]

Chapters 1–5 form an integrated argument, constituting a general exploration and defence of the suggestion above (which I'll refer to as the "symmetry-interpretation link"). The route we take to get there is gradual: we start with general considerations in philosophy of science (Chapter 1), then move to topics in metaphysics (Chapters 2 and 3), before applying the results of those investigations to physical theories (Chapters 4 and 5). Chapter 6 is somewhat independent: it is an extended study in applying the lessons of symmetries in the context of Newtonian gravitation.

Despite the presence of some fairly heavy-duty metaphysical topics in this thesis, the overall tenor is less strongly realist than is, perhaps, the contemporary average. In particular, one recurring theme will be a repudiation of metaphysical structure above and beyond what plays a role in the laws, and in the architecture of the space of possibilities. Most notably, I will resist making appeals to naturalness, fundamentality, or grounding. The reasons to avoid such structure are familiar themes from anti-realist thought: they risk emplacing the nature of the world forever beyond our epistemic or semantic ken. In the contemporary philosophical literature, such concerns are apt to be decried as atavistic positivism—we've learnt by now, the story goes, that naïve empiricist scruples can be overcome by judicious appeal to theoretical virtues such as explanatory power. But the concern is precisely whether all such scruples are necessarily naïve. If we are to make a genuine assessment of whether we should ever

[1] One methodological observation should be made at the outset: in general, the logico-linguistic resources I bring to bear on this question are semantic in character. For an approach to symmetries that makes more use of syntactic notions, see [Saunders, 2016].

postulate more metaphysics than the laws need, presumably we should be comparing the explanatory virtues of one metaphysical scheme against the other. So the present work could be considered an experiment, in what kind of picture of the world is possible without such extra-nomological structure—and especially, in whether a stable resting place can be found between positivism and uncritical realism.

Chapter 1 was originally going to be this introduction: the intention was merely to introduce the notion of symmetry, and offer some brief remarks about the sense in which I am thinking of the project of theory interpretation. However, the remarks took on something of a life of their own, and so the chapter turned into a more general discussion of how interpretation, ontological commitment, and possibility interact with one another. The key idea is that interpreting a theory is a matter of determining when the models of that theory represent the same possible world or not; but this is as much a definition of the notion of a possible world as it is of interpretation. In effect, possible worlds are a certain kind of abstraction from models of (non-modal) theories.

Chapter 2 concerns the debate over haecceitism, that is, the debate over how individuals are identified across possible worlds. In my terms, this arises from a debate over how to interpret isomorphic models of a theory: do two isomorphic models represent the same possible world, or distinct such worlds? The bulk of the chapter is about how, if we take the former option, we are to account for *de re* modal claims, concerning how things could go for specific individuals. Deprived of one natural way of giving a semantics for claims of that kind, we must use counterpart theory instead.

In Chapter 3, we look at how *properties* are individuated—the issue that metaphysicians know as *quidditism*. My characterisation of this debate is somewhat different to that more standardly found. I argue that this issue pertains, in the first instance, to the identification of properties across *laws*, and only derivatively to the identification of properties across worlds. Moreover, I claim that we can give this a precise content by using the notion of translation found in model theory. I then look at how this can be applied to a certain special case: that of translating a theory into itself. This, I claim, is a manifestation of the idea of a symmetry.

This suggestion is followed up in Chapter 4, where I look at internal symmetries. The main argument is that the symmetry-interpretation link for internal symmetries can be justified by the same sorts of considerations we used to justify anti-quidditism—once we attend properly to the analogies between the model-theoretic formalism and the formalism of physical theories. I also expand on a contrast first noted in Chapter 3, between two methods of implementing that interpretational lesson ("sophistication"

and "reduction").

In Chapter 5, we look at external symmetries. Here, I argue that the lessons of Chapter 2 *and* Chapter 3 need to be brought to bear in order to justify the symmetry-interpretation link. I then consider some of the issues to which this gives rise: the question of determining what spacetime structure is most appropriate to a theory, and some of the subtleties which attend defining "the" external symmetry group of a theory.

As mentioned above, Chapter 6 stands somewhat independently of the rest (although it illustrates and applies several general themes from preceding chapters, especially Chapter 5). Its jumping-off point is the application of the symmetry-interpretation link to the external symmetries of Newtonian gravitation. In particular, it considers the "accelerative shift" symmetry that one finds in Newtonian gravitation, and seeks a formulation of the theory which embodies applying the symmetry-interpretation link to it (in the terminology of Chapters 4 and 5, it seeks a sophisticated version of the theory). In effect, this amounts to showing how Newtonian gravitation, in its field formulation, can be expressed entirely in terms of relative rather than absolute acceleration.

1 Symmetry and interpretation

> We battle on in words, as always,
> mere words,
> and what's the cure? We cannot find
> a thing.
>
> Homer, *The Iliad*

In this chapter, I introduce the issue that will be our concern throughout this thesis: what is a symmetry of a theory, and what are the implications of symmetries for the interpretation of those theories? First, I introduce a toy theory. This will illustrate some general ideas: most importantly, what I mean by a symmetry of a theory. Then, I go on to discuss the notion of interpretation that I am working with. Finally, I relate this apparatus of interpretation to modality.

1.1 Introduction to symmetries

Consider a single Newtonian particle, confined to one dimension, moving against the backdrop of some fixed potential. The theory of such a particle is an elementary component of any physics education. First, we have a pair of dynamical variables: one *independent variable* of time, t, and one *dependent variable* of position, x. Each of these ranges over a real-valued space; let us use X to denote the range of x, and T to denote the range of t. We also introduce a real-valued parameter m to characterise the particle's *mass*. Finally, we introduce a function $V : X \to \mathbb{R}$, to represent the potential at various points in space (which we identify with the possible locations of the particle). The content of the theory is then captured in the following equation:

$$m\frac{d^2x}{dt^2} = -\frac{dV}{dx} \tag{1.1}$$

The sense in which this equation summarises the physics of such a particle is as follows: any physically possible history for the particle is represented by a *solution* of the equation. A solution, here, is a function $f : T \to X$ such that at every $t \in T$, the above equation is satisfied. For instance, in the case of a free particle ($V = 0$), all solutions are those functions of the form

$$f(t) = at + b \tag{1.2}$$

for $a, b \in \mathbb{R}$.

This theory, simple though it is, already illustrates the core features of theories that will concern us in what follows. First, we introduce some kind of formal language: in this case, the language is just that of ordinary differential equations. Second, we stipulate the kinds of mathematical structures that will be put to representational work, and the way in which they can make sentences of the language true or false: in this case, the constructs are real-valued functions of one real argument, which may satisfy or fail to satisfy those differential equations. Finally, some kind of conditions (in the formal language) are specified, which those constructs may satisfy or fail to satisfy: in this case, the differential equation (1.1). This serves to pick out some of the constructs as privileged, i.e. those which do indeed satisfy the specified conditions: in this case, the solutions (1.2) of (1.1).

We will see, as we go, that this form is ubiquitous. A natural question, for those familiar with the literature in philosophy of science, is whether this form is better assimilated to the *syntactic* (sometimes: received) or *semantic* view of theories. In sloganeering form, that distinction runs as follows: the syntactic view maintains that theories are sets of sentences, whereas the semantic view identifies them with sets of models. If we take these slogans at face value, then I don't think that either is a good fit for the way I am understanding theories, i.e., for thinking of theories as a set of syntactic conditions for semantic constructions of a certain kind. I think it matters what semantic constructions which are taken to be the subject of the syntactic conditions; I certainly don't want to require that the theory's content be specifiable in terms of some kind of purely syntactic proof-procedure. Equally, it matters that the models of the theory are not an arbitrary set of mathematical structures, but rather a set of structures answering to some specific set of conditions. Moreover, I am quite happy with the idea that these models are "yoked to a particular syntax":[1] the spaces T and X are explicitly

[1] [van Fraassen, 1989, p. 366]

labelled (by the variables t and x respectively), in order to make manifest how to assess whether the condition (1.1) holds of a given function. That said, I don't wish to rule out the notion that some more subtle conception of these positions is consistent with this way of thinking about theories—indeed, I expect that one could render it consistent with a sufficiently thoughtful version of either view.[2] I merely wish to signal that it does not, so far as I can see, coincide with thoughtless versions of either.

The *symmetries* of such a theory are a special kind of *transformation*. A transformation, in this case, consists of a diffeomorphism (smooth bijection) of type $T \times X \to T \times X$ which admits of being specified by a pair of functions

$$\tau : T \to T \tag{1.3a}$$

$$\xi : T \times X \to X \tag{1.3b}$$

Transformations naturally induce a map on the space of functions from T to X: any function $f : T \to X$ is mapped to the function $\widetilde{f} : T \to X$ such that

$$\widetilde{f}(\tau(t)) = \xi(t, x(t)) \tag{1.4}$$

Formally, a symmetry transformation is a transformation which maps solutions to solutions: i.e. which is such that f is a solution of (1.1) iff $\widetilde{f}$ is. For example, again consider the free-particle case. Then the transformation

$$\tau(t) = t + c \tag{1.5a}$$

$$\xi(t, x) = vt + d \tag{1.5b}$$

for $v, c, d \in \mathbb{R}$ is a symmetry. However, there are two slightly different ways of characterising what it is that's striking about symmetry transformations, corresponding to two different ways of thinking about the effect of a general transformation.

On one conception—what we might call the *passive* conception—the transformation (1.3) represents a change in the *description*. That is, it is a change in the representational conventions governing the mathematics, such that (as a matter of stipulation) $\widetilde{f}$ represents whatever particle-history was previously represented by f. Now, if such a change is to be consistent, then it ought to be accompanied by a change in the differential

[2] Some (highly defeasible) evidence for this claim: when describing this view, I have been told both that it is clearly best thought of as an appropriately careful version of the semantic view, and (by others) that it is clearly best thought of as an appropriately careful version of the syntactic view.

equations being used. After all, we don't think that a change of representation ought to make any difference to the physics! So it had better be the case that the same physical histories get picked out as before; given that those histories are now represented by different functions, we want to change the equations so as to pick out that new set of functions. If the old set of differential equations was Δ, then we want a new set $\widetilde{\Delta}$ such that $\widetilde{f}$ satisfies $\widetilde{\Delta}$ iff f satisfies Δ. On the other conception, which we shall call the *active* conception, the transformation (1.3) is just a map on the space of functions: nothing more, nothing less. In particular, the representational conventions are held invariant across the application of this change. As such, there is no concomitant change to the differential equations.

These two conceptions yield slightly different ways of thinking about symmetries, and why they might be interesting. On the passive conception, we can characterise symmetry transformations as those transformations which do *not* require any alteration to the differential equations being used. For, if f satisfies Δ iff $\widetilde{f}$ satisfies Δ, then f satisfies Δ iff f satisfies $\widetilde{\Delta}$: and hence, $\widetilde{\Delta} = \Delta$. On the active conception, symmetry transformations are simply those transformations under which the space of solutions is invariant. In other words, both conceptions agree that in a symmetry transformation, both the (mathematical) theory in play, and the dynamical status of any model, remain unchanged. However, on the passive conception, the latter condition comes "for free" (since it holds for *any* transformation), and so the distinctive thing about symmetry transformations in particular is their satisfaction of the former condition. On the active conception, by contrast, it is the former condition that is upheld across all transformations, and so the distinctive nature of a symmetry transformation is that it also meets the latter condition.

This lets us be a little clearer about the claim earlier. The claim that I seek to defend in this thesis is that a pair of symmetry-related models, considered as models of the same theory and without changing our representational conventions, should not be interpreted as representing distinct possibilities. This makes clear that it is not just the trivial observation that we can change how the mathematics is to represent the physics; rather, it is a claim about the proper way to extract physical content from a theory. In many cases, such a claim can yield quite striking results. For example, note that *any* pair of solutions of the form (1.2) can be mapped to one another by a symmetry transformation of the form (1.5). So in the case of a free Newtonian particle, the claim entails that there is a *unique* physically possible history. The reasons to believe something like this (and some reasons why it is less surprising than it might sound,

once properly understood) follow, I claim, from general reflection upon interpretational practice. It will therefore be helpful to spend some time thinking about such practice, and its relationship to scientific theorising.

1.2 Interpretation and science

In order to assess what interpretation *is*, it is well to begin by considering what interpretation *does*. That is, we should ask what role the notion of interpretation is supposed to play in our scientific and philosophical practice. Having done so, we can then look at whether such-and-such an account of what interpretation involves does, in fact, describe an activity that instantiates that role.

Firstly, being willing (indeed, compelled) to engage in interpretation is one of the marks of being a scientific realist. To be a realist about some scientific theory is a commitment to the (approximate) truth of the theory,[3] and to be committed to the truth of those statements under a realistic semantics for theoretical terms.[4] It is the first factor which commits the realist to interpretation as a process or project. If the theory's statements are to be asserted, and asserted *as true*, then the realist cannot rest content with uninterpreted or partially interpreted theories: for uninterpreted sentences are not the sort of thing that can be true (or false). The second factor is a constraint on what kind of interpretation the realist can accept (i.e., one which gives a realistic semantics—whatever that might mean). So the compulsion towards interpretation, and towards interpretation of a particular sort, is one of the things which distinguishes the realist from her rivals.

Of course, this isn't to imply that interpretation is unimportant for anti-realists: mandatorily caring about interpretation is a necessary condition for realism, not a sufficient one. As the above taxonomy of positions makes clear, the reductive empiricist is also required to interpret theories; they merely disagree with the realist over what kind of interpretation is appropriate. And there are good reasons for the constructive empiricist to care about interpretation, since they take the provision of a realistic semantics to be part and parcel of presenting a theory for acceptance. But the overall idea remains: attitudes towards the practice of interpretation (compulsory vs.

[3] Unlike constructive empiricists, who maintain that acceptance of a theory as empirically adequate is sufficient to licence its assertion.

[4] Unlike reductive empiricists—such as instrumentalists—who may acknowledge the truth of scientific claims, but only because such claims are understood as "secretly" being claims about observable entities.

supererogatory vs. ill-advised), and towards what kinds of interpretation that practice should seek (realistic vs. deviant), are one of the ways in which different positions in the debate over realism distinguish themselves from one another.

Secondly, the notion of interpretation is not only a means of marking territory within the realism debate; it also bears upon the dialectic of that debate. For consider the virtues which, the realist contends, are such as to warrant a (truth-based) commitment to a scientific theory: explanatory power, unificatory strength, etc. Put aside the issue of whether these virtues do indeed warrant such a commitment, and instead merely note that these are virtues of *interpreted* theories.[5] The first point (above) was intended to show that interpretation is a precondition of the *coherence* of realism. This observation shows that even if one doesn't buy that, it's nevertheless the case that interpretation is a precondition of the *plausibility* of realism. Without interpretation, theories simply would not have the kinds of features which the realist takes as justifications for the realist attitude.

Third, interpreting a theory is a necessary component of figuring out the theory's *commitments*, both ontological and ideological. An uninterpreted theory is just that: a symbolic calculus, with (perhaps) rules governing how the elements of the calculus may be manipulated, but with no indication of how the calculus is of any greater representational significance than a game of Go. In particular, belief in an uninterpreted theory does not bring with it any commitments regarding the nature of the world (indeed, it is not even clear that an uninterpreted theory is the sort of thing which is apt to be the subject of doxastic attitudes). If it was uniquely determined what commitments *would* be involved, in the event that one takes the realist plunge and decides to believe a theory, then we could perhaps claim that the mere application of such a calculus is sufficient to "count" as taking on those commitments. But at least prima facie, there are choices over how a given formal calculus ought to be interpreted.[6] Maybe, after analysis, we will succeed in showing that there is no such multiplicity of interpretative options—but doing so will only be possible after the application of some philosophically rich account of interpretation, so we are still required to develop such an account.

Finally, and relatedly, determining relations of equivalence between theories requires that those theories be interpreted. The heart of the notion of theoretical equivalence is a certain sort of *ecumenicism* with regards to equivalent theories: if theories A and B

[5] I take this observation from [Ruetsche, 2011, Chapter 1].
[6] cf. [Jones, 1991].

are equivalent, then there is no question about which of them one ought to commit oneself to, since advocating the one induces the same commitments as advocating the other. This is why determining conditions for equivalence is interesting and important, as those conditions will tell us when we do or don't need to make choices amongst theories. But it also makes clear that A and B must be *interpreted* theories in order for the question of equivalence to be well-posed, since (as was just observed) it is only an interpreted theory which has an unambiguous roster of commitments. (Note that *conditions* for theoretical equivalence could nevertheless be interpretation-independent, at least to some degree: for such conditions could indicate merely that *if* A and B are interpreted according to the same interpretative scheme, *then* they will come out equivalent—regardless of which scheme one employs.)[7]

So, what kind of process or project could interpretation be, which brings about such results? I wish now to briefly outline one approach to interpretation which is widespread, but—I contend—flawed. I have in mind understandings of theory-interpretation which take it to be analogous to the interpretation of a passage written in a foreign language. In such cases, interpretation is a matter of translating the foreign passage into some antecedently understood tongue. By analogy, then, these approaches to interpretation take some fixed language as "transparent"—as having its meaning already fixed—and conceive of the task of interpreting a theory as being that of translating it into the transparent idiom. Of course, what idiom is taken as transparent will change with the philosophical inclinations of whoever happens to be pursuing this approach at a particular time. For the logical positivists, it was to be a language of sense-data, or a catalogue of verification methods; for Quine, the language of single-sorted first-order logic;[8] for the primitive ontologists, a language describing "local beables".[9] But the underlying conception of interpretation is, I claim, the same in each case.

There are two problems with such an approach. First, since this approach involves pretheoretically privileging some particular model of description, it gives rise to nat-

[7] This puts me at odds with the analysis of [Coffey, 2014], who argues that the only possible account of conditions for theoretical equivalence is the trivial account, according to which A and B are equivalent if their interpretations coincide. I return to this below.

[8] [Quine, 1948]. That said, it may not be entirely fair to present Quine as an advocate of transparent interpretation, in the sense defined here: Quine's view is plausibly better thought of as the claim that *the most scientifically respectable theories* will be those presented in first-order logic, for reasons internal to the purposes and practice of science. If this is correct, then my remarks below should be construed as directed at those who want to apply Quine's method as a form of transparent interpretation, rather than at Quine himself.

[9] See e.g. [Maudlin, 2007d], [Allori et al., 2008].

uralist concerns. Insisting that any acceptable theory must be translatable into the transparent idiom requires imposing constraints on science which have been derived entirely (or almost entirely) from *a priori* philosophical reflection. This concern becomes particularly acute when the demand for transparency is used to direct or constrain the search for theories: for instance, when primitive ontologists demand that any acceptable quantum theory *must* take a certain form.[10] We should be extremely skeptical that the *a priori* reflections of philosophers will offer a better mechanism for theory choice in science than the practice of science does.

Second, there are reasons to be prima facie concerned about the coherence of the project. To offer a translation of the theory into the transparent idiom is to exhibit a theory formulated in the transparent idiom that is claimed to be equivalent to the original theory. But how are such relations of equivalence to be adjudicated? As discussed above, equivalence and inequivalence are relations that hold, properly speaking, between *interpreted* theories. Even granting, for the sake of argument, that any theory in the transparent idiom stands in need of no interpretation, the source theory is—by definition—uninterpreted. So what fixes whether a proposed translation into transparency is acceptable or not?[11]

1.3 Interpretation and semantics

Towards developing an alternative, let's return to how I introduced the various spaces in our toy Newtonian theory. My informal introduction of the variables ranging over those spaces included a specification of what kind of ontological constituent they were to represent: $\mathbf{x}$ as a point of space, t as the time, etc. If we understand interpretation as translation, then we could treat these statements as *stipulations*, whose function is to imbue the theory with meaning—as telling us what the theory is about. In Lockean terminology, these serve as the "nominal definitions"—fixing reference to the subject-matter of a discourse—with the dynamical laws then serving as (attempts at)

[10] e.g. [Egg and Esfeld, 2014], [Esfeld et al., 2014]

[11] [Coffey, 2014] claims that there is no interesting problem of theoretical equivalence, since being theoretically equivalent is just a matter of two theories' having the same interpretation. As indicated above, I agree that equivalence is a relation that holds between interpreted theories: for the reason discussed here, however, I demur from Coffey's claim that such interpretation is to be cashed out in terms of "picturing" the two theories (and especially, from the claim that the process of interpretation renders equivalence a moot or uninteresting topic). We will discuss theoretical equivalence in more detail in Chapters 3–5.

"real definitions", which state the nature of the things so referred to.[12] I have already indicated my disagreement with such an approach. So what's the alternative? To hold that *a theory is the arbiter of its own meaning*: that ultimately, the terms deployed in a theory must derive their meaning from the role they play in that theory. So it is not that we stipulate the ontological categories of what the theory postulates in advance and *ex cathedra*; rather, it is by attending closely to the dynamics that we can try to figure out what ontology the theory proclaims.[13]

The challenge is then to cash this out: to articulate a notion of what interpreting a scientific theory is which amounts to more than mere gesturing, but which is sufficiently minimal to avoid prejudicing questions we might want to ask. The solution is to be serious about the idea of "interpretation". Interpreting a theory is no more and no less than the project of delineating the internal semantic relationships within the theory, by indicating how the (formally conceived) representational structures of the theory line up with one another. Is this sentence synonymous with that sentence? Is this model equivalent to that model? What are the inferential relationships amongst these sentences? And so on and so forth.

In order to explain what I mean by this, it will be helpful to consider the case of theories phrased in terms of first-order logic (rather than differential equations). For, despite my disagreements with Quine or the positivists, I do not wish to disavow the application of first-order logic to problems in philosophy of science—far from it. However, rather than approaching first-order logic as a formalism into which other theories must (for the purposes of interpretation) be put, I think it is valuable as a formalism to which other theories may be *compared*. That is, I think that first-order theories are most helpfully thought of as doing the same sort of things as theories couched in terms of fibre bundles, or Hilbert spaces, or whatever; it is just that they do so using somewhat simpler and more rudimentary means.

To see this comparison, I briefly review the formalism of first-order model theory (as much to fix notation as anything else).[14] The logical vocabulary consists of the following:

[12] This way of putting things is due to Tim Maudlin.

[13] Note that doing so should—if all goes well!—mean that the theory "has a unique interpretation", in the sense that one interpretation will stand out as that favoured by reflection upon the theory. I don't take that to be in tension with my earlier remark that multiple interpretations are *prima facie* possible for a theory: that one interpretation is the best available doesn't mean that other interpretations are not possible in some more minimal sense (i.e. there may well be other, less good, ways of making sense of the theory).

[14] I draw primarily on [Hodges, 1997].

- A set Var of *variables*
- The unary logical connective $\neg$ (*negation*)
- The binary logical connective $\wedge$ (*conjunction*)
- The *universal quantifier* $\forall$
- The *parentheses* (and)

Every model-theoretic language[15] we will consider agrees on the logical vocabulary. Thus, the languages are distinguished from one another by their non-logical vocabularies. We will refer to a choice of non-logical vocabulary as a *signature*: a set Σ of monadic and polyadic predicates.[16] Given a signature Σ, one can define the set $\mathrm{Form}(\Sigma)$ of well-formed Σ-*formulae*, using the standard compositional rules of predicate logic:

1. If $\Pi \in \Sigma$ is of arity $n \in \mathbb{N}$, and if $\xi_1, \ldots, \xi_n \in \mathrm{Var}$, then $\Pi\xi_1 \ldots \xi_n$ is a formula
2. If ψ is a formula, then $\neg\psi$ is a formula
3. If ψ_1 and ψ_2 are formulae, then $(\psi_1 \wedge \psi_2)$ is a formula
4. If ψ is a formula, and $\xi \in \mathrm{Var}$, then $\forall\xi\psi$ is a formula

We will use the logical connectives $\vee$ and $\rightarrow$, and the quantifier $\exists$, as abbreviations:

- $(\psi_1 \vee \psi_2)$ abbreviates $\neg(\neg\psi_1 \wedge \neg\psi_2)$
- $(\psi_1 \rightarrow \psi_2)$ abbreviates $\neg(\psi_1 \wedge \neg\psi_2)$
- $\exists\xi\psi$ abbreviates $\neg\forall\xi\neg\psi$

The set of Σ-*sentences* is the set of closed Σ-formulae (formulae with no free variables).

The semantics for a language with signature Σ is given by Σ-*pictures*.[17] A Σ-picture consists of

- A set $D_{\mathcal{S}}$

[15] Object language, that is; the metalanguage will just be mathematical English throughout.

[16] We will only consider *relational* model theory, i.e., model theory without constants or function-symbols. This is simply in order to avoid certain complications that arise from function-symbols, especially when (as in Chapter 3 below) we consider the notions of definitions and translations. See [Hodges, 1997] for details.

[17] This terminology is non-standard. The more standard term is a Σ-*structure*: I have changed the terminology in order to avoid confusion between informal use of "structure" and its use as a term of art.

- For each $n \in \mathbb{N}$, for each n-ary predicate $\Pi \in \Sigma$, a function $\Pi_{\mathcal{S}} : \mathcal{D}_{\mathcal{S}}^n \to \{0, 1\}$ (the *characteristic function* of Π in $\mathcal{S}$)

Note that this is a little non-standard; it is more common to see a Σ-picture defined as comprising

- A set $D_{\mathcal{S}}$ (the *domain* of $\mathcal{S}$)
- For each $n \in \mathbb{N}$, for each n-ary predicate $\Pi \in \Sigma$, a set $\Pi^{\mathcal{S}} \subseteq D_{\mathcal{S}}^n$ of n-tuples of members of $D_{\mathcal{S}}$ (the *extension* of Π in $\mathcal{S}$)

However, any subset S of a set A is interchangeable with its characteristic function: the function $k_S : A \to \{0, 1\}$ defined by

$$k_S(a) = \begin{cases} 1 & \text{if } a \in S \\ 0 & \text{otherwise} \end{cases} \tag{1.6}$$

So all we are doing here is using characteristic functions in place of extensions. The main reason we do this is that it helps make clearer the analogy between model theory and physical theories (although it will also be useful in doing some expository work in Chapter 2, and leads to a marginally more elegant clause for the truth-values of atomic formulae).

Given a Σ-picture $\mathcal{S}$, a *variable-assignment for* $\mathcal{S}$ is a map $g : \mathrm{Var} \to D_{\mathcal{S}}$. A Σ-picture $\mathcal{S}$, together with a variable assignment g for $\mathcal{S}$, determines a truth-value $|\phi|_g^{\mathcal{S}}$ for every Σ-formula ϕ. These truth-values are determined as follows:

(a) $|\Pi\xi_1 \ldots \xi_n|_g^{\mathcal{S}} = \Pi_{\mathcal{S}}(g(\xi_1), \ldots, g(\xi_n))$

(b) $|\neg\psi|_g^{\mathcal{S}} = 1$ iff $|\psi|_g^{\mathcal{S}} = 0$

(c) $|(\psi_1 \wedge \psi_2)|_g^{\mathcal{S}} = 1$ iff $|\psi_1|_g^{\mathcal{S}} = 1$ and $|\psi_2|_g^{\mathcal{S}} = 1$

(d) $|\forall\xi\psi|_g^{\mathcal{S}} = 1$ iff for every $a \in \mathcal{D}_{\mathcal{S}}$, $|\phi|_{g_a^\xi}^{\mathcal{S}} = 1$

where the variable-assignment g_a^ξ in clause ((d)) is defined by

$$g_a^\xi(\zeta) = \begin{cases} g(\zeta) & \text{if } \zeta \neq \xi \\ a & \text{if } \zeta = \xi \end{cases} \tag{1.7}$$

If $\mathcal{S}$ together with g makes a formula ϕ true, we write $\mathcal{S} \models_g \phi$; if ϕ is a sentence, then the variable-assignment no longer matters, and we write simply $\mathcal{S} \models \phi$.

A *theory* $\mathbb{T}$ in the signature Σ (for short, Σ-theory) is a set of Σ-sentences.[18] A Σ-picture $\mathcal{M}$ is said to be a *model* of $\mathbb{T}$ if it satisfies each member of $\mathbb{T}$; we denote the class of all models of $\mathbb{T}$ by $\mathrm{Mod}(\mathbb{T})$. Finally, $\mathbb{T}$ *entails* a Σ-sentence ϕ just in case $\mathcal{M} \models \phi$ for every $\mathcal{M} \in \mathrm{Mod}(\mathbb{T})$; this will be denoted by $T \vDash \phi$.[19] Hence, a first-order theory exhibits the same tripartite structure that we saw in section 1.1. There is a specification of the kinds of mathematical structures that will be used for representation (i.e. Σ-pictures). There is a collection of syntactically given conditions (i.e. $\mathbb{T}$). There is a subclass of the representational structures, privileged in virtue of fulfilling the stated conditions (i.e. $\mathrm{Mod}(\mathbb{T})$).

Furthermore, we can even see analogies between the intrinsic workings of the representational structures in either case. After all, each model of the theory in §1.1 consists of the following:

- A space $T \cong \mathbb{R}$
- A function $f : T \to X$

Comparing this to a Σ-picture $\mathcal{S}$, we can see a certain degree of analogy (although with some differences). The base set $D_{\mathcal{S}}$ of $\mathcal{S}$ is analogous to the base space T, although T carries significantly more structure than a mere set—the structure, indeed, of the real numbers. (This difference will become significant when we discuss external symmetries in Chapter 5.) Note also that whereas Σ-pictures have varying domains, the domain of individuals for the Newtonian theory is always the same (i.e., T). We'll see later that for more general physical theories, this need not be the case; indeed, we could well have allowed that models of the Newtonian theory take intervals in T, rather than the whole of T, as their domain. The function f is analogous to (one of) the functions $\Pi_{\mathcal{S}}$, indicating what each element of T is like. However, it takes values in X rather than in $\{0, 1\}$. Intuitively, this difference is because whereas predicates in model theory are standardly assumed to be binary (i.e. to either apply or fail to apply to a given n-tuple of individuals), position is a *determinable* predicate: there are as many ways for a pointlike individual to be, position-wise, as there are points of space.[20]

Nevertheless, these differences notwithstanding, we can see a certain level of analogy between the two formalisms: we can think of a model of the Newtonian theory as

[18] In accordance with standard practice in model theory, I don't require theories to be deductively closed.

[19] The symbols for satisfaction and entailment are unfortunately similar: the former is $\models$, whilst the latter is $\vDash$. Context will make clear what is meant in any given case, however.

[20] For discussion of some of the (problematic) influences of this feature of first-order logic on metaphysical theorising, see [Hall, 2012].

describing the distribution of a monadic determinable property (position) over some set of individuals (particle-stages).[21] Hence, the methodology for this thesis will be to initially confine attention to the formalism of model theory, and only after treating that to turn to issues arising from formalisms arising in more realistic physical theories. There are three advantages to doing this. First, various of the ideas that I wish to explore can be developed in model theory: hence, we will be able to investigate these issues without having to worry about anything more than elementary mathematics. Secondly, this will allow us to make use of the well-developed and -understood criteria of translation and equivalence that are found in model theory, before we engage in trying to extend such notions to more complex settings. Thirdly, this will also make it easier for us to make contact with the philosophical literature on topics such as anti-haecceitism or anti-quidditism. The conceptual framework for much of this literature is that of objects, standing in networks of (non-determinable) properties and relations: i.e., a conceptual framework that is most naturally formalised by first-order model theory.

I can now say more about how I am thinking of interpretation. We've seen already that part of what it is to give a theory is to provide a semantics for it: i.e., some class of mathematical structures which systematically bestow truth-values upon the sentences of the theory's language. However, we need not treat that semantics as immediately codifying all of a theory's commitments; the semantics provided as part of an (uninterpreted) theory is merely a *putative* semantics, whose role is to characterise the background logic. But this means that we can understand interpreting a theory as a matter of providing a *genuine* semantics, i.e., of providing a class of structures (systematically constructed from the structures of the putative semantics) which do precisely capture the commitments of the theory. Throughout this thesis, what that will amount to is giving some kind of equivalence relation on the putative models, and taking the genuine models to correspond to the equivalence classes thereby obtained.[22]

For many purposes, such an interpretative task will only constitute *part* of what interpreting a theory involves. The assimilation of the theoretical vocabulary of a scientific theory to the quotidian vocabulary of day-to-day usage, and the corresponding

[21] This exploiting the fact that T can equally well be thought of as representing time, or as representing the instantaneous stages of a particle (along the lines of the "stage theory" defended by [Sider, 1996]); it seems more natural to take such stages, rather than instants of time, to be the subject of predication here.

[22] This isn't to say that interpretative work could not be done by specifying other inter-model relationships: for example, that certain models are limiting or reducing cases of others. But these other relationships will not be our concern here.

connection of the theoretical architecture to our overall picture of the world, plays an enormously significant role in endowing the theory with semantic content. Of particular importance (both conceptually, and in terms of the history of philosophy of science) is the particular kind of assimilation that places the theory within the grip of experimental science: which identifies certain terms of the theory, for example, with the kinds of things that precipitate certain regularities in the behaviour of our detectors. This formulation is intended to capture both the trivial case in which certain terms of the theory are held to be *directly* detectable (say, when we identify "x" as the position of Jupiter) and the non-trivial case where such detection is mediated by an apparatus of one sort or another (say, the characterisation of "S_x" as a quantity which—*inter alia*—affects the behaviour of a Stern-Gerlach apparatus). Note that the latter kind of stipulation is usually[23] appropriate only if one has some theory of the apparatus which demonstrates and explains this mediation;[24] and *that* theory will be brought into epistemic contact with us by the trivial kind of assimilation—identifying its "pointer variables" as, for example, the position of the pointer.

I agree that this exogenous form of interpretation is important and interesting, and worthy of analysis. Where I differ from the advocate of interpretation by translation is in allowing that it is not the only form of interpretation: it is also important and interesting to explicate the semantic architecture of our theories, and complement whatever exogenous analysis we give with an *endogenous* account of their representational character. In this thesis, I will confine my attention to the latter, and leave the task of exogenous interpretation to future work—and without denying the importance of such work. In particular, giving a full analysis of symmetries of subsystems[25] would require an analysis of exogenous interpretation.

The main consequence of this stance is a difference in how claims such as "x represents the position of the particle" bear upon the interpretation of the theory, and (correlatively) the kind of authority they are taken to enjoy. In exogenous interpretation, these claims possess a particular kind of semantic authority: they are *definitions*, and so serve to bind the term to express a certain kind of thing. On the endogenous view, these claims differ only in degree from assertions such as (1.1), not in kind. In particular, claims of this kind are *fallible*. If the postulated dynamics for a theoretical term is not appropriate to a certain kind of ontological category, then you can swear until you are

[23] The exception would be where a theoretical term is introduced "phenomenologically", as whatever substance or property is responsible for some observed phenomenon.

[24] This is, of course, just the theory-ladenness of observation (or at least, one version of that phenomenon).

[25] See e.g. [Kosso, 2000], [Brading and Brown, 2004], [Healey, 2009], [Greaves and Wallace, 2014].

blue in the face that it is of this kind—but to no avail. We will see a concrete example of this phenomenon in Section 5.5, when we consider issues about determining which structures in a theory count as the spacetime structures.

1.4 Model theory as modal metaphysics

Finally, I wish to say a little more about the sense in which the genuine semantics characterises the commitments of the interpreted theory (i.e., the commitments which the theory is being interpreted as having). It does so in the following sense: by providing an account of what is possible according to the theory, and of what kinds of variations there are between the possibilities allowed by the theory. To explain this, we must (briefly) engage with some of issues from the philosophy of modality. As a preliminary, recall one of the major motivations for engaging in modal metaphysics (i.e., reflections on the nature of possible worlds): the desire to obtain a coherent account of modal thought and talk.

However, what constitutes such talk? Here, it is helpful to follow Quine[26] in distinguishing three degrees to which our talk might be modal. The first (and weakest) degree is that of using modality as a "semantical predicate", applicable to sentences for which we have names, e.g.

"$9 > 5$" is necessary.

The second degree is that of admitting modality as an operator which may take sentences (closed formulae) as arguments, thereby yielding other sentences. As Quine puts it, in ordinary English this is marked by the shift from the predicate "is necessary", which may be combined with a noun to yield a sentence, to the adverb "necessarily", which combines with a sentence to form a sentence, as in

Necessarily, $9 > 5$.

The third degree admits modality as an operator which may take *open* formulae as arguments. This enables the formation of open modalised formulae, and consequently formulae whose modal operator falls within the scope of a quantifier, e.g

There is something which is necessarily greater than 5.

[26] [Quine, 1953b]

For now, let us suppose that we are only interested in accounting for the first two degrees to which modality may be entangled in our language; we will worry about the third degree in the next chapter. And for the purposes of concreteness, suppose that the language with which this modality has become involved is a first-order language with signature Σ. So we are envisioning a (somewhat peculiar) language, constituting a fragment of quantified modal logic (QML). Its logical vocabulary consists of the logical vocabulary of first-order predicate logic, together with the necessity operator $\Box$; its syntactic formation rules are the same as those for the first-order extensional language based on Σ, but supplemented by

5. If σ is a sentence, then $\Box\sigma$ is a formula (in fact, a sentence).

The possibility operator, $\Diamond$, will be taken as a defined symbol: $\Diamond\phi$ abbreviates $\neg\Box\neg\phi$.

Suppose we now ask how truth-values could be assigned to the formulae of this language, in such a way that they do not clash with the strictures imposed by the meanings of the non-logical vocabulary. Such strictures include, for example, the injunctions to never assign truth to $(P \wedge \neg P)$ to always assign truth to $\forall x(x = x)$; and to not assign truth to $\Box(P \to Q)$ but falsity to $(\Box P \to \Box Q)$. And the answer is that it is both necessary and sufficient for obeying such strictures that the truth-values be generated by a Σ-*based Kripke-picture*, where a Σ-based Kripke-picture consists of the following data:

- A set W (of *worlds*)
- A relation $R \subseteq W \times W$
- For each world $w \in W$:
  - A set D_w
  - For each n-ary predicate Π, a function $\Pi_w : (D_w)^n \to \{0, 1\}$

A Σ-based Kripke-picture generates a truth-value for ϕ at w, relative to a variable-assignment $g : \mathrm{Var} \to D_w$, as follows:

(a)′ $|\Pi\xi_1 \ldots \xi_n|^w_g = \Pi_w(g(\xi_1), \ldots, g(\xi_n))$

(b)′ $|\neg\psi|^w_g = 1$ iff $|\psi|^w_g = 0$

(c)′ $|(\psi_1 \wedge \psi_2)|^w_g = 1$ iff $|\psi_1|^w_g = 1$ and $|\psi_2|^w_g = 1$

(d)′ $|\forall\xi\psi|^w_g = 1$ iff for every $a \in D_w$, $|\phi|^w_{g^\xi_a} = 1$

(e)′ $|\Box\sigma|^w_g = 1$ iff for every $v \in W$ such that wRv, $|\sigma|^v_g = 1$

Thus, we reach the following conclusion: insofar as our modal thought and talk manifests only the first or second degree of modal involvement, and supposing the underlying extensional language to be a first-order Σ-language, giving a possible-worlds story sufficient to account for that thought or talk means giving a Kripke-picture where each of the worlds has the structure of a Σ-picture. (There may be other ways of accounting for modal thought or talk besides a possible-worlds story, though.) It is easy to see how this generalises to cases where the underlying extensional apparatus is not a first-order Σ-language, but some other formalism: just change the worlds from Σ-pictures to a picture of the language in question. For instance, if the underlying language is that appropriate to representing the motion of a single Newtonian particle, we want a Kripke-picture whose worlds are functions of type $T \to X$. In other words, a coherent account of modal thought and talk—insofar as that talk can be characterised by taking a language L and modalising it to the first or second degree—just is the provision of an L-based Kripke-picture, in the sense that each world of the Kripke-picture has the structure of an L-picture. Any such Kripke-picture will assigns truth-value to the elements of the talk in a way respecting their logico-grammatical relationships. However, merely being coherent (in this sense) is not enough. What we want is a commitment as to which (modal) sentences are true and which are false. Thus, I will understand the project of modal metaphysics as that of explaining how one ought to go about finding a Kripke-picture which gets the modal facts right; i.e., of finding a Kripke-picture which correctly represents modal reality.

However, which Kripke-picture of that kind should we be advocating? In this thesis, my main interest is not in general metaphysical possibility, but rather in nomic possibility: the sense in which something is possible or impossible according to the laws. Or, to be a little more accurate, I am interested in the subgenres of nomic possibility associated to *different* sets of laws, rather than (just) the variety associated to the one true set of laws (whatever they may be, and should such a thing exist). Given that, there is a very natural candidate for what Kripke-picture to advocate as the correct depiction of modality relative to a given set of laws: namely, the Kripke-picture whose worlds are the (genuine) models of the the (interpreted) theory expressing those laws.[27] This expresses the fact that we generally explicate theory-relative possibility by looking

[27]Note that "models" here is being used in a general sense, encompassing both Tarski-structures satisfying a given first-order theory and objects that form the solutions to theories expressed by differential equations (in line with my remarks in Section 1.3).

to what sorts of things are true in some model or other of the theory. Is it possible, according to General Relativity that black holes exist? Yes, because there are models of the theory according to which black holes exist. Is it possible, according to quantum mechanics, for a particle to spontaneously accelerate? No, because there is no model of the theory in which that is the case. Thus, to (endogenously) interpret a theory is to provide an account of possibility according to that theory: declaring which models of the theory (under the putative semantics) correspond to the same possible world (i.e., the same model under the genuine semantics).

Let us consider a specific example. Suppose that the theory in question is that expressing the philosopher's favourite law of nature, "all Fs are G": this is a theory $\mathbb{T}_0$ of signature $\{F, G\}$, whose sole axiom is

$$\forall x(Fx \to Gx) \tag{1.8}$$

Moreover, take the "literal" interpretation of $\mathbb{T}_0$, according to which we identify the possible worlds (possible, that is, according to $\mathbb{T}_0$) with the models of $\mathbb{T}_0$. This class of models is an $\{F, G\}$-based Kripke-picture, according to which (for instance) $\Box\forall x(Fx \to Gx)$ is true, but $\neg\Diamond\forall xFx$ is false.

This view is, of course, reminiscent of Carnap's proposal to explicate necessity in terms of logical truth.[28] More specifically, Carnap proposes adopting the following convention for his necessity operator, N (where "L-true" means "logically valid"):

> For any sentence '…', 'N(…)' is true if and only if '…' is L-true.[29]

However, there are a few important differences. One is that Carnap is explicit that this is intended as an analysis of *logical* necessity, not nomological necessity. As such, Carnap's proposal would correspond to the special case of the above scheme where the theory in question is the empty theory (so that the models of the theory are *all* the pictures of the appropriate type). More significantly, however, Carnap appears to propose this convention *as a semantics for modal logic*, i.e., as a means of determining the validity or invalidity of modal inferences. For, he claims, the above convention enables us to acclaim certain sentences of modal logic as L-true (or L-false): if that convention, together with his conventions for non-modal logic,[30] suffice to determine the sentence as true (respectively, false), then the sentence in question is L-true (respectively, L-false).

[28] [Carnap, 1956]
[29] [Carnap, 1956, p. 174]
[30] Which are the same as those used here, except that Carnap employs a substitutional account of quantification.

Here, I definitely part company from Carnap. As I stressed, the account of possible worlds given here is intended to furnish us with a *specific* Kripke picture, i.e., that in which the worlds are the models of the associated theory. But attention to one Kripke picture in particular is not sufficient for correctly characterising the notion of validity in modal logic, any more than attention to one Tarski picture is sufficient for characterising the notion of validity in non-modal logic. It is for this reason that Carnap's account of validity in modal logic is defective. For instance, it is a consequence of his analysis that "Every sentence of the form 'N…' is L-determinate [i.e., logically valid or logically invalid]."[31] This results from the fact that he is considering only one Kripke picture $\mathcal{K}$ (that in which the worlds are all and only the Tarski pictures), and identifies logical validity with truth in $\mathcal{K}$, rather than truth in all Kripke pictures. So, because P is true in some Tarski-pictures, $\Diamond P$ is true in $\mathcal{K}$, and so is held by Carnap to be *logically valid*. This is not only implausible in itself, but has the consequence that logical validity is not closed under uniform substitution:[32] $(Q \wedge \neg Q)$ is false in all Tarski-pictures, so $\Diamond(Q \wedge \neg Q)$ is false in $\mathcal{K}$, and hence is not logically invalid (indeed, is logically invalid). The point is that a sentence such as $\Diamond P$ is not true *in virtue of their form alone*, or true *independently of the meaning assigned to* P: if P is assigned to a necessarily false proposition, then $\Diamond P$ is false.

Therefore, as an analysis of logical inference—of what inferences are good, or what sentences valid, independently of the meanings of their terms—Carnap's convention is no good. However, this does not impugn its status as an analysis of which modal sentences are *true*. Indeed, understood in those terms, it surely has to be correct: what else could it be for a first-order sentence σ to be logically necessary than for it to be logically valid, i.e., true in all Tarski-pictures? Consider the non-modal analogue. It would be a disaster to hold that there is some special Tarski picture, $\mathcal{S}$, such that a sentence is logically valid if and only if it is true in $\mathcal{S}$. But there is, of course, no problem with affirming that $\mathcal{S}$ is the correct representation of what the facts are, that a sentence is (actually) true if and only if it is true in $\mathcal{S}$. One obvious difference is that the means by which we come to affirm $\mathcal{S}$ as a good representation of the actual facts will presumably be empirical in nature, whereas the means by which we come to affirm $\mathcal{K}$ as a good representation of the modal facts (concerning logical modality) are not. But that should not be especially surprising. As empiricists are wont to remind us, there is no obvious means by which we could gain empirical access to the modal facts. One virtue of the

[31] [Carnap, 1956, p. 175]
[32] I take this observation from [Williamson, 2013, §2.8].

analysis proposed here is that it provides a means by which we could have any access to modal facts at all.[33]

Despite its naturalness (especially, the way it meshes with the way working scientists tend to talk of possibility), this view of possible worlds has not been very popular amongst metaphysicians. Indeed, I am not sure that it has been explicitly defended. Its closest relative, so far as I am aware, is a view Lewis calls "pictorial ersatzism", and describes thus:

> Perhaps, apart from its abstractness, an ersatz world should be like a picture in a generalized sense: a sense in which a statue counts as a three-dimensional picture, and and a working model counts as a four-dimensional picture. And let it be an idealised picture, infinite in extent and many-dimensional if need be, which represents the concrete world in its entirety and all its detail. [...]
>
> A picture represents by isomorphism. [...]
>
> With our ordinary pictures, the isomorphism is limited. [...] These limitations of isomorphism are patched over by means of extensive and complicated conventional understandings; at this point pictorial representation is language-like. We don't want that. We were looking for an *alternative* to linguistic ersatz worlds [i.e., worlds conceived of as sets of sentences], so let us have something as different as we can. Since we think of our new ersatz worlds as *idealised* pictures, we may safely suppose that they represent entirely by isomorphism. [...]
>
> Isomorphic representation, pure and simple, works by composition of parts and by identity of properties and relations. So our pictorial ersatz world must consist of parts, with divese properties, arranged in a certain way. Thereby by it represents the concrete world as isomorphic to it: as consisting of corresponding parts, with the same properties, arranged in the same way.[34]

[33] Not that such access will be full or complete: Williamson observes that there is no recursive procedure by which one could determine what sentences are true or false in $\mathcal{K}$ [Williamson, 2013, §2.8]. He presents this as a further problem for Carnap's analysis, since it means that first-order modal logic would have no sound and complete axiomatisation. On the view here, however, it is not terribly surprising—why expect that we could fully discover what all the modal facts are, even given a definite criterion for what those facts are?

[34] [Lewis, 1986, p. 166]

Pictorial ersatzism seems to generally be reckoned to not have many attractions or defenders.[35] However, there are a number of differences between pictorial ersatzism and the view outlined here.

First, there is a greater element of language-dependence here than on Lewis' account, since the worlds are to be taken as abstractions from models—and as I stressed in section 1.1, models are manifestly constructions that depend upon the language to hand (as they must be, if their capacity to generate truth-values for the language is to work). That said, we will see in Chapter 3 that one of the things we abstract away from will be this dependence on language. So although we employ language-dependent ingredients in building the worlds, the finished structure is language-independent.

Second, there is no requirement that any model represents how things could be "in its entirety and all its detail". As indicated above, this is intended as an account of the various and shifting modalities that we get out of the use and application of theories, rather than the grand and transcendent modality in which Lewis is interested.

The most important difference is that the pictorial worlds are supposed, by Lewis, to literally manifest the same properties as the world they represent. This is significant, since it means that worlds in our sense are not subject to Lewis' most serious complaint against pictorial ersatzism: that pictorial ersatzism is no less ontologically profligate than his own modal realism, instead just being less honest about it. For, Lewis complains, the pictorial ersatz worlds are inhabited by objects having all the same properties as actual objects (redness, wisdom, charge, what have you) save for some mysterious property of "concreteness". So if anything, the pictorial ersatzer has *more* commitments than Lewis: they believe in all the same structural stuff as Lewis does, but also in this magical concreteness property! As a complaint against pictorial ersatzism, I think this is quite right. As a complaint against the view here, however, it misses the mark. The parts of a model do not literally possess the same properties as the objects that they would represent, in the event that the model represented actuality; hence, nor do the worlds abstracted from them. The price of using worlds in our sense is that it is no longer immediate, given a world, what properties it represents. In Chapter 3, we will consider some of the consequences of this.

[35] For example, "[Pictorial ersatzism is] an odd, hybrid view that, I suspect, no one has or ever will hold" [Bricker, 2006, p. 42]; "pictorial ersatzism is a puzzling view, and may have no actual adherents" [Nolan, 2015, p. 64].

2 Anti-haecceitism

> The city streets are littered
> with the casualties;
> the could haves,
> and the should haves,
> and the would have beens...
>
> Pulp, *Tomorrow Never Lies*

2.1 Isomorphism and equivalence

To begin: suppose that we are considering how best to interpret a theory given in the formalism of first-order logic, such as the theory (1.8) encountered at the end of the previous chapter. The putative semantics for such a theory is that we saw in §1.3: each model is a set, equipped with extensions for the predicates in Σ.

For our purposes, the notable feature about this semantics is the following: it will be the case that we obtain distinct pairs of *isomorphic* models. Recall that an isomorphism from one Σ-picture, $\mathcal{M}$, to another, $\mathcal{N}$, is a bijection $h : D_{\mathcal{M}} \to D_{\mathcal{N}}$ such that, for every $n \in \mathbb{N}$, every n-ary $\Pi \in \Sigma$, and any $a_1, \ldots, a_n \in D_{\mathcal{M}}$,

$$\Pi^{\mathcal{M}}(a_1, \ldots, a_n) = \Pi^{\mathcal{N}}(h(a_1), \ldots, h(a_n)) \tag{2.1}$$

If $\mathcal{M}$ and $\mathcal{N}$ are isomorphic to one another, we will write $\mathcal{M} \cong \mathcal{N}$. What is the difference between such a pair of models? If they correspond to distinct worlds, then what do these worlds disagree on? Prima facie, they agree both on the question of how many individuals exist, and on the distribution of (qualitative) properties and relations. That is all we need for now; we will shortly consider what, if anything, they could be taken to disagree on. In this chapter, I defend the following claim: the right interpretation of such a theory is one in which isomorphic models of the putative semantics are equivalent (i.e., correspond to a single element of the genuine semantics).

Or, relating it to the conception of possibility introduced in §1.4: we should not identify possible worlds with models, but rather with equivalence classes of such models under isomorphism. For the purposes of convenience, I will refer to such equivalence classes as WORLDS. This is consistent with the claim that some non-isomorphic models are equivalent: isomorphism of models is sufficient for equivalence, but not necessary (as we shall see in the next chapter).

Why is such an interpretation to be preferred? I offer the following considerations. First, there is the fact that the theory itself draws no distinctions between isomorphic models. Indeed, in an important sense the entire language of the theory draws no such distinction: if $\mathcal{M} \cong \mathcal{N}$, then for any sentence σ, $|\sigma|^{\mathcal{M}} = |\sigma|^{\mathcal{N}}$ (i.e., isomorphism is sufficient for elementary equivalence). As a result, equivalence classes of isomorphic models will do just as well as models in fixing unambiguous truth-values for closed sentences; and hence, are just as well-suited to be the points of a Σ-based Kripke picture capable of dealing with the first two grades of modal involvement.[1]

That said, we should note that the treatment of *open* formulae here is a little trickier. Even if we relativise to a variable-assignment g, open formulae do not (in general) receive unambiguous truth-verdicts from WORLDS. For example, suppose that $g(x) = a \in D_v$; and suppose further that $v \cong w$, but $P^v(a) \neq P^w(a)$ (which could be because a lies outwith the domain of P^w). Then $|Px|^v_g \neq |Px|^w_g$ (again, possible because the latter is undefined). Consequently, $[v]$ does not unambiguously determine a truth-value for Px.

The solution to this problem is to recognise that variable-assignments aren't really doing their job, in a context in which we are using WORLDS as the representational constructs of the semantics. The point of a variable-assignment, recall, is to map each variable to an individual. But a variable-assignment $g : \mathrm{Var} \to D_w$ isn't doing that in any meaningful sense: the elements of D_w are just used in an arbitrary representative of $[v]$. Instead, variables should be assigned to the domain of a specific WORLD; and they should be assigned to individuals *qua* qualitative roles, rather than *qua* the entities underpinning those roles.[2] To implement this idea, define an INDIVIDUAL of W to be a map $A : w \in W \mapsto A_w \in D_w$ such that for any $w, v \in W$, there is some isomorphism

[1] One concern with this argument is that it overgeneralises, and would have us run together elementarily equivalent models. But isomorphic models are not just elementarily equivalent in the original language; they will remain elementarily equivalent under arbitrary strengthenings of the logical or expressive capacities of the language. (Relatedly, the whole point of non-isomorphic yet elementarily equivalent models is that they may be distinguished using the resources of the metalanguage; that is, they are not elementarily equivalent in the metalanguage.)

[2] cf. [Rynasiewicz, 1994]

$f : w \to v$ such that

$$A_v = f(A_w) \tag{2.2}$$

Intuitively, the idea is this: suppose that a plays some particular qualitative role in v. If $w \cong v$, then it need not be the case that a plays the same role in w as it does in v, but it is guaranteed that *some* individual in w plays the same qualitative role in w that a does in v. Such an individual will be the image of a under some isomorphism $f : v \to w$. Say that a set of INDIVIDUALS of some WORLD W is *coherent* if for any $w, v \in W$, there is some isomorphism $f : w \to v$ such that for every INDIVIDUAL A in the set, $A_v = f(A_w)$. For W, let D_W be a maximal coherent set of INDIVIDUALS for W.[3]

We can then define a variable-ASSIGNMENT for W to be a map $G : \mathrm{Var} \to D_W$. Equivalently, it may be defined as a map $G : w \in W \mapsto G_w$, where $G_w : \mathrm{Var} \to D_w$ is a variable-assignment for w, such that given any $w, v \in W$, there is some isomorphism $f : w \to v$ such that

$$G_v = f \circ G_w \tag{2.3}$$

That is, if a pair of variable-assignments $g : \mathrm{Var} \to D_w$ and $g' : \mathrm{Var} \to D_v$ are such that $g' = f \circ g$, then g' and g agree on what qualitative role any variable ξ gets mapped to (e.g. if $g(\xi) = a$, then $g'(\xi) = f(a)$). So we relativise variable-assignments to the selection of some particular set of individuals to populate the roles—i.e., to the choice of some particular Kripke-world $w \in W$. With this new apparatus, WORLDS can evaluate open formulae too (relative to variable-ASSIGNMENTS).

The second consideration in favour of identifying isomorphic models as equivalent is that drawing strong distinctions between isomorphic mathematical objects runs counter to accepted practice in mathematics.[4] One reason to be suspicious of such distinctions is that we seemingly have no means of referring determinately to some one mathematical object, rather than to any of its isomorphic cousins.[5] As just discussed, there is no closed description available (in this or any other language) which could pick out one of them in particular. So insofar as our reference to such objects is fixed or brought about by appeal to closed descriptions, we cannot guarantee reference. Nor is the problem solved by supposing that some more fine-grained description is encoded in subtle features of our communal linguistic practice: *any* description of this kind will apply equally well to any isomorphic model. Finally, note that we cannot directly ostend the models themselves, given that they are abstract acausal structures. So there

[3] The appeal to coherence is necessary to allow for cases where W admits non-trivial automorphisms.
[4] [Weatherall, 2016] argues for this idea, although for somewhat different reasons.
[5] cf. [Benacerraf, 1965]

is a threat of semantic underdetermination, but one which can be resolved by taking isomorphic models to be equivalent: by taking isomorphic models to correspond to the same possible world, we recover determinate reference to the possible worlds.

Of course, this is not to say that doing so is the only conceivable way of resolving the underdetermination. Perhaps our reference-seeking resources are richer than the above argument allows: one could hold, for example, that indexicals, names, or other means of direct reference might give us a way to pick out one model in particular. I defer discussing this proposal until §2.2, since it depends on some tricky issues about *de re* representation. (I will, however, assume throughout that the semantic facts must be fixed by linguistic practice, externalist considerations, or some such non-semantic matters: that is, I assume "semantic supervenience" and deny "semantic sovereignty".[6] If the facts about reference can outstrip the non-semantic facts, then determinate reference to anything you like is, of course, easy to come by.)

Third, asserting the inequivalence of isomorphic models leads us into epistemic difficulties. Intuitively, the problem is that isomorphic possible worlds are epistemically indistinguishable: at least *prima facie*, there is no way of attaining knowledge that one is in one such world rather than another. This will certainly be the case if our evidence consists solely of sentences, given the elementary equivalence of isomorphic models: it immediately follows that any evidence will be consistent with the one possible world iff it is consistent with the other. So believing that there are worlds which are distinct yet isomorphic commits one to believing in facts that are in principle unobtainable—namely, facts about which world this is, out of the array of isomorphic options. Crude empiricism or positivism, which rejects unknowable facts outright, is rightly unpopular. But epistemic inaccessibility (especially the strong form of inaccessibility encountered here) is a prima facie reason to be suspicious of such facts, and to desire a way to do without them if possible. In short, if a metaphysical view admits of *insoluble ignorance*, then that is a defect of that view (albeit, presumably, a defeasible one).

One can also put the point in terms of propositions rather than possibilities. Associating propositions with sets of possible worlds, the issue is that we cannot know any proposition whose associated set of worlds is not closed under isomorphism: such as, for example, the proposition corresponding to it being some particular individual playing the role of Donald Trump, rather than any other. This way of putting things, however, raises the worry that (as with the semantic concern) richer accounts of our

[6] For defences of semantic sovereignty, see [Breckenridge and Magidor, 2010] or [Kearns and Magidor, 2012].

knowledge-seeking resources could resist this line of argument. In particular, it is tempting to appeal to certain forms of direct reference or externalism about content. Again, I defer discussion until the next section, where we will have the resources necessary to make out the appeal.

Before turning to that, we should note that there is a subtlety about nailing down the notion of ignorance at play here. Precisely because isomorphic models are elementarily equivalent, there is no sentence S such that we are ignorant of whether S is true or false: the ignorance is *inexpressible*.[7] However, I don't think that the inexpressibility is problematic. After all, we have a perfectly clear sense of the manner in which, independently of how much evidence (in the form of closed sentences) is gathered, there will remain multiple epistemic possibilities: that indicated above, of understanding epistemic possibilities as possible worlds satisfying the evidence-sentences. Of course, this analysis will be no good if we try to cash out possible worlds as (say) maximal consistent sets of sentences.[8] But since that's not how we're understanding possible worlds in general, there is no compulsion to adopt it as a means of understanding epistemic possibility. Moreover, we have an explanation of why this ignorance does not translate into sentential ignorance, namely, the semantic underdetermination that accompanies the epistemic underdetermination. So we have the resources to both explicate this as a species of ignorance, and to explain why it does not manifest as some more familiar species of ignorance. That is surely sufficient for us to account it as genuine ignorance.

2.2 *De re* representation

I now turn to an important objection to identifying isomorphic models. I observed in Chapter 1 that the characterisation of possible worlds as entities capable of rendering sentences true or false sufficed for the first two grades of modal involvement. However, it is not straightaway sufficient for the third grade. That is, suppose that we were to *extend* the modal language considered in the previous chapter, by amending the syntactic formation rule 5 to

5.′ If ϕ is a formula, then $\Box\phi$ is a formula.

To do so is to admit the third grade of modal involvement in a language: that of a sentential operator *permitted to take open formulae in its scope*, as in the formulae $\Box Px$

[7] [Dasgupta, 2015]; cf. [Maudlin, 1993].
[8] This yields [Chalmers, 2006]'s way of understanding epistemic possibility.

or $\exists x \Diamond Qx$. Clearly, in evaluating such a formula at a world $w \in \mathcal{W}$, it is insufficient to merely know what truth-values the worlds assign to closed formulae. Speaking roughly, we require not just information about the variation in what a *sentence* is like across possibilities, but information about what the variation in what an *individual* is like across possibilities: that is, we must plumb not just the *de dicto* modal facts, but the *de re* modal facts.

Do we need to care about formulae of this kind, though? Or could we follow Quine and reject *de re* modality as confused or incoherent?[9] The use of *de re* modal locutions is a widespread and robust component of ordinary speech: "someone at this party could have been famous", say. It's somewhat less clear that these locutions are needed for scientific purposes, but there is at least a prima facie case: "one of the planets in the solar system could have been larger" certainly seems like a scientifically appropriate thing to say (under the right evidential circumstances). Of course, one might think that such statements can be analysed away, so that verifying them only ever requires the demonstration that certain *de dicto* claims are true. But the best way to provide such an analysis is to first provide an account of the semantics needed for these claims, and then a demonstration that that semantics can be reduced to or constructed from something else.[10]

So suppose that we are interested in coherently assigning truth-values to the formulae of our extended language; and suppose, to begin with, that we are seeing if some arbitrary (Σ-based) Kripke picture will let us do so. Formally, what we need is some rule which tells us how to evaluate a formula such as $\Box Px$ at w, presumably relative to some assignment g of the variable x. The natural suggestion is to simply carry over the rule (e)′ from Chapter 1, but applied to formulae rather than merely to sentences:

(e)″ $|\Box\phi|^w_g = 1$ iff for every $v \in \mathcal{W}$ such that wRv, $|\phi|^v_g = 1$

Natural as this may be, however, it is incomplete as it stands. In evaluating $|Px|^v_g$, for instance, we run into the problem that if $g(x) \notin D_v$, then $P_v(g(x))$ is not defined. Prima facie, there are three responses to this problem available.

1. Extend the domain of P^v from D_v to $D := \bigcup_{w \in \mathcal{W}} D_w$; more generally, for any world $w \in \mathcal{W}$ and $\Pi \in \Sigma$, take Π^w to be a map $D^n \to \{0, 1\}$ (rather than $D^n_w \to \{0, 1\}$).

[9] Quine sounds some concerns about this grade in his [Quine, 1953b], but the *locus classicus* of his criticisms is [Quine, 1953a].

[10] As far as the dialectic of this thesis goes, there is another reason to be invested in *de re* modality. When we get to Chapter 5, we will need to exploit the apparatus of counterpart functions (developed below) in order to justify taking models related by an external symmetry as equivalent.

Heuristically, one could describe this option as requiring that objects have or lack properties, even at worlds where they do not exist. If this sounds unpalatable, one could seek to ameliorate it by stipulating that (for any $w \in \mathcal{W}$, $\Pi \in \Sigma$ and $a \in D$) if $a \notin D_w$ then $\Pi^w(a) = 0$—i.e., by stipulating that atomic predications of non-existent objects are always false. Alternatively, one could introduce a third truth-value of $1/2$, and stipulate that (for any $w \in \mathcal{W}$, $\Pi \in \Sigma$ and $a \in D$), $\Pi^w(a) = 1/2$. This would require, of course, committing to some particular treatment of such truth-values—e.g. the Lukasiewicz or Kleene truth-tables.

2. Require that for any $w, w' \in \mathcal{W}$, $D_w = D_{w'}$. Heuristically, one could describe this option as requiring that the same objects exist at every possible world (*necessitism*).[11] (That said, one could also think of the first option as a way of implementing necessitism, combined with opting for restricted quantifiers—say, quantifiers restricted to ranging over concrete things. So P^v takes a in its domain because a exists at v; if $a \notin D_v$, that just indicates that a is non-concrete at v.)

3. We could exploit the apparatus of accessibility relations.[12] First, we relativise accessibility to sets of individuals in worlds: given a set $A \subseteq D_w$ for some $w \in \mathcal{W}$, for any world $v \in \mathcal{W}$ say that wR_Av iff $A \subseteq D_v$. Second, given a variable-assignment $g : \mathrm{Var} \to D_w$, we define the set of ϕ-relevant individuals according to g as

$$g[\phi] := \{a \in D_w : \text{for some } x \text{ free in } \phi,\ g(x) = a\} \tag{2.4}$$

We now modify clause (e)" to

- $|\Box\phi|^w_g = 1$ iff for every $v \in \mathcal{W}$ such that wRv and $wR_{g[\phi]}v$, $|\phi|^v_g = 1$

This guarantees that, when evaluating a modalised open formula such as $\Box Px$ relative to assignment g, we need only consider the truth-value of Px at worlds in which $g(x)$ exists. (Note one curious feature: despite offering an intuitively contingentist metaphysics, a semantics along this lines will validate the Barcan Formula.) Although I have left the "original" accessibility relation R in alongside the individual-relative accessibility, under these circumstances it would be natural to trivialise R by supposing it equal to $\mathcal{W} \times \mathcal{W}$.

Although the third option is my preferred choice, for the purposes of this chapter I will remain neutral; where necessary, I will indicate how the choice of one or another

[11] [Williamson, 2013]

[12] This follows [Hazen, 1979] and [Russell, 2013] (note that in both those cases, the proposed solution was part of a counterpart-theoretic semantics).

of the above might bear on things.

For the claim that isomorphism suffices for equivalence, however, the semantics encoded by the clause (e)" is unacceptable (regardless of how it is completed): for such a semantics makes essential use of the distinctions between isomorphic models. In evaluating $\Box|Px|^w_g$, for instance, we look at what $g(x)$ is up to in all possible worlds: and a pair of isomorphic worlds v and v' may well nevertheless disagree on what role $g(x)$, in particular, is playing within their qualitatively indiscernible productions. Thus, if we are to preserve this (now semantically relevant) data, we may not collapse together isomorphic models. We have already seen that isomorphic models agree on the qualitative facts. The semantics considered here gives a positive account of what it is they disagree on: the facts about *which individual is playing which qualitative role*.

Hence, one advantage of *not* identifying isomorphic models is that it meshes with a fairly natural account of *de re* modality. Moreover, it might seem that this account can be used to undermine some of the concerns raised above, thereby undercutting the case for identifying isomorphic models in the first place. As regards the semantic concern, the thought would be that it is only a problem insofar as we are considering worlds comprised of individuals who do not actually exist.[13] For possible worlds whose only inhabitants are also denizens of the actual world (this thought goes), we can certainly pick out one of the various isomorphic possibilities: having given its qualitative description, we go on to assert that it is Alice who is playing this qualitative role, Bob who is playing another, etc.—where "Alice" and "Bob" are rigid designators whose reference we have fixed in actuality (whether by ostension or by description). Similarly, the epistemic problem could perhaps be blunted by a certain externalism about content.[14] By thinking '*this* man is playing the qualitative role of Donald Trump' whilst demonstrating Donald Trump, I succeed in believing a proposition to the effect that it is that very individual (and no other) instantiating the Trump role. If so, the thought continues, I surely have knowledge of that proposition, given that this method for coming to believe this proposition is reliable: had someone else been playing the Trump role, my singular thought would have concerned them instead.

These externalist manoeuvres do not succeed, however. The problem is that in both cases, they succeed in showing how *if* we had a determinate means of singling out a unique model corresponding to the actual world, then that would straightway extend to a means of singling out other unique models. But the antecedent clause is false. We

[13]Thanks to Shamik Dasgupta for raising this concern.
[14][Dasgupta, 2015] outlines this argument, as does [Maudlin, 1993].

can stipulate that we intend to be talking about a model whose qualitative structure matches that of actuality, but that does not serve to pick out one model in particular. And direct ostension of actual concrete objects does not help either. These objects are not literally components of any of the models, and they could have only a conventional association to the abstract elements of the models' domains. As such, there is no means of singling out one model as that corresponding to *the* actual world.

Similarly, there is no proposition that is uniquely eligible to be the content of my singular thought regarding Trump, since the concrete individual before me has no privileged association to the abstract domain-elements that differentiate the isomorphic candidate propositions (one being a set of models all containing one element, one being a set of models all containing another, etc.). Instead, the multitude of isomorphic propositions are all equally well-qualified to be the content of that thought. Thus, the only available line to take would be that when we think that *this* individual has the qualitative attributes of Trump, we indeterminately and simultaneously think all these propositions, not any one of them in particular. So we cannot use such externalism to believe—much less know—the specific proposition we were after.

2.3 Counterpart theory

Nevertheless, the other observation still stands: that the semantics above for *de re* modality cannot be appealed to if we interpret isomorphic models as equivalent. Let us consider the following argument that if isomorphic worlds are interpreted as equivalent, then *no* semantics of *de re* modality can be given at all. For, consider a world W_0 containing two qualitatively identical iron spheres.[15] Surely it's true, of one of the spheres, that it might have been different; that it might, for example, have had a subtly different chemical composition, or even that it might have been annihilated entirely.[16] Hence, there is a possible world W_1 according to which that sphere is different in the relevant respect. But surely the same is true of the other sphere. So there is also a possible world W_2 according to which *that* sphere is different in the relevant respect.

[15] The two-spheres world was introduced by [Black, 1952]. One might worry that such a world violates the (intra-world) principle of the identity of indiscernibles; however, if we allow that the PII may appeal to qualitative relations and not just qualitative properties, then the world can be made consistent with it (see [Saunders, 2003]). At any rate, there will be worlds of this sort that we will want to admit (i.e., worlds containing a high degree of symmetry)—so any version of the PII which rules them out will simply be presumed to not be a credible principle.

[16] The latter example is drawn directly from [Adams, 1979], whilst the former is—following [Pooley, 2013a]—adapted from it.

This can't be the *same* possible world as W_1, since it describes a different way for things to have gone: W_1 and W_2 disagree over which sphere is different, or over which sphere gets annihilated. Yet W_1 and W_2 are qualitatively identical. So—it seems—if we want to be able to represent how things could have gone for individuals, rather than just how things might for the world as a whole, then we are bound to reject anti-haecceitism; we need qualitatively identical yet distinct possible worlds, in order to represent the full plenitude of possibilities for distinct individuals.

The reply to this problem is well-known, however: one uses *counterpart theory*.[17] Take as given some INDIVIDUAL A in a WORLD W. The idea is that if we want to represent how things go for A according to some other WORLD W', the bare data comprised by the pair of WORLDS W and W' is insufficient: we need also to know which individual in W' has the job of representing what A would be up to in W'—what is known as a *counterpart* of A. This extra data is therefore coded up in a *counterpart relation* between the INDIVIDUALS in W and those in W'. Once we admit this extra data, then we can (at least in principle) see how to overcome the problem of the two spheres. Let us suppose that the entertained alteration is some chemical change, say from iron to an iron alloy containing a trace amount of lead. Both W_1 and W_2, therefore, contain one sphere of pure iron and one sphere of alloy; in W_1 the alloy sphere is a counterpart of one of the iron spheres in W_0, whilst in W_2 the alloy sphere is a counterpart of the other iron sphere in W_0. Now, if the relation "being a counterpart of" is an equivalence relation, then we must indeed admit W_1 and W_2 as distinct possibilities, on pain of claiming that we started with one iron sphere rather than two. However, if we allow that it may have a more flexible logical character, then we need not keep them distinct: we can allow that the alloy sphere (in W_1 *aka* W_2) is counterpart to *both* of our original iron spheres, and that it therefore does double duty for both our *de re* claims at once.

Let us seek to make this precise. To take a specific example, let us suppose that we are evaluating whether $\Box Px$ holds of some INDIVIDUAL A in a WORLD W: that is, whether $|\Box Px|^W_G = 1$, where G is a relative variable-assignment such that $G(x) = A$. Suppose that V is a WORLD accessible from W. According to the counterpart theorist, there is some INDIVIDUAL B of V who is a counterpart of A, which gives us a proxy means of assessing whether Px holds of A according to V: we look at whether $P^V(B) = 1$ (i.e. $P^v(B_v) = 1$, for any $v \in V$).

However, B need not be the only counterpart of A. This is widely discussed (indeed, it is generally taken to be one of the central innovations of counterpart theory), but is

[17] [Lewis, 1968], [Lewis, 1986], [Hazen, 1979]

actually rather ambiguous: there are (at least) three senses in which this could be the case.

First, it could be that there are *different counterpart relations*, which disagree over which INDIVIDUAL in V is the counterpart of A. In particular, one might think that counterparthood is tied to (qualitative) similarity, in which case one expects different counterpart relations as the context makes salient different varieties of similarity. For example, suppose that A is a bronze picture frame, B a wooden picture frame, and C a bronze statue. Then one context might emphasise similarity of function, so that A's counterpart in V is best taken to be B, whilst another context might emphasise similarity of material, so that A's counterpart in V is best taken to be C.

Second, it might be that even having fixed the context, and so the relevant counterpart relation, there is more than one individual in V who bears the relation of counterparthood to A. Our two-spheres world offers an illustration of this: as we remarked earlier, the alloy sphere (in W_1, i.e. W_2) is a counterpart of both iron spheres in W_0 *in the specific sense that* it is true of one iron sphere that the alloy sphere is its counterpart, and also true of the other iron sphere that the alloy sphere is *its* counterpart.

But even this is not quite the same as the third and most radical sense in which an object might have multiple counterparts. That sense would be that of A having B and C as counterparts *simultaneously*, so that assessing what is true of A according to V proceeds by consulting both B and C and seeing if the claim in question is true of either. Thus, for example, this notion of simultaneous counterparts holds that if B satisfies P and C does not, then both Px and $\neg Px$ hold of A according to V. Or, in a more general principle: a (one-place open) formula $\phi(x)$ holds of A according to V just in case V contains a counterpart of A which satisfies $\phi(x)$.

I claim that we should allow INDIVIDUALS to have multiple counterparts in the first and second senses, but not the third. In general, what A is up to at V is underspecified by the data contained in V (and a specification of the context); one also needs to fix on some *particular* counterpart of A as being "the" representative of A for the purposes of this assessment. So the context might select out those INDIVIDUALS in V which are candidates to be counterparts of A; but any actual assessment of what A is like according to V involves the (arbitrary) selection of some particular candidate. Note that this dependence on a choice of representative can be made to come out in the wash when assessing whether claims such as $\Box Px$ hold of A (at W): we simply stipulate, as seems natural, that $\Box Px$ holds of A if at every V accessible from W, and on every choice of A-representative B from V, Px holds of B according to V.

There is one more issue we need to navigate, before we are in a position to specify how clause (e)" should be amended. Suppose that we want to assess whether a *relational* claim Rxy holds, according to V, of some pair of INDIVIDUALS (A, A') from W. In order to assess this claim, we need some notion of the counterpart of a pair. Now, one option would be to say that a pair (B, B') is a counterpart of (A, A') just in case B is a counterpart of A and B' is a counterpart of B'. But this seems like a bad idea. Even if we suppose the context fixed, the similarity of the pair (B, B') to (A, A') need not supervene upon the similarity of B to A and B' to A'. Again, take the two spheres as an example. As we've seen, the alloy sphere and iron sphere in W_1 are each candidate counterparts of each of the iron spheres in W_0. But it shouldn't follow from this that the pair consisting of the alloy sphere and itself is a potential counterpart of the pair of iron spheres, nor the converse. For then we would get the claim that (according to this choice of counterparts) W_1 represents W_0's pair of iron spheres as being identical to one another; and conversely, that W_0 represents W_1's alloy sphere as being distinct from itself. But neither of these claims is plausible: it is not possible that a pair of distinct individuals could have been identical, nor that a pair of identical individuals could have been distinct.[18]

We implement these ideas using the notion of a *counterpart-function* from one set of INDIVIDUALS to another.[19] Given sets $\mathbf{A} \subseteq D_W$ and $\mathbf{B} \subseteq D_V$, a counterpart-function is a bijection $C : \mathbf{A} \to \mathbf{B}$. Given a pair of worlds W and V, let $\mathcal{C}(W, V)$ be the set of all counterpart-functions from some subset of D_W to some subset of D_V. I will assume that $\mathcal{C}(W, W)$ always contains Id_{D_W}. The idea is that if we are evaluating what A is up to at V, then we must look at its image $\kappa(A)$ under each counterpart-function $\kappa \in \mathcal{C}(W, V)$ in turn. And if we want to know what the pair (A, B) is up to, we consider $(\kappa(A), \kappa(B))$ for each $\kappa \in \mathcal{C}(W, V)$; and so on for longer sequences of individuals.

So much for cases of multiple counterparts. However, what if a WORLD contains *no* counterparts of A? Again, this is somewhat similar to the problem discussed in section 2.2 above. The analogue of solution (1) is just to forbid this from ever happening: i.e. insisting that for any pair of WORLDS V and W, there is at least one counterpart-function $C : D_V \to D_W$. The analogue of solution (2) is to do the same, but suppose that the quantifiers may range, at a WORLD W, over some subset of D_W. Finally, the analogue of solution (3) is to restrict accessibility only to WORLDS containing appropriate

[18] I don't have space here to defend this pair of theses (the theses of necessary distinctness and identity, as they're known). However, they are widely accepted: see [Kripke, 1971] for a defence (although [Gibbard, 1975] resists some of these arguments).

[19] [Hazen, 1979]

counterparts. That is, given a set $\mathbf{K} \subseteq D_W$ and any WORLD $V \in {}^{\mathcal{W}}\!/_{\cong}$, say that $W\mathbf{R}_{\mathbf{K}}V$ iff there is some counterpart-function $C \in \mathcal{C}(W, V)$ such that $\mathbf{K} \subseteq \mathrm{dom}(C)$; the idea is to evaluate Px (relative to G) only over WORLDS which are accessible to the ϕ-relevant INDIVIDUALS according to G, defined as

$$G[\phi] := \{A \in D_W : \text{for some } x \text{ free in } \phi,\ G(x) = A\} \tag{2.5}$$

All three solutions, however, are consistent with a counterpart-theoretic version of clause (e)", namely

(E) $|\Box\phi|^W_G = 1$ iff for every $V \in {}^{\mathcal{W}}\!/_{\cong}$ and every $C \in \mathcal{C}(W, V)$, if $G[\phi] \subseteq \mathrm{dom}(C)$ then $|\phi|^V_{C \circ G} = 1$

If one opts for (the analogues of) solutions (1) or (2), then the antecedent clause "if $G[\phi] \subseteq \mathrm{dom}(C)$" is guaranteed to be fulfilled.

2.4 Three dimensions of modal disagreement

We now have two semantics adequate to the third grade of modal involvement: that is, two accounts of how *de re* representation is carried out. The disagreement between these semantics is often referred to in terms of *haecceitism* and *anti-haecceitism*—labels which I have avoided so far, since they have meant rather different things to different authors. The labels originally trace back to the following passage of Kaplan (edited so that his notation matches mine):[20]

> When we construct a model of something, we must distinguish those features of the model which represent features of that which we model, from those features which are intrinsic to the model and play no representational role. The latter are *artifacts of the model*. [...] given any distinct elements w and w' of $\mathcal{W}$, some definite relation, either of overlap or disjointness, will hold between D_w and $D_{w'}$. [...] Thus, the overlaps (or disjointness) between such pairs as D_w and $D_{w'}$ is a definite feature of our model. Is it an *artifact of the model* or a *feature of the metaphysical reality* being modeled?
>
> [...] there seems to be some disagreement as to whether we can meaningfully ask whether a possible individual that exists in one possible world also

[20] Although [Kaplan, 1975, p. 725] credits the "epithet" of haecceitism to Robert Adams.

> exists in another without taking into account the attributes and behavior of the individuals that exist in the other world. The doctrine that holds that it does make sense to ask—without reference to common attributes and behavior—whether *this* is the same individual in another possible world, that individuals can be extended in logical space (i.e. through possible worlds) in much the same way we commonly regard them as being extended in physical space and time, and that a common "thisness" may underlie extreme dissimilarity or distinct thisnesses may underlie great resemblance, I call *Haecceitism*. [...]
>
> The opposite view, *Anti-Haecceitism*, holds that for entities of distinct possible worlds there is no notion of trans-world being. They may, of course, be linked by a common concept and distinguished by another concept [...] but there are, in general, many concepts linking any such pair and many distinguishing them. Each, in his own setting, may be clothed in attributes which cause them to *resemble* one another closely. But there is no metaphysical reality of sameness or difference which underlies the clothes. [...] Although the Anti-Haecceitist may seem to assert that no possible individual exists in more than one possible world, that view is properly reserved for the Haecceitist who holds to an unusually rigid brand of metaphysical determinism.[21]

Confusion over the label begins here, for the above passage actually bundles together three distinct doctrines. One is the issue of whether the overlap or disjointness of world-domains should be accorded semantic or representational significance (as Kaplan puts it, whether the identity of individuals across models is a mere artefact). This has been my main concern so far; let us describe proponents of the two available positions as (respectively) *identitarians* and *counterpart theorists*. The second issue concerns how closely transworld identity is tied to qualitative character. Do the relations of transworld identity entirely float free of the qualitative architecture of the worlds whose domains they put into correspondence, so that there is no systematic relationship at all? Or, at the other end of the spectrum, do those relations just supervene on the qualitative facts, so that the latter entirely fix the former? Or is it some combination, where the qualitative facts constrain the transworld identity relations, without fully determining them? Finally, there is the question of whether there is some uniquely privileged

[21] [Kaplan, 1975, pp. 722–723]

relationship of transworld identity, or if there are many such relations, with our use of one rather than another varying (presumably) with contextually driven salience, particular interests and attitudes, etc. Let us say that advocates of the former view are *monists*, and that advocates of the latter view are *pluralists*.

With these distinctions in hand, we can see that the position that Kaplan identifies as Haecceitism is (in truth) a combination of identitarianism with monism and the (full) independence of the singular from the qualitative; and that the position he identifies as Anti-Haecceitism combines counterpart theory, pluralism, and the denial of such independence. However, these three distinctions are mostly independent of one another. Mostly independent, not fully independent: identitarianism entails monism, given that there can be only one identity relation between the elements of two sets (and conversely, pluralism entails counterpart theory). And pluralists need not hold determinately to any one view about how tightly the transworld identity facts must track the qualitative facts: different counterpart relations may be correlated to qualitative similarity to different degrees. Beyond these interactions, however, any combination is possible. Counterpart theory may be combined with monism, by holding that there is some particular counterpart relation that enjoys a privileged status; identitarians may hold that the transworld identities (as indicated by domain-overlap) show a total or partial correlation with the qualitative facts; and counterpart theorists may allow counterpart relations that exhibit no systematic relationship to qualitative similarity at all.[22]

We can bracket the issue of pluralism versus monism, as we have done in the above, by supposing that we have fixed the context sufficiently that only one counterpart relation (if there is more than one) is salient. But that still leaves us with two theses associated to the term "haecceitism": either the thesis that transworld identity is encoded by identities rather than counterpart relations, or the thesis that transworld identities may float (somewhat) free of qualitative constraints. I now announce, therefore, that I will mean the *conjunction* of these theses whenever I speak of "haecceitism". This has the consequence that the haecceitist believes in distinct yet isomorphic worlds: because transworld identity does not fully supervene on qualitative character, there are worlds with the same qualitative character (i.e. which are isomorphic) but which do not codify the same haecceitistic facts. I am allowing both moderate essentialists and radical recombinatorialists to count as haecceitists: the latter think that the transworld identities float entirely free of the qualitative facts, whilst the former just think that

[22] Lewis does consider this possibility, but rejects it on the grounds that the alleged non-qualitative counterpart relation is unintelligible ([Lewis, 1986, pp. 229–230]. But all we're doing here is canvassing logical possibilities, not assessing their plausibility.

they are not fully determined by them.

Since haecceitism is a conjunctive thesis, there are various ways to oppose it. One could be a counterpart-theorist who thinks that what available counterpart relations are around is not something to be fully determined by qualitative facts. (This is the view that Lewis decried as incoherent.) One could be an identitarian who believes that the qualitative structure of two worlds guarantees the transworld facts, i.e. (for you) the facts about which individuals are in them: this is an extreme form of essentialism, Kaplan's "unusually rigid brand of metaphysical determinism". Finally, one could deny both halves of the haecceitistic creed, and be a counterpart theorist who also takes the qualitative facts to determine what counterpart relations are available. Note that although Lewis held this view, he was also quite happy with the existence of distinct yet isomorphic worlds. So the denial of such worlds is an additional commitment. I'll refer to this package of views—counterpart theory, the supervenience of counterpart relations upon qualitative facts, and the denial of distinct yet isomorphic worlds—as *qualitativist anti-haecceitism*, or just *qualitativism* for short. Defending it will be my main concern for the rest of this chapter. (That said, the supervenience of the counterpart relations upon the qualitative facts won't play much of a role in this chapter, so what I say here could extend to a view where the counterpart relations aren't tied so tightly to qualitative character. It's only when we start looking at specific theories—especially in Chapter 5—that I'll start making use of this idea.)

Finally, a brief remark about how these relate to a rather different distinction: that between *individualism* (the view that the fundamental facts include individual facts) and *generalism* (the view that the fundamental facts do not include any individual facts).[23] The individual facts are described as those facts which depend "on how things stand with a particular individual (or individuals)";[24] the general facts are all the non-individual facts. I am a little skeptical of the notions of ground being appealed to in this debate, but I will put that to one side for the moment. The question that concerns us is how the individualist/generalist debate interacts with the disputes just canvassed.

First, it seems clear that generalism entails that the transworld identity facts supervene upon the qualitative facts: facts regarding how individuals are to be identified across possibilities are paradigmatically individual, and it is generally reckoned that grounding entails supervenience (so that if the A-facts ground the B-facts, the B-facts

[23][Dasgupta, 2009], [Paul, 2012], [Dasgupta, 2014a], [Dasgupta, 2016], [Russell, 2015], [Russell, 2016].
[24][Dasgupta, 2016, p. 1].

must supervene upon the A-facts). Second, it seems very plausible that generalists should be counterpart theorists. Exactly how generalism should be carried out (and whether it can be carried out) remains contentious, but the idea typically seems to be that individuals are either absent (with talk about individuals a mere *façon de parler* for talk describing the general facts), or are to be somehow constructed from generalistically acceptable constituents.[25] If the former, then it surely cannot be the case that individuals have modally robust identities (given that they don't exist, speaking literally); if the latter, then it is hard to see how entities constructed from general, qualitative structure could possess intrinsic, modally robust identities.

Thus, we have a good case that generalism entails qualitativism (in the sense defined above). What about the converse? That is, does qualitativism commit one to generalism? I don't think this question has a definitive answer, given the scope for contesting what an individual fact should be. Certainly, being an anti-haecceitist seems consistent with accepting that amongst the fundamental facts are facts about the existence of individuals (i.e., facts to the effect that so-many individuals exist, and have such-and-such properties). What will not be consistent is a combination of anti-haecceitism with an acceptance of facts regarding the existence of *some individual in particular* (in contrast to the existence of some other individual). For the only way I can see to cash out that idea would be to admit intrinsic, internal identities of just the kind denied by the anti-haecceitist (*qua* counterpart theorist). That said, note that although such an individualism would entail identitarianism, it doesn't seem to entail the non-supervenience of the singular on the qualitative. This corresponds to the fact that grounding theses are generally taken to be *strictly* stronger than supervenience theses: that allows one to deny that the individual facts are grounded in the general facts, yet admit that the singular facts supervene upon the qualitative facts.

2.5 Haecceitism in disguise?

I now turn to a slightly different concern: that qualitativism could be effectively equivalent to moderate haecceitism. Here's the concern. We can imagine a circumstance under which one has a pair of haecceitist worlds w, v, and a counterpart-function $\kappa : [w] \to [v]$, such that for any $A \in D_W$, $\kappa(A) = B$ iff $B_v = A_w$. In other worlds, it certainly seems plausible that at least sometimes, the manner in which a counterpart-function could relate the denizens a pair of WORLDS could be precisely analogous to the way in which

[25] In Chapter 5, I'll briefly review a way in which this might be done in the context of spacetime theories.

the identity relation relates the denizens of some haecceitist worlds. The concern is that if this kind of thing were sufficiently widespread, then perhaps some kind of equivalence could be demonstrated between the two formalisms, showing the one to just be a notational variant on the other.

Well, to begin with, the two accounts are formally distinct: there are anti-haecceitist models which correspond to no haecceitist model. The reason for this is that the counterpart relations need not behave like identity. (This is despite the fact that I forced them to be injective functions, thereby making them already somewhat identity-like; a counterpart theorist who rejected the necessity of identity or the necessity of distinctness would have an even stronger disagreement with the haecceitist than I.) In particular, the counterpart relations are not, as things stand, required to be symmetric: $\kappa \in \mathcal{C}(V, W)$ does not entail that $\kappa^{-1} \in \mathcal{C}(W, V)$. Nor are they required to be transitive: $\kappa \in \mathcal{C}(U, V)$ and $\kappa' \in \mathcal{C}(V, W)$ do not jointly entail that $\kappa' \circ \kappa \in \mathcal{C}(U, W)$, even if $\mathrm{cod}(\kappa) = \mathrm{dom}(\kappa')$.

There is a more difficult question regarding whether one can demonstrate an equivalence, provided that sufficient conditions are imposed upon the formal character of the counterpart relations. Presumably, the relevant result here would be the exhibition of some pair of systematic translations between haecceitist and (suitably constrained) counterpart-theoretic models, such that a model (of either kind) verifies a modal formula iff its translation does. Trying to prove such a result is not straightforward, and depends on the details of the *de re* semantics that have been proposed: that is, generating a proof will depend on whether solution (1), (2) or (3) has been adopted.[26]

Even if such a proof could be offered, however, I maintain that we can distinguish the haecceitist and the anti-haecceitist—and that the anti-haecceitist comes out better off. First, there is still an important sense in which the counterpart theorist is committed to less structure than the haecceitist. Consider a pair of (let's say, non-isomorphic) haecceitistic worlds. There is a privileged (partial) bijection between their domains: the identity. The corresponding piece of structure on the counterpart theorist's view is that of a pair of WORLDS, together with some particular counterpart function between them. But for the haecceitist, the bijection is an *internal* relation: it arises because of the existence of certain *intrinsic* facts about each world (namely, which individual plays each qualitative role). For the counterpart theorist, by contrast, the counterpart function is an external relation, merely showing how the individuals of one WORLD could be used to represent those of the other WORLD—without there being some underlying

[26] That said, see [Bacon, 2014] for a result along these lines.

intrinsic data which generates the counterpart function.

Second, the counterpart-theorist's picture offers a cleaner account of the distinction between *de dicto* and *de re* possibility. For the counterpart theorist has two kinds of data in their models. There are the WORLDS, each with their own qualitative character; if all you want is *de dicto* modalities, then this data is all you need. It is only if one also desires *de re* modal claims that there is a need for the counterpart functions, overlaid on top of the WORLDS like guiding threads. The haecceitist, by contrast, has just one kind of entity—the worlds, enriched with intrinsic haecceitistic facts—playing double duty, simultaneously encoding how things could be for specific individuals, and how they could be for entire worlds. Insofar as we want to distinguish our two kinds of possibility, it is helpful to have a formal apparatus which lets us factorise them in the neat manner of the counterpart theorist.

Finally, suppose that one buys the idea that the *de re* facts ought to supervene upon the qualitative facts. I contend that the haecceitist and anti-haecceitist have available different accounts of how to understand this fact: specifically, that the anti-haecceitist is explanatorily better off than the haecceitist. The anti-haecceitist can explain this by contending that the supervenience indicates a *reduction* thesis. All that the counterpart relation really was, it turns out, was some kind of qualitative relation. So the fact that the transworld identity facts track the qualitative facts isn't strange or surprising, any more than it's strange or surprising that the thermodynamic facts track the statistical-mechanical facts. By contrast, it does not seem that the anti-haecceitist can help themselves to this kind of reductionist story. After all, they are supposed to believe that there are intrinsic, internal facts associated to each world about what individuals are present in that world. It is very hard to see how the fact that some individual is the very individual that it is could be reduced to facts about its qualitative character. The only way I can see to cash this idea out would be to advocate generalism; but as discussed in §2.4 above, it seems plausible that generalism entails anti-haecceitism. So such a move possesses no dialectical force.

In fact, I think that the anti-haecceitist can, under appropriate circumstances, *use the haecceitist formalism* as a means of representing some particular counterpart relation. That is, suppose that we wish to discuss some specific counterpart-relation $\kappa : W \to V$. We know that κ must be a bijection between certain subsets of D_W and D_V; we also know that we have the freedom to choose any Kripke-worlds from within W and V as representatives. So we may as well choose a pair of Kripke-worlds w and v such that $\kappa(A) = B$ iff $B_v = A_w$. In effect, we use the identity relation on the base sets

as a representation of the counterpart-function κ, since it will have the same formal character.

2.6 Determinism

In this final section, I turn to one final topic: the relationship between modality and determinism. Here, I have less of an axe to grind. The main point is to distinguish two varieties of determinism, and explain how they are cashed out by the haecceitist and anti-haecceitist respectively; we will then return to this distinction in Chapter 5. In thinking about determinism, we are supposing that some characterisation of time has been built into our formalism. I won't worry about exactly how this has been done (e.g. whether we have captured it by extensional means, or have extended the language with tense operators). We will just suppose that it has been added somehow, in such a way that we can speak of what models are like at certain times. So a model $\mathcal{M}$ will be supposed to come with a stock $T_{\mathcal{M}}$ of times; given $t \in T_{\mathcal{M}}$, let $\mathcal{M}|_t$ denote what $\mathcal{M}$ is like at t.[27]

Now consider the following definition of determinism, for an interpreted theory $\mathbb{T}$.

Definition 1. $\mathbb{T}$ is deterministic if: for any models $\mathcal{M}, \mathcal{N}$ of $\mathbb{T}$, if there is some $t \in T_{\mathcal{M}}$ and $t' \in T_{\mathcal{N}}$ such that $\mathcal{M}|_t$ is equivalent to $\mathcal{N}|_{t'}$, then $\mathcal{M}$ is equivalent to $\mathcal{N}$.

In other words: if two worlds agree at some time, then they agree at all times.

Since the haecceitist and the anti-haecceitist disagree over the criteria for equivalence of models, they will take the above definition to translate into different criteria for determinism. For the haecceitist, equivalence is just identity, so the definition becomes

Definition 2. $\mathbb{T}$ is deterministic if: for any models $\mathcal{M}, \mathcal{N}$ of $\mathbb{T}$, if there is some $t \in T_{\mathcal{M}}$ and $t' \in T_{\mathcal{N}}$ such that $\mathcal{M}|_t = \mathcal{N}|_{t'}$, then $\mathcal{M} = \mathcal{N}$.

For the anti-haecceitist, on the other hand, equivalence is a matter of isomorphism, and so they obtain

Definition 3. $\mathbb{T}$ is deterministic if: for any models $\mathcal{M}, \mathcal{N}$ of $\mathbb{T}$, if there is some $t \in T_{\mathcal{M}}$ and $t' \in T_{\mathcal{N}}$ such that $\mathcal{M}|_t \cong \mathcal{N}|_{t'}$, then $\mathcal{M} \cong \mathcal{N}$.

[27] There's a bit of a subtlety here: it may be better to think of t as an arbitrarily short interval of time, rather than an instant of time. I gloss over this issue, in the interests of simplicity.

As one would expect, these two definitions do not coincide. For example, consider the following theory:[28] it governs particles which manifest a single determinable property, the determinates of which may be labelled continuously by the real numbers between 0 and 1 (inclusive). Call this property the particle's "charge". For any given particle, its charge uniformly decreases with time, until it hits 0, at which point the particle decays into two other particles with (initial) charge 1. This theory is deterministic according to Definition 3: given any distribution of particles with charges, we can predict exactly when each particle will decay, what the outcomes of that decay event will be, and what further events it will be followed by. However, it is indeterministic according to Definition 2. For example, consider the following pair of models $\mathcal{M}$ and $\mathcal{M}'$ of this theory. At time t, both models feature the same particle a, and both agree that a at t has charge 1. At time $t+1$, both models have a decaying into a pair of particles. However, the pair of objects into which a decays in $\mathcal{M}$ are distinct from the pair of objects into which a decays in $\mathcal{M}'$. So despite the fact that they agree at t, $\mathcal{M}$ and $\mathcal{M}'$ are distinct models—and so, according to the haecceitist, represent distinct possible worlds.

Now, it seems fair to remark that determinism as studied by science is better captured by Definition 3, rather than Definition 2: the kind of indeterminism exhibited by this theory is not apt to strike us as scientifically interesting. However, it's not clear that this is a dreadful burden for the haecceitist. All they have to do is allow that scientists happen to be interested in a somewhat derivative notion of determinism to that captured by Definition 1: what they might call *qualitative determinism*, the issue of whether the qualitative facts at one time suffice to determine the qualitative facts at all times. Sure, this isn't the most natural definition of determinism, given how things stand metaphysically (according to them). But it's perfectly definable: indeed, the haecceitist will presumably hold that it is captured by Definition 3. This position is, perhaps, a little non-naturalist (since it holds that the most metaphysically natural definition of determinism is distinct from that of interest to science), but it certainly doesn't seem incoherent.

Moreover, the haecceitist might claim that there are at least some cases where we *do* want to allow a certain sense of indeterminism, which cannot be captured by Definition 3. For example, suppose that we have a world that consists of nothing but a single tower on a spherical, homogeneous, and otherwise empty planet.[29] At some time

[28] This is broadly similar to an example discussed by (*inter alia*) [Belot, 1995], [Rynasiewicz, 1994, p. 419], [Melia, 1999, p. 647], and [Leeds, 1995, pp. 428–429].

[29] The example comes originally from [Wilson, 1993]; it is discussed by [Belot, 1995], [Brighouse, 1997], [Melia, 1999], [Brighouse, 2008], and [Arntzenius, 2012, pp. 179–181]. I follow the version given by

(again, determined by the laws of nature), the tower topples to the ground; however, the laws say nothing about the direction in which it is to fall. As a consequence, the world is claimed to be indeterministic:

> Now, are there not many ways in which the tower could have fallen? Surely, given the symmetry of the initial conditions, the tower could have toppled in a different direction and come to rest upon some other part of the planet. It didn't happen—but it might have.[30]

One can still deny that this example is indeterministic, or at least that it is indeterministic in any physically relevant way—[31]perhaps with some extra philosophical manoeuvring to explain away why they seem intuitive. Both [Brighouse, 1997] and [Rickles, 2008, p. 97] suggest that the reference to individual entities in the tower example is illegitimate: unless we imagine ourselves in the scenario, we cannot determinately refer to the particular parts of the ground to say that the tower might or might not fall there. But even if we grant that no linguistic act can determinately refer to only one bit of ground rather than another, that doesn't show that we cannot quantify over all of them to say that for each such bit of ground, it is indeterminate what will happen to it (whichever bit "it" should happen to be).

So, as with the third grade of modality, it would be nice for the anti-haecceitist to have *some* way of recognising the force of such examples. The solution is to recognise that they, too, can distinguish two ways in which a world could be deterministic. In the tower example, the point is that different ways of (as it were) placing the individual into the model, consistent with the state at a time, lead to different outcomes for that individual. In terms of models, this means that a theory is indeterministic if there is a way of putting both models into isomorphic correspondence at some time which does not translate into isomorphic correspondence at all times. This motivates the following definition of determinism:

Definition 4. $\mathbb{T}$ is deterministic if: for any models $\mathcal{M}, \mathcal{N}$ of $\mathbb{T}$, if there is some $t \in T_{\mathcal{M}}$ and $t' \in T_{\mathcal{N}}$ such that $\mathcal{M}|_t \cong \mathcal{N}|_{t'}$, then $\mathcal{M} \cong \mathcal{N}$; *and* for any isomorphism $f : \mathcal{M}|_t \to \mathcal{N}|_{t'}$, there is an isomorphism $g : \mathcal{M} \to \mathcal{N}$ such that $g|_t = f$ (i.e., such that g is an extension of f).[32]

Melia.

[30][Melia, 1999, p. 649]

[31]This is, more or less, the response advocated by [Brighouse, 1997].

[32]This definition corresponds to [Belot, 1995]'s "definition 2", and to [Melia, 1999]'s "second resolution". On the basis of Melia's paper, one might worry that I should have opted for his, supposedly

This definition does characterise the tower-world as indeterministic (although not the particle-decay world; more on that in a moment): just taking a single model of the world for convenience, note that only the identity map is a global automorphism, but any rotation around the tower as an axis is an automorphism of the pre-collapse segment. So both the haecceitist and the anti-haecceitist have the resources to distinguish two varieties of determinism. This suggests the distinction is of independent interest, and worth thinking about a little more. Melia's way of expressing the distinction is as follows:

> On Lewis' definition, whether or not there is more than one way in which a world can evolve is a global matter: two duplicate futures count as one physically possible way a world could evolve. But this conflicts with the way we think a deterministic world should evolve at a *local* level. If a theory is deterministic then it should be the case that any two parts of two worlds satisfying this theory that share the same qualitative properties and relations and that have duplicate histories will go on to share qualitative properties and relations in the future.[33]

It seems to me that this is the right way of cashing out the distinction, though unlike Melia, I'm not sure that one of these kinds of determinism is "really" determinism; I'd rather just recognise that there are distinct kinds of determinism, and suggest we distinguish between them. And rather than the "local" and "global" phrasing, I would suggest that we borrow pre-existing terminology: it seems to me that what we are seeing is a distinction between (what we will call) *determinism de dicto* and *determinism de re*.[34]

Determinism *de dicto* is a question of whether there is more than one way the future could go for a *world*, given how it is at a particular time; determinism *de re* is whether there is more than one way the future could go for an *individual*, given how it (and its environs) are at a time. I think this helps to stress that in fact, with a little hindsight, it is *utterly unsurprising* that there should turn out to be two concepts of determinism. Determinism is a matter of whether there is one possibility or more consistent with

inequivalent, first resolution: this accounts a theory indeterministic if there is some isomorphism f that maps the objects of S to those of S' but no global isomorphism which agrees with f on the mapping of objects. In fact, however, the two resolutions are equivalent after all: see [Pooley, 2002, pp. 108–109].

[33] [Melia, 1999, p. 652]

[34] cf. the idea in [Skow, 2005, §3.2] that we might distinguish different kinds of determinism on the basis of whether Laplace's demon is being given (and being asked to produce) only qualitative, or qualitative and non-qualitative, information.

things being a certain way at a certain time; we have two species of possibility, *de dicto* and *de re*; so as a consequence, there are two species of determinism. Of course, if one is sceptical of *de re* modality in general, then that scepticism will naturally be extended to its associated determinism. On the other hand, if one is happy to accept that one can meaningfully talk not only of ways for a world to be but also ways for an individual to be, then one should be happy to extend this treatment to ways for an individual to evolve.

Moreover, understanding the two kinds of determinism as linked to our two kinds of modality helps us to see why the haecceitist and the anti-haecceitist understand the two kinds in somewhat different ways. For the haecceitist, determinism framed in terms of worlds captures *de re* determinism, since the worlds already encode *de re* data; defining *de dicto* determinism requires abstracting away from certain features of possible worlds. For the anti-haecceitist, worlds (considered by themselves) are apt only to capture *de dicto* data, and so it is *de dicto* determinism that gets captured by a criterion on worlds. The *de re* data necessary to capture the associated notion of determinism comes from attending to the persistence and identity conditions of individuals within models.

However, there is still a slight asymmetry between the two positions. Definition 4 is available to the haecceitist (again, as a matter of abstracting away from some of the full portion of facts). But Definition 2 is not available to the anti-haecceitist, or at least not in any natural way: after all, anti-haecceitists don't think that identity of models (or parts thereof) is representing anything of significance. As such, there is *no* sense in which the anti-haecceitist will admit the particle-decay world as indeterministic. There is no uncertainty about how the global qualitative facts will pan out; and at no time is there any individual whose own particular future is uncertain. So if one did find the particle-decay world compellingly indeterministic, then that would be a strike against the anti-haecceitist. But if not, then this becomes a feature not a bug: the anti-haecceitist can then *explain* why this example strikes us as less compelling than the tower world.

3 Anti-quidditism

I'm a charged body,
But I don't know which kind

Totemo, *Host*

In the previous chapter, we were concerned with how the qualitative character of worlds constrained the kinds of variation they could exhibit. In this chapter, we consider how the *nomological* character of worlds constrains their variation: specifically, the extent to which worlds can differ qualitatively whilst manifesting the same kind of nomological structure.

In the metaphysics of properties, there are a contrasting pair of answers given to this question. One answer, generally known as quidditism, says something in the neighbourhood of the following theses: what makes a property the very property that it is, is an internal, intrinsic feature of that property (its "quiddity"); properties can be identified across possibilities, without reference to the roles that they happen to be playing in the laws governing those worlds (their *nomological roles*); properties play these roles contingently; it is no part of the nature of a property that it plays the nomological role that it does; that for any given nomological role that is realised by a particular property in a particular possible world, that role could have been realised (in some other possible world) by a different property.

The other answer (or perhaps answers), known variously as nomic structuralism, or causal essentialism, or anti-quidditism, says something in the neighbourhood of the following theses: what it is for a property to be the very property that it is, is the nomological or causal role of that property; properties can only be identified across possible worlds with reference to the causal roles that they play; properties play their causal roles necessarily or essentially; it is part of the nature of a property that it play the causal role that it does; if a nomological role is realised by some property in one possible world, it is impossible for that role to be realised by any other property.

This chapter is a defence of (a version of) the second view: I will refer to this view

as "structuralist anti-quidditism" (for short, SAQ). In the next section, I explain what structuralist anti-quidditism is. In sections 3.2 and 3.3, I consider the reasons to be a structuralist anti-quidditist. In section 3.4, I remark on the implications of anti-quidditism when we consider a special case: that of *symmetries*. Finally, section 3.5 looks at how to incorporate those implications into the formal structure of the theory.

3.1 Structuralist anti-quidditism

In the previous chapter, I defended structuralist anti-haecceitism: the view that isomorphic Kripke-worlds should be taken to represent the same possible world. It is widely recognised that the debate over quidditism and anti-quidditism is helpfully thought of as the "higher-order" analogue of the debate over haecceitism and anti-haecceitism. Just as haecceitism countenances distinct possible worlds that agree on the distribution of qualitative structure, and differ only over which individuals are the ones instantiating that structure, so quidditism countenances distinct possible worlds that agree on the distribution of *causal* or *nomological* structure and differ only over which properties are the ones instantiating that structure.

So, making out the debate will require us to get clear on the notion of a world's nomological structure. Naturally, this risks opening up a whole can of worms as to the nature of laws. I don't intend to get involved in that debate, so I will here just provide an account of laws that will serve our purposes nicely, and not worry about its virtues or vices relative to other accounts. In effect, we will take laws to stand to theories[1] in the same relationship that possible worlds stand to models: just as possible worlds are abstractions from models (so that worlds may be identified with equivalence classes of models), so laws are abstractions from theories (so that laws may be identified with equivalence classes of theories). In fact, we can relate it to the terms of Chapter 1, by taking laws to be intepreted theories. This picture of laws is clearly an unapologetic form of anti-Humeanism, although it's not clear to me that it aligns especially closely with extant anti-Humean accounts (necessetarianism, dispositionalism, etc.). Insofar as I am avoiding any kind of reductive account, the view is likely best characterised as a form of primitivism about laws.[2] But it means that we can connect the dispute over whether properties are tied to their laws to questions about how to interpret

[1] Where theories are understood in the sense described in Chapter 1, i.e., as syntactic conditions over semantic structures—whether in the formalism of first-order logic, or differential geometry, or whatever.

[2] [Maudlin, 2007b]

theories: the issue become a dispute over whether theories ought to be interpreted as inequivalent, if the (allegedly distinct) laws thereby obtained differ only over which properties play which nomological roles.

Thus, we need to say a little about the circumstances under which this arises. So suppose that $\mathbb{T}_1$ and $\mathbb{T}_2$ are two theories; let's label the law-systems which they represent, or to which they give rise, as Λ_1 and Λ_2. What would it be for the properties and relations of Λ_1 to manifest the same nomological structure as those of Λ_2? Presumably, the answer is just: if there is some kind of "nomological isomorphism" between the two sets of properties and relations, which maps each property of Λ_1 to a property of Λ_2 with the same nomological role. So let's see if we can make sense of this. The first step in defining an appropriate notion of isomorphism is typically to define some appropriate notion of homomorphism from the set of properties and relations for Λ_1 to the set of properties and relations for Λ_2. For the sake of convenience, I'll just speak of "properties" from here on as an abbreviation for "properties and relations".

Let's use Ξ_1 to denote the properties for Λ_1, and Ξ_2 to denote the properties for Λ_2. Our first task is just to get a grip on these sets. Again, we'll treat properties as abstractions of a certain sort: in this case, as abstractions from open formulae. The reason to take open formulae as the proper raw materials for the abstraction, rather than just the predicates, is simply in order that "complex" properties (e.g. *being both red and round*) are available. We therefore need to decide on the conditions for equivalence for such formulae (i.e. on the conditions under which two formulae should count as representing the same property).

One pretty clear candidate is logical equivalence: it is very implausible that the formulae Fx and $\neg\neg Fx$ should represent distinct properties. We will also, however, reckon as equivalent those formulae which are equivalent modulo the ambient theory. For instance, if a theory $\mathbb{T}$ contained the axiom $\forall x(Px \leftrightarrow Qx)$, then Px and Qx will be taken to represent the same property. Note that doing so amounts to taking (nomologically) necessary coextension as sufficient for property-identity. We will use $[\phi]$ to denote the property abstracted from the open formula ϕ, since we are identifying it with the equivalence class of ϕ under the relation of equivalence modulo $\mathbb{T}$. The above stipulations mean that there are well-defined "grammatical" relationships amongst the properties: relationships of the form, for example, of one property's being the negation of another, or being the conjunction of two others. Such relationships are determined by the analogous relationships amongst the open formulae, so that the property $[(Fx \wedge Gx)]$ is the conjunction of $[Fx]$ and $[Gx]$, etc.

Hence, Ξ_1 may be identified with the partition of $\mathrm{Form}(\Sigma_1)$ under the relation of equivalence modulo $\mathbb{T}_1$ (which we'll abbreviate $\equiv_1$), and the same (*mutatis mutandis*) for Ξ_2. We now want to define the notion of a nomological homomorphism from Ξ_1 to Ξ_2. Such a homomorphism should at least preserve the grammatical relationships above: if one property is the conjunction of two others, it ought to get mapped to the conjunction of whatever its conjuncts get mapped to. But more than that, it should preserve the "nomological relationships" in which the members of Ξ_1 stand. Here is a suggestion for how to enact that: I claim that the causal-nomological relationships within Ξ_1, according to Λ_1, are exactly the relationships that Λ_1 asserts to hold amongst the various members of Ξ_1. That suggests that the "mark" of a nomological homomorphism will be that for any such relationship asserted to hold by Λ_1 of some sequence of properties in Ξ_1, Λ_2 asserts the same relationship to hold amongst the sequence of properties to which they get mapped.

To make this more precise, let's return to our representation of the laws by theories. Since Ξ_1 and Ξ_2 are represented (redundantly) by $\mathrm{Form}(\Sigma_1)$ and $\mathrm{Form}(\Sigma_2)$, we expect that such a homomorphism will be representable by a suitable mapping $\mathrm{D} : \mathrm{Form}(\Sigma_1) \to \mathrm{Form}(\Sigma_2)$. Now, if we want to be sure that D preserves the grammatical relationships, then the simplest thing is to require that it commutes with the syntactic formation rules, i.e. that

- $\mathrm{D}(\neg\phi) = \neg\mathrm{D}(\phi)$
- $\mathrm{D}(\phi \wedge \psi) = (\mathrm{D}(\phi) \wedge \mathrm{D}(\psi))$
- $\mathrm{D}(\forall\xi\phi) = \forall\xi\mathrm{D}(\phi)$

Note that as a result, fixing the effect of D on all the *atomic* formulae serves to fix its effect on all formulae. Call a map which commutes with the rules in this fashion a *syntactic map* from the signature Σ_1 to the signature Σ_2.

Second, if we desire that D preserves the nomological relationships, then we should require that any consequence of $\mathbb{T}_1$ is turned into a consequence of $\mathbb{T}_2$, i.e., that if $\mathbb{T}_1 \vDash \phi$, then $\mathbb{T}_2 \vDash \mathrm{D}(\phi)$. (Note that a consequence of this is that we obtain a well-defined map $\Xi_1 \to \Xi_2$, in the sense that it does not depend on the choice of representative for a property: if $\phi \equiv_1 \psi$, then $\mathrm{D}(\phi) \equiv_2 \mathrm{D}(\psi)$.) But what we have come up with is a well-known model-theoretic construction: that of a *translation* from one theory to another.

Definition 5. Let $\mathbb{T}_1$ be a Σ_1-theory, and $\mathbb{T}_2$ a Σ_2-theory. A *translation from* $\mathbb{T}_1$ *to* $\mathbb{T}_2$ is a syntactic map D from Σ_1 to Σ_2 such that for every Σ_1-formula ϕ, if $\mathbb{T}_1 \vDash \phi$ then $\mathbb{T}_2 \vDash \mathrm{D}\phi$.

Thus, we will regard nomological homomorphisms as abstractions from translations. From here, defining a nomological isomorphism is straightforward: it is just an invertible nomological homomorphism. At the level of theories and formulae rather than laws and properties, this means that we can find a pair of translations D and D' which represent such a pair of inverse nomological homomorphisms. The necessary and sufficient condition for them to do so is *not* that they are themselves inverse to one another, but merely that they are *almost* inverse: the compositions of the two translations need not take every formula back to itself, but must take it to a formula which is equivalent (modulo $\mathbb{T}_1$ or $\mathbb{T}_2$, as appropriate). More precisely:[3]

Definition 6. A pair of translations $\mathrm{D} : \mathbb{T}_1 \rightarrow \mathbb{T}_2$ and $\mathrm{D}' : \mathbb{T}_2 \rightarrow \mathbb{T}_1$ comprise a *translational equivalence* between $\mathbb{T}_1$ and $\mathbb{T}_2$ iff for any Σ_1-formula $\phi(x_1, \ldots, x_m)$, and any Σ_2-formula $\psi(x_1, \ldots, x_n)$,

$$\mathbb{T}_1 \vDash \forall x_1 \ldots \forall x_m(\phi(x_1, \ldots, x_m) \leftrightarrow \mathrm{D}'\mathrm{D}\phi(x_1, \ldots, x_m)) \tag{3.1a}$$

$$\mathbb{T}_2 \vDash \forall x_1 \ldots \forall x_n(\psi(x_1, \ldots, x_n) \leftrightarrow \mathrm{D}\mathrm{D}'\psi(x_1, \ldots, x_n)) \tag{3.1b}$$

We will say that a pair of theories $\mathbb{T}_1$ and $\mathbb{T}_2$ are *translationally equivalent* if there is a translational equivalence between them. Note that it is crucial that D and D' be translations. For instance, suppose that Σ_1 and Σ_2 are a pair of signatures such that D is a one-to-one arity-preserving bijection between them (or rather, is the dictionary map corresponding to such a bijection), and that D' is the inverse (strictly, is the dictionary map corresponding to the inverse). Then the conditions (3.1) will be satisfied with respect to *any* pair of theories $\mathbb{T}_1$ and $\mathbb{T}_2$; but D and D' will not, in general, be translations.

Finally, let us relate this apparatus to notions of worlds and pictures: as ever, the models of T represent the nomologically possible worlds (according to the law-system T expresses). Here is an important observation about the relationship between syntactic maps and pictures: any syntactic map $\mathrm{D} : \Sigma_1 \rightarrow \Sigma_2$ naturally induces a dual map D^* from Σ_2-pictures to Σ_1-pictures (note that D^* goes in the "opposite direction" to D). Intuitively, given any Σ_2-picture, we add on the extensions of the Σ_1-predicates by taking the extension of $\Pi \in \Sigma_1$ to be all and only those n-tuples satisfying $\mathrm{D}(\Pi x_1 \ldots x_n)$.

[3]The below definition is due to [Barrett and Halvorson, 2015].

Formally, given any Σ_2-picture $\mathcal{S}$, $\mathrm{D}^*\mathcal{S}$ is the Σ_1 picture given by

$$D_{\mathrm{D}^*\mathcal{S}} = D_{\mathcal{S}} \tag{3.2a}$$

$$|\Pi x_1 \dots x_n|_g^{\mathrm{D}^*\mathcal{S}} = |\mathrm{D}(\Pi x_1 \dots x_n)|_g^{\mathcal{S}} \tag{3.2b}$$

It is easy to establish by induction that for any Σ_2-picture $\mathcal{S}$, any Σ_1-formula ϕ, and any variable-assignment g over $D_{\mathcal{S}}$,

$$|\mathrm{D}\phi|_g^{\mathcal{S}} = |\phi|_g^{\mathrm{D}^*\mathcal{S}} \tag{3.3}$$

In the event that D is a translation, we get a neat mapping from the models of $\mathbb{T}_2$ to those of $\mathbb{T}_1$: that is,

Proposition 1. Suppose that D is a translation $\mathbb{T}_1 \to \mathbb{T}_2$. Then for any model $\mathcal{M}$ of $\mathbb{T}_2$, $\mathrm{D}^*\mathcal{M}$ is a model of $\mathbb{T}_1$.

Proof. Suppose that $\mathrm{D}^*\mathcal{M} \not\models \mathbb{T}_1$. So for some $\phi \in \mathbb{T}_1$, $0 = |\phi|^{\mathrm{D}^*\mathcal{M}} = |\mathrm{D}\phi|^{\mathcal{M}}$. But since D is a translation, $\mathbb{T}_2 \vDash \mathrm{D}\phi$. And since $\mathcal{M}$ is a model of $\mathbb{T}_2$, $|\mathrm{D}\phi|^{\mathcal{M}} = 1$. So we have a contradiction: hence, by reductio, $\mathrm{D}^*\mathcal{M} \models \mathbb{T}_1$. □

Thus, for D a translation, we may regard D^* as representing a map from worlds that are possible according to Λ_2 to worlds that are possible according to Λ_1. And in the event that D is a translational equivalence, we get an even closer relationship between the two sets of models: [4]

Proposition 2. Suppose that we have translations $\mathrm{D} : \mathbb{T}_1 \to \mathbb{T}_2$ and $\mathrm{D}' : \mathbb{T}_2 \to \mathbb{T}_1$. Then D and D′ implement a translational equivalence between $\mathbb{T}_1$ and $\mathbb{T}_2$ iff D^* is a bijection $\mathrm{Mod}(\mathbb{T}_2) \to \mathrm{Mod}(\mathbb{T}_1)$, with $(\mathrm{D}')^*$ as its inverse.

Proof. First, assume first that D and D′ implement a translational equivalence between $\mathbb{T}_1$ and $\mathbb{T}_2$. I show that for any $\mathcal{M} \in \mathrm{Mod}(\mathbb{T}_2)$, $(\mathrm{D}')^*\mathrm{D}^*\mathcal{M} = \mathcal{M}$, i.e., that $(\mathrm{D}')^*\mathrm{D}^*$ acts on $\mathrm{Mod}(\mathbb{T}_2)$ as the identity. The proof that $\mathrm{D}^*(\mathrm{D}')^*$ acts on $\mathrm{Mod}(\mathbb{T}_1)$ as the identity goes similarly.

So consider any such $\mathcal{M}$. We have immediately that $D_{(\mathrm{D}')^*\mathrm{D}^*M} = D_{\mathrm{D}^*M} = D_M$. So

[4] This result is standard (although the statement of it has been tweaked to mesh with the above definition of translational equivalence): see, for example, [de Bouvère, 1965, Theorem 2] or [Hodges, 1997, p. 54].

now consider any $\Pi \in \Sigma_2$. By the above lemma,

$$\begin{aligned} |\Pi x_1 \dots x_n|_g^{(\mathrm{D}')^* \mathrm{D}^* \mathcal{M}} &= |\mathrm{D}'(\Pi x_1 \dots x_n)|_g^{\mathrm{D}^* \mathcal{M}} \\ &= |\mathrm{D}\mathrm{D}'(\Pi x_1 \dots x_n)|_g^{\mathcal{M}} \end{aligned}$$

But since $\mathcal{M} \models \mathbb{T}_2$ and D, D' implement a translational equivalence,

$$\mathcal{M} \models \forall \mathbf{x}(\Pi \mathbf{x} \leftrightarrow \mathrm{D}\mathrm{D}'\Pi \mathbf{x})$$

and so $|\mathrm{D}\mathrm{D}'\Pi \mathbf{x}|_g^{\mathcal{M}} = |\Pi \mathbf{x}|_g^{\mathcal{M}}$. Thus, $\Pi_{(\mathrm{D}')^* \mathrm{D}^* \mathcal{M}} = \Pi_{\mathcal{M}}$. Thus, $(\mathrm{D}')^* \mathrm{D}^* \mathcal{M} = \mathcal{M}$.

Second, assume that D^* and $(\mathrm{D}')^*$ are mutually inverse. I show that for any Σ_2-formulae ψ,

$$\mathbb{T}_2 \vDash \forall \mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathrm{D}\mathrm{D}'\psi(\mathbf{x})) \tag{3.4}$$

So suppose that (3.4) did not hold. Then there would be some model $\mathcal{M}$ of $\mathbb{T}_2$ such that $\mathcal{M} \not\models \forall \mathbf{x}(\psi(\mathbf{x}) \leftrightarrow \mathrm{D}\mathrm{D}'\psi(\mathbf{x}))$; i.e., such that for some $\mathbf{a}$ from $D_{\mathcal{M}}$, either $\mathcal{M} \models \psi[\mathbf{a}]$ and $\mathcal{M} \not\models \mathrm{D}\mathrm{D}'\psi[\mathbf{a}]$, or vice versa. But by the above lemma, $\mathcal{M} \models \psi[\mathbf{a}]$ iff $\mathcal{M} \models \mathrm{D}\mathrm{D}'\psi[\mathbf{a}]$. So by reductio, (3.4) holds. By similar reasoning, we can show that the parallel claim for $\mathbb{T}_1$ holds; hence, the pair $(\mathrm{D}, \mathrm{D}')$ is a translational equivalence. □

With this formal apparatus in hand, we can identify the following two kinds of attitudes towards interpreting a pair of translationally equivalent theories $\mathbb{T}_1$ and $\mathbb{T}_2$. Quidditists will hold that such a pair will, in general, represent different possible systems of laws even if there is a translational equivalence $(\mathrm{D}, \mathrm{D}')$ between them; and correspondingly, that a given model $\mathcal{M}$ of $\mathbb{T}_2$ and its "corresponding" model $\mathrm{D}^* \mathcal{M}$ of $\mathbb{T}_1$ will represent different possible worlds. By contrast, the anti-quidditist holds that $\mathbb{T}_1$ and $\mathbb{T}_2$ represent the *same* system of laws, and that $\mathcal{M}$ and $\mathrm{D}^* \mathcal{M}$ represent the *same* possible world.

Having now given our debate precise content, I go on to consider why one should be an anti-quidditist. The next section discusses a semantic problem for the quidditist; section 3.3 discusses an epistemic problem.

3.2 The semantic concern

In this section, I want to outline a semantic objection to quidditism. However, the nature of that objection depends on the account of possible worlds being appealed to. First, suppose that you reject the account of possible worlds given in Chapter 1:

so possible worlds are represented by models, but are not identical to them (or to equivalence classes of them). Suppose that D is a translational equivalence from $\mathbb{T}_1$ to $\mathbb{T}_2$, and consider any model $\mathcal{M}$ of $\mathbb{T}_2$. The quidditist maintains that the two models $\mathcal{M}$ and $\mathrm{D}^*\mathcal{M}$ respectively represent distinct possible worlds W and W', differing from one another in virtue of featuring different properties. In other words, what we have is a putative correspondence between the mathematical models and the physically possible worlds; let's call this correspondence I. However, there is now a problem: we can put forward a systematic reinterpretation of the models which permutes which worlds they represent. Specifically, whereas $P \in \Sigma_1$ is supposed to represent some property $I(P)$ according to the correspondence I, we instead take it to represent $I^*(P) := I(\mathrm{D}(P))$. Conversely, for any $Q \in \Sigma_2$, we set $I^*(Q) := I(\mathrm{D}'(Q))$.

Under I^*, $\mathcal{M}$ represents W' and $\mathrm{D}^*\mathcal{M}$ represents W. What makes this interpretation scheme any worse than I? You might say: because we *stipulated* that $\mathcal{M}$ was to represent the possible world W (and $\mathrm{D}^*\mathcal{M}$ to represent W'). But all that does is push the problem back to the label "W", whose connection to any given possible world is just as questionable as $\mathcal{M}$'s. Alternatively, one could try appealing to the two sets of laws: $\mathcal{M}$, being a model of $\mathbb{T}_2$ (rather than $\mathbb{T}_1$), represents a world governed by the laws Λ_2 expressed by $\mathbb{T}_2$ (rather than by the laws Λ_1 expressed by $\mathbb{T}_1$). But similarly, this just brings us to the question of how it comes to be that $\mathbb{T}_2$ expresses Λ_1, rather than Λ_2 (and vice versa for $\mathbb{T}_1$). In effect, we are exploiting the fact that because $(\mathrm{D}, \mathrm{D}')$ is a translational equivalence, interpreting "is P" as "is $\mathrm{D}(P)$" is an acceptable reinterpretation. The equivalence of nomological structure ensures that the property (allegedly) expressed by $\mathrm{D}(P)$ is exactly as eligible to be the referent of P as is the property (allegedly) expressed by P.

This argument is not bulletproof, however. The chink in the armour is as follows: it presumes that eligibility is exhausted by nomological role. As such, it may be resisted by someone who believed that semantic eligibility can depend upon extra-nomological factors. For instance, suppose one believes that in addition to the nomological (and logical) relationships amongst properties, there are facts about which properties are *natural*.[5] Such a person could then claim that a natural property is more eligible to be the referent of an atomic predicate (such as P) than a non-natural property would be; conversely, they could claim that a non-natural property would be a more appropriate candidate to be the referent of a complex predicate (such as, in general, $\mathrm{D}(P)$). They could then claim that W and W' disagree over precisely these kinds of

[5]Such a proposal is most famously associated with Lewis: see [Lewis, 1983].

facts about which properties are natural, and which are derivative; correlatively, they would be able to say that one of I or I^* is privileged by virtue of assigning the atomic predicates to the natural properties. Moreover, the appeal to naturalness in particular is inessential: the same move can be made using other extra-nomological accolades (e.g. fundamentality). In brief, this solution involves two moves: the introduction of extra-nomological structure as a means of differentiating between nomologically equivalent worlds, and then the use of the structure of definability as the model-theoretic correlate of this extra structure.

A response has to accept that these show how one could make the representation relation determinate, after all. However, it bears emphasising that they do so only by importing resources external to the theory itself. For all the theory cares, one can use I, or I^*, or whatever correspondence one likes; and that observation stands, regardless of whether extra-theoretical considerations of naturalness get invoked to privilege one correspondence over the others. Therefore, we should be extremely wary of the casual appropriation of terms like "natural" or "fundamental" for these purposes. If we are at all naturalistically inclined, we should very strongly resist the idea that this structure—which as I've stressed, must go above and beyond the structure codified in the natural laws—is in any way "natural"! Thus, insofar as one is interested in accounting for the meaning imbued upon the parts of a theory by their role within that theory, the reference of these models remains indeterminate—unless, that is, they are all taken to refer identically, to one and the same possible world.

Suppose, on the other hand, that one accepts the story about possible worlds I suggested in Chapter 1. Then the quidditist could maintain that the possible worlds are to be identified with the models $\mathcal{M}$ and $\mathrm{D}^*\mathcal{M}$ directly, rather than with equivalence classes of such models. As a result, there is no problem about which possible world $\mathcal{M}$ or $\mathrm{D}^*\mathcal{M}$ represents: they represent themselves! Nor is there the kind of problem we saw in Chapter 2, where there is a disconnect between the models and the sentences: being models of theories with different signatures, it is straightforward to have $\mathcal{M}$ be picked out by different sentences to those picking out $\mathrm{D}^*\mathcal{M}$.

However, this points to a different kind of semantic problem with possible worlds conceived of in this fashion: they are unacceptably language-dependent. The reason why a Σ_2-sentence can only be satisfied by $\mathcal{M}$ rather than $\mathrm{D}^*\mathcal{M}$ is because only $\mathcal{M}$ contains Σ_2-themed structure, i.e., extensions for Σ_2-predicates rather than Σ_1-predicates. But this kind of structure—structure concerning how extensions get labelled—seems like a clear example of something which is an "artifact of the model" (in [Kaplan, 1975]'s

phrase) rather than something of genuine representational significance. And to reject that structure will involve collapsing haecceitistic distinctions. In other words, the quidditist faces a dilemma: the more "metaphysical" they take the quidditistic stucture to be, the more opaque it becomes to our representational grasp; the more "representational" they make it, the less it seems like the kind of thing that should be postulated as a real difference in the world.

That said, this also opens the way for rival anti-quidditist views to the one explicated here, corresponding to different ways in which the models might be rendered "language-independent". In particular, the literature on philosophy of science suggests two alternative ways in which linguistic structure might be expurgated; so we should assess these proposals, and see how they compare. On the syntactic side, one might argue that we should work with Ramseyfied theories instead. As a reminder, recall that Ramseyfying a Σ-theory $\mathbb{T}$ consists of four steps. First, one thinks of $\mathbb{T}$ as a set of second-order rather than first-order sentences. Second, one forms the conjunction $\bigwedge \mathbb{T}$ of all elements of $\mathbb{T}$ (if $\mathbb{T}$ is not finite, then this will require upgrading to infinitary logic). Third, for each $\Pi_i \in \Sigma$, we uniformly substitute a distinct second-order variable X_i (of the appropriate arity), to obtain $\bigwedge \mathbb{T}(X_1/\Pi_1, \dots)$. Finally, we prefix an existential second-order quantifier for each variable thus introduced, to finally obtain the Ramsey sentence of the theory,

$$\mathbb{T}^R := \exists X_1 \cdots \bigwedge \mathbb{T}(X_1/\Pi_1, \dots) \tag{3.5}$$

Again, infinitary resources will be necessary (this time, to allow infinitary quantification) unless only finitely many elements of Σ occur in $\mathbb{T}$.

However, Ramseyfication leads to difficulties. First, note that unless $\mathbb{T}$ is a fairly simple theory, we will require a significant escalation in the background expressive and logical resources: from finitary first-order logic to infinitary second-order logic. Second, it is plausible that Ramseyfying leads to an excessive loss of structure. In particular, we face the *Newman problem*:[6] if we use the standard semantics for second-order logic, then a Ramsey sentence $\mathbb{T}^R$ will be satisfied by any standard Σ-picture of the same cardinality as some model of $\mathbb{T}$. To my mind, this observation just constitutes a good reason to think that Ramseyfication should not be coupled with adherence to the standard second-order semantics, but rather with something like Henkin semantics. Although I don't have the space to explore this here, my suspicion is that doing so would lead to a position more or less equivalent to that defended here: the characteristic feature of a Henkin-picture, after all, is that the second-order quantifiers are only

[6] [Newman, 1928]; see [Ketland, 2004] or [Ainsworth, 2009] for contemporary discussion.

required to range over the *definable* relations (rather than, as in a standard second-order picture, over *all* relations).

Alternatively, on the semantic side, one might propose just stripping off the labels from the predicate-extensions in the models. Here is one way this could be made precise.[7] Given a Σ-picture $\mathcal{M}$, define the *H-picture corresponding to* $\mathcal{M}$, $\mathcal{M}^H$,[8] to consist of the following data:

- The set $D_{\mathcal{M}}$
- The multiset[9] $[\Pi^{\mathcal{M}} : \Pi \in \Sigma]$ containing the extensions of the predicates in Σ

So, one could claim that H-pictures (rather than Σ-pictures) ought to be identified as possible worlds, with two Σ-pictures $\mathcal{P}$ and $\mathcal{Q}$ being equivalent just in case $\mathcal{P}^H = \mathcal{Q}^H$ (or perhaps, just in case $\mathcal{P}^H \cong \mathcal{Q}^H$). The idea is that by (as it were) forgetting which extension belongs to which predicate, we have purged the dependence on language.

How does this compare with the above? There is one sense in which this retains more structure than the view I am defending, and another sense in which this gets rid of more structure than my view. Here is the sense in which more structure is retained: in general, for a translational equivalence $\mathrm{D} : \mathbb{T}_1 \to \mathbb{T}_2$ and model $\mathcal{M}$ of $\mathbb{T}_2$, it is not the case that $\mathcal{M}^H = (\mathrm{D}^*\mathcal{M})^H$. In fact, the only translations for which $\mathcal{M}^H = (\mathrm{D}^*\mathcal{M})^H$ are "mere relabelling" translations, in which every predicate in Σ_1 is mapped to an atomic formula of Σ_2, and vice versa. The basic reason is that delabelling a model, in the manner envisaged above, gets us the extensions only of the *atomic* predicates, rather than complex predicates: so if $\mathrm{D}(P)$ is a complex formula, then its extension in $\mathcal{M}$ will not be "visible" in $\mathcal{M}^H$. Here is the sense in which more structure is expunged: if an entire class of models is de-labelled, then we lose the data about which extensions in one model correspond to those in another (where prior to de-labelling, that correspondence was encoded in the fact that *this* extension in one model and *that* extension in another were extensions of the same predicate). As a result, two theories can yield the same class of H-models even if they are not translationally equivalent.[10]

The relevant question is then whether these differences are an advantage or a disadvantage. The former difference could be an advantage for the advocate of naturalness

[7] Due to [Lutz, 2015].

[8] The term "H-picture" (following Lutz's terminology of "H-structure") is a reference to [Halvorson, 2012], who looks at the (problematic) consequences for the semantic view of theories of understanding theories to be sets of H-models.

[9] A multiset is like a set, except that the same element can recur more than once (see [Blizard, 1988]).

[10] This pair of observations forms the core of [Halvorson, 2012]'s critique of explicating the semantic view of theories in terms of H-models.

or fundamentality: by retaining the distinction between atomic and complex predicates in this way, we can retain the (alleged) distinction between the natural/fundamental properties, and the unnatural/derivative ones. I don't find this compelling, since (for the reasons outlined above) I am skeptical of the appeal to such extra-nomological virtues as these. The latter feature seems to be more unambiguously a problem, since there is good intuitive reason to reject the criterion of theoretical equivalence that results from this. Consider the following example (due to [Halvorson, 2012]). Let $\Sigma_P = \{P_i\}_{i\in\mathbb{N}}$ and $\Sigma_Q = \{Q_i\}_{i\in\mathbb{N}}$. Let the Σ_P-theory $\mathbb{T}_P$ consist of the sole axiom

$$\exists x((x = x) \wedge \forall y(y = x)) \tag{3.6}$$

asserting that there is exactly one thing. Let the Σ_Q-theory $\mathbb{T}_Q$ consist of the following axioms

$$\begin{gathered}\exists x((x = x) \wedge \forall y(y = x)) \\ \forall x(Q_0 x \rightarrow Q_i x)\end{gathered} \tag{3.7}$$

where the second line is an axiom schema, with the metalinguistic variable Q_i ranging over all the predicates $Q_0, Q_1, \ldots$. Thus, $\mathbb{T}_Q$ asserts not only that there is exactly one thing, but that there is a special property such that if the thing has that property then it has every other property. Intuitively, $\mathbb{T}_P$ and $\mathbb{T}_Q$ are inequivalent theories: they say different things about the properties found in the world. In particular, it seems right to say that the nomological role of Q_0 in $\mathbb{T}_Q$ is not the same as the nomological role of any P_i in $\mathbb{T}_P$. Yet the class of H-models of $\mathbb{T}_P$ is the same as the class of H-models of $\mathbb{T}_Q$: for each such model consists of just a single-membered domain, together with a countable infinity of subsets of that domain. By contrast, $\mathbb{T}_P$ and $\mathbb{T}_Q$ are not translationally equivalent (precisely because there is no way of translating Q_0 into Σ_P).

Clearly, more work needs to be done to fully explore these alternative ways in which we might seek to make our account of possible worlds language-independent. For now, however, I conclude that explicating language-independence in the manner here—i.e., of identifying models related by a translational equivalence—is the most promising.

3.3 The epistemic concern

Secondly—and analogously to haecceitism—there is a concern that quidditism suffers from an epistemic defect. Now, the haecceitistic argument for this conclusion proceeded simply from the observation that a given world w and its permuted variant w' have the

same distribution of qualitative properties, i.e. (as we put it) are isomorphic. So, given that our evidence about what the world is like appears to be isomorphism-invariant, we immediately got out the conclusion that we could not have knowledge of whether we are in w or w'. Clearly, in the quidditistic case things will have to be a little more subtle. Suppose that $\mathcal{M}$ and $\mathrm{D}^*\mathcal{M}$ (for some translational equivalence $\mathrm{D} : \mathbb{T}_1 \to \mathbb{T}_2$) correspond, according to some quidditist, to distinct possible worlds W and W'. To claim that W and W' agree on the distribution of qualitative properties is just to beg the question against the quidditist. So we want an argument that even supposing W and W' to be qualitatively distinct, they are nevertheless epistemically equivalent: that we can never have knowledge that we are in W rather than W'.

We can have such an argument.[11] The idea is to make precise the intuitively compelling idea that our knowledge of the intrinsic properties of objects is entirely "mediated" by their causal/nomological roles—and, hence, that we can have no knowledge of the difference between cases in which the same nomological structure obtains. To do so, we reason as follows. Suppose that quidditistic knowledge was, in fact, obtainable. That is just to say that it is possible for such knowledge to be obtained. So let W be a possible world verifying this possibility. Thus, let us imagine that W is a world containing some kind of "noumenometer": a device which, when exposed to a property P of some kind, generates a certain output depending on what the quiddity of the property is. Call this output O.

Now, in order that the noumenometer count as generating *knowledge* of the quiddities, it has to exhibit a certain level of reliability. It had better be the case that had it been exposed to some other property P', it would not have generated the output O, but instead some other output O'. Moreover, if this really is to be a noumenometer—sensitive to P's internal, quidditistic nature, rather than merely to its nomological role—then it had better do so *counterlegally*, that is, independently of what role P plays in the laws. In particular, the noumenometer ought to generate a different output even if it its exposure to P is the result of a *merely quidditistic* change. Just to fix things, let us suppose that $\mathcal{M}$ represents the world W; that $\mathbb{T}_2$ represents the laws governing W; that we have a translational equivalence $\mathrm{D} : \mathbb{T}_1 \to \mathbb{T}_2$ (with "inverse" translation D'), where $\mathbb{T}_1$ represents a (merely quidditistically) different set of laws than $\mathbb{T}_2$; and that $\mathrm{D}'(P) = P'$ (abusing notation by letting P and P' stand for both properties and predicates). So $\mathrm{D}^*\mathcal{M}$ represents a world W' that differs from W in the prescribed

[11] The argument outlined here is an adapted version of an argument originally given by [Roberts, 2008]; see [Dasgupta, 2011], [Dasgupta, 2014b] for discussion.

manner: P', plays the nomological role of P in the laws, and (roughly speaking) is found in W' wherever P was found in W. So, in particular, the noumenometer is exposed to P' rather than P in W'.

Thus, if the noumenometer is to be reliable, then it had better be the case that it not generate O as output in W', but generates (say) O' instead. Thus, if $\mathrm{D}'(O) = O$—that is, if O is *invariant* under the quidditistic change—then the noumenometer is not reliable, and hence does not count as a way to obtain knowledge of quiddities. Moreover, note that all we require is that there is *some* translational equivalence E (with inverse E') such that $\mathrm{E}'(P) \neq P$ and $\mathrm{E}'(O) = O$. For even if $\mathrm{D}'(O) \neq O$, with the result that the noumenometer reliably distinguishes W from W' (at least with respect to discriminating P from P'), the existence of E indicates that there is some world W'' (i.e. that represented by $\mathrm{E}^*\mathcal{M}$) in which the noumenometer indicates O despite not having been exposed to P. And the latter is all that we need to impugn the reliability of the noumenometer, in such a way that it fails to be a generator of quidditistic knowledge.

Therefore, the only way for the noumenometer to function is if there is no transformation that leaves O invariant whilst changing P. But the only way that this could be true is if O and P are related by what I earlier called the "grammatircal structure". For instance, suppose that O is the property of being P and Q. Then any quidditistic change to P necessarily brings about a change to O. Or, even more directly, suppose that O just *is* P! Then it is very clear that O is invariant only if P is invariant. Either way, we find that the noumenometer acts as a kind of *trivial* detector: the property it uses as a signal of the presence of P is defined in terms of P. And the problem in *this* case is that this "detection" is clearly bogus. An instance of P cannot serve as evidence that P, or at least not as *mediating* evidence for P. Unless we already have some means of directly detecting P, then having a noumenometer that delivers P or $(P \wedge Q)$ as output is of no help.

Hence, we have at least reduced the problem before us: if we are not capable of *directly* perceiving merely quidditistic differences, then we are not capable of indirectly doing so. In other words, it shows that the unobservability of such differences is sufficient for their undetectability. The question then becomes: is it plausible that such differences are unobservable? I don't think that a decisive answer to this question can be given. This is unfortunate but unsurprising. The question amounts, in effect, to asking whether there might be something to the direct, intrinsic phenomenology of property-perception that is left underdetermined by whatever nomological process underpins that perception. But of course, that question is the subject of a vast literature in the

philosophy of mind; indeed, whether the causal-nomological workings of the brain uniquely fix the phenomenological character of qualia is one of the central research topics in contemporary philosophy of mind.[12]

Clearly, this is not the place to fully settle this question, but it is worth reminding ourselves of the costs of subscribing to such epiphenomenal qualia—or at least of supposing that such qualia are capable of generating knowledge. That is, suppose that we were directly sensitive to quiddities. What would that be like? Presumably, it would mean that the nature of our experience would depend directly on which property it was we were experiencing, independently of what causal role that property is playing. So, for example, suppose that our experiences of colour are directly tied to colour-quiddities in this way.[13] When I see an object which is green, the distinctive qualia that I associate with that experience is a result of it being that very property (rather than any other) which I am experiencing. In particular, there is a difference between the phenomenological experiences I actually have, and those I would have in a world differing merely quidditistically from actuality: say, in a world in which greenness plays the role of blueness, and vice versa. But note that such "extra-nomological" phenomenological profiles will, of necessity, be independent of any manifestation of experience in external behaviour. Clearly, my propensity to say "I can see green!", or to classify the seen object with other objects of the same colour, or to indicate it as an answer to the question "what kind of colour is green?", etc., is the same in both worlds, independently of which qualia I experience. (Moreover, it seems that I would be *right* to engage in all of these activities, in either world—more on this in a moment.)

In other words, we can restore the correlation between the presence of certain quiddities and my internal mental state, but only at the cost of severing any correlation between that internal mental state and overt behaviour. One can maintain such a theory if one likes: but note how scrawny a notion of detectability has been rescued. Yes, intrinsic quiddities can be detected; but they can't be communicated through normal physical channels, or used to guide action, or do any of the other things that we usually think of detections as there to do. Our original concern was that quiddities were metaphysical danglers; all that this manoeuvre has gained us, it seems, is the trading of metaphysical for mentalistic danglers.

Finally, it should be noted that—again, as with haecceitism—there seem to be good reasons to regard this kind of ignorance as inexpressible. After all, as discussed in

[12] See [Chalmers, 1997] and references therein.

[13] Historically, this position is associated with Russell [Russell, 1910]; [Hildebrand, 2015] defends this position within the modern debate on quidditism.

§3.2, there seems to be no way of determinately referring to these properties except via their causal roles.[14] (It is for this reason that, even if associated to distinct internal phenomenological experiences, I am right in treating whatever plays the green-role as the bearer of the term "green".) But, as with haecceitism, I don't believe that this should be regarded as the source of any great disquiet. By treating epistemic possibilities in line with metaphysical possibilities—i.e., as the sorts of things represented by models—then we straightforwardly see that quidditism generates multiple epistemic possibilities, which (by the above reasoning) cannot be cut down by any piece of evidence. So we have insoluble ignorance, of the kind which motivates a change to the metaphysics that has generated it.

3.4 Anti-quidditism and symmetries

The above constitutes my defence of the claim that in general, we should not take there to be any worlds which differ merely quidditistically. I now turn to considering a special case of this, of great importance for the remainder of this thesis. We have seen that the notion of worlds differing merely quidditistically may be given precise content by using the apparatus of translational equivalences between a pair of theories $\mathbb{T}_1$ and $\mathbb{T}_2$. The special case is when $\mathbb{T}_1 = \mathbb{T}_2$: that is, when we are considering translations of a theory into itself. I will refer to a translational equivalence from a theory into itself as a *symmetry* of the theory (the connection to physical symmetries will become clear in the next chapter).

Of course, for any theory, the trivial translational equivalence $(\mathrm{Id}, \mathrm{Id})$ is a symmetry. But many theories have non-trivial such symmetries. For example, consider the following theory $\mathbb{T}_H$:

$$\forall x(Lx \vee Rx) \tag{3.8a}$$

$$\forall x \neg (Lx \wedge Rx) \tag{3.8b}$$

We may heuristically think of this as a (very simple) theory about worlds in which there is nothing but gloves: in such worlds everything is either left-handed or right-handed,

[14] cf. [Lewis, 2009], [Langton, 2004], [Dasgupta, 2015].

but nothing is both. Now consider the syntactic map S such that

$$\mathrm{S}(L\xi) = R\xi \tag{3.9a}$$
$$\mathrm{S}(R\xi) = L\xi \tag{3.9b}$$

for any variable ξ. It is easy to see that S is a translational equivalence between $\mathbb{T}_H$ and itself, with S as its own inverse.

The important observation is that if a theory admits symmetries, then there are multiple models *of that theory* which represent worlds differing, at most, merely quidditistically. For instance, consider the following pair of models $\mathcal{N}$ and $\mathcal{N}'$ of $\mathbb{T}_H$.

$$D_{\mathcal{N}} = D_{\mathcal{N}'} = \{0, 1, 2\} \tag{3.10a}$$
$$L^{\mathcal{N}} = R^{\mathcal{N}'} = \{0\} \tag{3.10b}$$
$$R^{\mathcal{N}} = L^{\mathcal{N}'} = \{1, 2\} \tag{3.10c}$$

Clearly, $\mathcal{N}' = \mathrm{S}^*\mathcal{N}$, and $\mathcal{N} = \mathrm{S}^*\mathcal{N}'$. Hence, the considerations adduced in §§3.2 and 3.3 apply, and we have good reason to deny that $\mathcal{N}$ and $\mathcal{N}'$ represent distinct possible worlds: if they did, then we would be unable to determinately refer to one such world rather than the other, and no inhabitant of such a world could have knowledge that they were in one world rather than the other. (Exactly how much knowledge gloves could have at all is, for our purposes, a moot point.)

To be clear, this does not mean interpreting the theory so that the two handedness properties are *identified* with one another.[15] Rather, the view is that when there *are* two properties with the same profile, there is no fact of the matter about which property-instantiation in a given possible world is an instantiation of which property. In each world there are two classes of congruence counterparts, each of which is the extension of a handedness property; but there is no preferred way of matching up a congruence class in one world with one in another world, that is, of identifying such pairs of congruence classes as the extensions of "the same" handedness property as one another. That said, *relative* to an (arbitrarily chosen) identification of the congruence-class in one world with a congruence-class in another, there is a privileged way of identifying the remaining congruence-classes: they had better be identified with each other, since the distinction between the classes in each model has to be preserved.

Thus, we find that anti-quidditism has the following consequence:

[15]Compare the discussion in [Hawthorne, 2001, Part Three].

> The symmetry-interpretation link. For a theory containing symmetries, we should not interpret that theory in such a way that the symmetry-related models (i.e., models related by a map induced by a symmetry) represent distinct possible worlds.

In other words, we are led to a position that mirrors the anti-haecceitist, in denying that all distinct world-representations represent distinct possible worlds. Note that whereas this phenomenon was generic when considering anti-haecceitism, for anti-quidditism it only arises when dealing with symmetries (in the sense above). The reason for this is that a merely haecceitistic change—a permutation of the individuals—is generically consistent with keeping the laws the same, whereas it is only if the laws admit symmetries that (some) merely quidditistic changes are consistent with keeping the laws the same.

One interesting observation here is that such cases may be viewed as illustrating Quine's famous thesis of the indeterminacy of translation. After introducing the infamous field linguist trying his best to translate "gavagai" as either "rabbit" or "rabbit part", Quine notes that as he offers more and more hypotheses relating terms of the native language to those of English, and as those hypotheses continue to match sentences to sentences of similar usage, he will feel more and more confident that those hypotheses are correct. However, Quine continues,

> [...] it seems that this method, though laudable in practice and the best we can hope for, does not in principle settle the indeterminacy between "rabbit", "undetached rabbit part", and "rabbit stage". For if one workable system of analytical hypotheses provides for translating a given native expression into "is the same as," perhaps another equally workable but systematically different system would translate that native expression into something like "belongs with". Then when in the native language we try to ask "Is this *gavagai* the same as that?" we could as well be asking "Does this *gavagai* belong with that?" Insofar, the native's assent is no objective evidence for translating "gavagai" as "rabbit" rather than "undetached rabbit part" or "rabbit stage".[16]

If we examine the above carefully, we see that a remarkable amount of work is being done by that "perhaps": of course there *might* be some workable and coherent set of alternative translation hypotheses, but it is hard to see what makes Quine so sure.[17]

[16] [Quine, 1969, p. 33]
[17] cf. Dennett's remarks on "cryptographic constraints" [Dennett, 2000].

In the case of symmetries, by contrast, we can explicitly construct such cases of underdetermination. For instance, it is clear that if we were translating the theory $\mathbb{T}_H$ into a theory $\mathbb{T}'_H$, differing from $\mathbb{T}_H$ only in the use of the predicates F and G rather than L and R, then there are two equally well-qualified translations: that taking L to F and R to G, or vice versa. This in turn is just a result of the fact that there are two equally well-qualified translations from $\mathbb{T}_H$ to itself (the identity, or the translation S).

3.5 Reduction and sophistication

Now, we saw in the previous chapter that enacting anti-haecceitism—taking prima facie distinct world-representations to represent the same possibility—required work. Specifically, we had to show that the semantical purposes to which we had been putting the excess representations could equally well be served by a more perspicuous set of possibility-representations. The same holds true here. We should not simply declare that symmetry-related models are equivalent to one another; given the role that such models play in explaining the way in which meaning works in our theories, we need to show that a coherent account of theoretical meaning is still available following such a declaration. But there are two ways in which this might be done, which it will be fruitful to compare.

The first way is just to give a theory whose symmetry-related models are isomorphic to one another. To exhibit such a theory, I introduce the concept of a *hands-free picture*: a hands-free picture m comprises

- A set D_m
- A two-element set $\mathbf{2}^m$
- A function $\chi_m : D_m \to \mathbf{2}^m$

The point of doing so comes in the introduction of a new definition of "homomorphism" for such pictures: we take a homomorphism $h : m \to n$ to comprise a map $h_1 : D_m \to D_n$ and a bijection $h_2 : \mathbf{2}^m \to \mathbf{2}^n$, such that for any $a \in D_m$,

$$\chi_n(h_1(a)) = h_2(\chi_m(a)) \tag{3.11}$$

In other words, we relax the requirement that isomorphisms must preserve the extensions of predicates: instead, they may map the extension of one predicate to the

extension of the other. To compose a pair of such homomorphisms, simply compose the components.[18]

I now explain how hands-free pictures determine truth-values. It will no longer be the case that a picture determines an unambiguous truth-value for every sentence of the handedness language: for a sentence like $\exists x Lx$, for example, there is no privileged way to determine which of the two "extensions" in the picture ought to count as the extension of L. But this is as it should be, if we are really interested in doing away with the structure that is variant under the symmetry: sentences which are not invariant under the symmetry are defective, if we do not take symmetry-variant structure seriously. Instead, truth in a hands-free picture m is (generally) relativised to a bijection $V : \{L, R\} \to \mathbf{2}^m$. In a certain sense, it is as though the predicate-letters L and R are being treated as second-order variables (although they can only range over $\mathbf{2}^m$); we will therefore refer to the map V as a second-order variable-assignment. Relative to such an assignment V, and to a first-order variable-assignment g, the truth-values of atomic sentences in a model m are determined as follows:

$$\begin{aligned} |Lx|^m_{g,V} = 1 \text{ iff } \chi_m(g(x)) = V(L) \\ |Rx|^m_{g,V} = 1 \text{ iff } \chi_m(g(x)) = V(R) \end{aligned} \tag{3.12}$$

The clauses for non-atomic sentences are unchanged. (These semantics could fruitfully be compared to either second-order semantics or supervaluationist semantics.) We then obtain the following result.

Proposition 3. Suppose that ϕ is logically equivalent to $\mathrm{E}\phi$, let m be a hands-free picture, and let g be a first-order variable-assignment for m. Then for any second-order variable-assignments V and V' for m,

$$|\phi|^m_{V,g} = |\phi|^m_{V',g} \tag{3.13}$$

Proof. Clearly, there only are two second-order variable-assignments for m (since $\mathbf{2}^m$ has only two members); so if $V \neq V'$, then we have that $V(L) = V'(R)$ and $V(R) = V'(L)$.

[18] I thank Thomas Barrett for inquiring after composition.

Let $\mathcal{M}$ and $\mathcal{M}'$ be Σ_H-pictures defined as follows:

$$D_{\mathcal{M}} = D_{\mathcal{M}'} = D_m \tag{3.14a}$$
$$L^{\mathcal{M}} = V(L) \tag{3.14b}$$
$$R^{\mathcal{M}} = V(R) \tag{3.14c}$$
$$L^{\mathcal{M}'} = V'(L) = V(R) = R^{\mathcal{M}} \tag{3.14d}$$
$$R^{\mathcal{M}'} = V'(R) = V(L) = L^{\mathcal{M}} \tag{3.14e}$$

In other words, $\mathcal{M}' = \mathrm{E}^*\mathcal{M}$. But clearly, $|\phi|^m_{V,g} = |\phi|^{\mathcal{M}}_g$, and $|\phi|^m_{V',g} = |\phi|^{\mathcal{M}'}_g$. Hence:

$$\begin{aligned} |\phi|^m_{V,g} &= |\phi|^{\mathcal{M}}_g \\ &= |\mathrm{E}\phi|^{\mathcal{M}}_g \\ &= |\phi|^{\mathrm{E}^*\mathcal{M}}_g \\ &= |\phi|^{\mathcal{M}'}_g \\ &= |\phi|^m_{V',g} \end{aligned}$$

□

As a consequence, the truth-value of any parity-invariant formula is unambiguously determined by a hands-free picture (together with a first-order variable-assignment). Note that all the members of $\mathbb{T}_H$ are (of course) logically equivalent to their "swapped" versions. Hence, we can define the hands-free models of $\mathbb{T}_H$ as those hands-free pictures which make $\mathbb{T}_H$ true. We then obtain the following.

Proposition 4. Suppose that ϕ is equivalent modulo $\mathbb{T}_H$ to $\mathrm{E}\phi$, let m be a hands-free model of $\mathbb{T}_H$, and let g be a first-order variable-assignment for m. Then for any second-order variable-assignments V and V' for m,

$$|\phi|^m_{V,g} = |\phi|^m_{V',g} \tag{3.15}$$

Proof. As above, but restricting to models of $\mathbb{T}_H$. □

We can therefore take our new theory to be given by the same set of sentences as the theory $\mathbb{T}_H$, but where the (putative) semantics for those sentences is that just outlined (i.e. is done in terms of hands-free pictures, rather than handed pictures). Let us denote this theory by $\widetilde{\mathbb{T}}_H$. We then interpret this theory by taking the genuine pictures to comprise equivalence classes of hands-free pictures under isomorphism.

What is the relationship between $\mathbb{T}_H$ and $\widetilde{\mathbb{T}}_H$? One might be tempted to ask whether they are translationally equivalent, in the sense defined above. The difficulty is that we developed the criterion of translational equivalence for first-order theories equipped with a normal (rather than hands-free) semantics. So first, there is a question about how to apply the criterion: only invariant sentences receive determinate truth-values under the hands-free semantics, which complicates the notion of entailment used to define translational equivalence. But more significantly, there are good reasons to think that translational equivalence is an inappropriate criterion to use here, given that it is characterised in the first instance in syntactic terms. That is, translational equivalence looks at the syntactic conditions set out by the two theories, and asks if they can be appropriately correlated with one another. In shifting from $\mathbb{T}_H$ to $\widetilde{\mathbb{T}}_H$, however, we effectively leave those syntactic conditions alone, and instead alter the semantic structures used to evaluate them. This makes translational equivalence a poor way of comparing the two theories. Comparing the syntactic conditions of one theory with those of another is only a good means of comparison if the two theories are using the same kind of semantic structures. In particular, in a case such as this (where the syntactic conditions have been left alone), translational equivalence is likely to be "blind" to differences that are not visible at the level of the syntactic conditions.

A more useful way to get insight into their relationship is to consider $\mathrm{Mod}(\mathbb{T}_H)$ and $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ as categories: specifically, to consider the categories obtained by taking the models as objects and the homomorphisms as arrows.[19] This makes very clear the change in the semantics. Moreover, as we shall see in the next chapter, it is a method of analysis that is very general indeed: it can be extended to theories of genuine physical interest.

In this case, having represented their classes of models as categories, there is a very natural criterion of equivalence to apply: that of categorical equivalence.[20] Two categories $\mathcal{C}$ and $\mathcal{D}$ are equivalent if there exist a pair of functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{C}$ such that there exist natural isomorphisms $\varepsilon : FG \Rightarrow \mathrm{Id}_{\mathcal{D}}$ and $\eta : \mathrm{Id}_{\mathcal{C}} \Rightarrow GF$. Heuristically, this means that F and G are "almost inverse": the composition functors

[19] I will assume familiarity with the basic notions of category theory; for an introduction, see e.g. [Awodey, 2010]. For our purposes here, all that is really needed is the fact that a category consists of some class of *objects*, equipped with *arrows* between pairs of objects. The arrows "behave like functions", in the following sense: (i) for any two arrows $f : A \to B$ and $g : B \to C$ there is a third *composition* arrow $g \circ f : A \to C$; (ii) composition is *associative*, so $h \circ (g \circ f) = (h \circ g) \circ f$; and (iii) for any object A, there is a unique *identity* arrow $\mathrm{Id}_A : A \to A$ such that $h \circ \mathrm{Id}_A = h$ and $\mathrm{Id}_A \circ k = k$ (for any $h : A \to B$ and $k : C \to A$).

[20] The idea of seeing whether two theories have (categorically) equivalent categories of models is taken from [Weatherall, 2015c] and [Halvorson and Tsementzis, 2015].

FG and GF do not quite bring objects and arrows back to themselves, but they bring them to an isomorphic object or arrow. It can be shown a functor $F : \mathcal{C} \to \mathcal{D}$ is (one half of) a categorical equivalence beween $\mathcal{C}$ and $\mathcal{D}$ if and only if it is

- Full, i.e., the induced map $f \in \mathcal{C}(A, B) \mapsto Ff \in \mathcal{D}(FA, FB)$ is surjective;[21]
- Faithful, i.e., the induced map $f \in \mathcal{C}(A, B) \mapsto Ff \in \mathcal{D}(FA, FB)$ is injective; and
- Essentially surjective, i.e., for any object X of $\mathcal{D}$, there is some object A of $\mathcal{C}$ such that FA is isomorphic to X.

The relevant observation now is that there is an extremely natural functor $\mathrm{I}^* : \mathrm{Mod}(\mathbb{T}_H) \to \mathrm{Mod}(\widetilde{\mathbb{T}}_H)$. First, for any $\mathcal{M} \in \mathrm{Mod}(\mathbb{T}_H)$, let $\widetilde{\mathcal{M}}$ be the hands-free model defined as follows:

$$\begin{aligned} D_{\widetilde{\mathcal{M}}} &= D_{\mathcal{M}} \\ 2^{\widetilde{\mathcal{M}}} &= \{L, R\} \\ \chi_{\widetilde{\mathcal{M}}}(a) = 1 \text{ iff } L_{\mathcal{M}}(a) &= 1 \end{aligned} \tag{3.16}$$

We then take I^* to be the functor which acts on any object $\mathcal{M}$ in $\mathrm{Mod}(\mathbb{T}_H)$ by $\mathrm{I}^*\mathcal{M} = \widetilde{\mathcal{M}}$, and on any arrow $H : \mathcal{M} \to \mathcal{N}$, by $(\mathrm{I}^*H)_1 = H$ (considered as maps on sets), $(\mathrm{I}^*H)_2(L) = L$ and $(\mathrm{I}^*H)_2(R) = R$. Were $\mathrm{Mod}(\mathbb{T}_H)$ and $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ equivalent as categories, then we would expect I^* to be an equivalence between them. In fact, however, we can show that

Proposition 5. $\mathrm{I}^* : \mathrm{Mod}(\mathbb{T}_H) \to \mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ is not full.

Proof. Let $\mathcal{M}$ be as follows:

$$\begin{aligned} D_{\mathcal{M}} &= \{0, 1, 2\} \\ L^{\mathcal{M}} &= \{0\} \\ R^{\mathcal{M}} &= \{1, 2\} \end{aligned}$$

Since $\mathrm{I}^*\mathcal{M} = \mathrm{I}^*(\mathrm{E}^*\mathcal{M})$, we know that $\mathrm{Id}_{\mathrm{I}^*\mathcal{M}} \in \mathrm{Hom}(\mathrm{I}^*\mathcal{M}, \mathrm{I}^*(\mathrm{E}^*\mathcal{M}))$. If there was some $H : \mathcal{M} \to \mathrm{E}^*\mathcal{M}$ such that $\mathrm{I}^*H = \mathrm{Id}_{\mathrm{I}^*\mathcal{M}}$, then h would need to act as the identity on the underlying set $D_{\mathcal{M}}$. But there is no homomorphism from $\mathcal{M}$ to $\mathrm{E}^*\mathcal{M}$ which does this. So there is no such H; thus, the induced map is not surjective. □

[21] Here, $\mathcal{C}(A, B)$ denotes the set of arrows between A and B in $\mathcal{C}$ (known as the "hom-set" of A, B in $\mathcal{C}$, and alternatively denoted by $\mathrm{Hom}(A, B)$.)

Of course, this does not provide a fully rigorous proof that $\mathrm{Mod}(\mathbb{T}_H)$ and $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ are not equivalent as categories, since we have not ruled out that there is *some* equivalence between them.[22] But it would be very surprising if this were so, given that the functor I^* is such a standout candidate for such an equivalence: if it is not able to do the job, then what could? I therefore conclude, albeit somewhat provisionally, that $\mathrm{Mod}(\mathbb{T}_H)$ and $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ are *not* equivalent as categories, and hence that the literal interpretation of $\mathbb{T}_H$ is not equivalent to the literal interpretation of $\widetilde{\mathbb{T}}_H$; or in other words, that the interpretation of $\mathbb{T}_H$ if we apply the symmetry-interpretation link differs from its interpretation if that link is repudiated.

The theory $\widetilde{\mathbb{T}}_H$ gives one way in which we could make manifest the invariant content of $\mathbb{T}_H$. An alternative way to do so is to seek a theory which traffics only in quantities that are invariant under the symmetry. In this case, we could seek a theory formulated in terms of the *congruence* relation. That is, we start by introducing a relation C that is defined by

$$\forall x \forall y (Cxy \leftrightarrow ((Lx \wedge Ly) \vee (Rx \wedge Ry))) \tag{3.17}$$

Informally, congruence is just the relationship that holds between two objects iff they have the same handedness. Let us use θ_C as a shorthand for the formula (3.17). If we supplement $\mathbb{T}_H$ by this definition, then we get its definitional extension $\mathbb{T}_H^+ := \mathbb{T}_H \cup \{\theta_C\}$, in signature $\{L, R, C\}$. The first observation is that agreement on the congruence relation suffices for agreement on all invariant content, in the following sense: if $\mathcal{M}$ and $\mathcal{N}$ are two models of $\mathbb{T}_H^+$, such that $|\mathcal{M}| = |\mathcal{N}|$ and $C^{\mathcal{M}} = C^{\mathcal{N}}$, then either $\mathcal{M} = \mathcal{N}$, or else $\mathcal{M} = \mathrm{E}^*\mathcal{N}$.

Now consider the theory $\mathbb{T}_C$, in signature $\Sigma_C := \{C\}$, comprised by the following axioms:

$$\forall x Cxx \tag{3.18a}$$

$$\forall x \forall y (Cxy \rightarrow Cyx) \tag{3.18b}$$

$$\forall x \forall y \forall z ((Cxy \wedge Cyz) \rightarrow Cxz) \tag{3.18c}$$

$$\forall x \forall y \forall z ((\neg Cxy \wedge \neg Cyz) \rightarrow Cxz) \tag{3.18d}$$

Informally, this theory states that C is an equivalence relation, with at most two equiv-

[22] Here is a means by which one could seek to demonstrate it (suggested by Teruji Thomas): by showing that neither $\mathrm{Mod}(\mathbb{T}_C)$ nor $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ have a terminal object. Since $\mathrm{Mod}(\mathbb{T}_H)$ does have a terminal object (a/the model containing exactly one element that is L and exactly one that is R), doing so would suffice to show that the categories are inequivalent. This strikes me as a good proof-method, but one which I lack the categorical expertise to execute.

alence classes.

Models of $\mathbb{T}_C$ closely correspond to models of $\mathbb{T}_H^+$ (and hence, to models of $\mathbb{T}_H$). On the one hand, for any model $\mathcal{M}$ of $\mathbb{T}_H^+$, its reduct $\mathcal{M}|_{\Sigma_C}$ is a model of $\mathbb{T}_C$. Indeed, suppose that $\mathcal{M} \models \mathbb{T}_H^+$; then $\mathcal{M}$ satisfies the sentences (3.8) and (3.17); but the sentences (3.18) of $\mathbb{T}_C$ are simply a consequence of those sentences, and so $\mathcal{M}$ must make (3.18) true as well; since these refer only to C, it follows that $\mathcal{M}|_{\Sigma_C} \models \mathbb{T}_C$. On the other hand, for any model $\mathcal{N}$ of $\mathbb{T}_C$, there is a Σ_H^+-expansion $\mathcal{N}^+$ of $\mathcal{M}$ (i.e., a Σ_H^+-picture $\mathcal{N}^+$ such that $\mathcal{N}^+|_{\Sigma_C} = \mathcal{N}$) which is a model of $\mathbb{T}_H^+$. Indeed, if $\mathcal{N}$ is a model of $\mathbb{T}_C$, then it is clear from equations (3.18a)–(3.18c) that $C^{\mathcal{N}}$ is an equivalence relation over $D_{\mathcal{N}}$, and from (3.18d) that it partitions the domain into at most two equivalence classes. So just let $L^{\mathcal{N}^+}$ be one of these equivalence classes, and let $R^{\mathcal{N}^+}$ be the other (if such there be). It is then obvious that $\mathcal{N}^+$ satisfies (3.8) and (3.17), i.e. that $\mathcal{N}^+ \models \mathbb{T}_H^+$. Thus, there is a natural sense in which $\mathbb{T}_C$ captures the "invariant part" of $\mathbb{T}_H$. On the one hand, any models of $\mathbb{T}_H$ which agree with respect to all the structure invariant under E^* will correspond to a single model of $\mathbb{T}_C$; and on the other, every model of $\mathbb{T}_C$ corresponds to some (indeed, more than one) model of $\mathbb{T}_H$.

That said, it is *not* the case that $\mathbb{T}_C$ captures more than the invariant part of $\mathbb{T}_H$. One argument for this conclusion is the fact that $\mathbb{T}_C$ and $\mathbb{T}_H$ are not translationally equivalent. For, the only plausible translation from $\mathbb{T}_H$ to $\mathbb{T}_C$ would be the syntactic map

$$\mathrm{F}(C) = ((Lx \wedge Ly) \vee (Rx \wedge Ry)) \tag{3.19}$$

But, as is easily seen, F^* is not a bijection.

However, we can also prove a different result, in analogy to what we saw above: the two theories do not have equivalent categories of models. Again, the only plausible candidate for a functor between these two categories is F^* considered as a functor: i.e., the functor which acts on models by $\mathcal{M} \mapsto \mathrm{F}^*\mathcal{M}$, and on homomorphisms as $h \mapsto h$ (this prescription works because $D_{\mathrm{F}^*\mathcal{M}} = D_{\mathcal{M}}$). This is not an equivalence of categories: more specifically,

Proposition 6. $\mathrm{F}^* : \mathrm{Mod}(\mathbb{T}_H) \to \mathrm{Mod}(\mathbb{T}_C)$ is not full.

Proof. Let $\mathcal{M}$ be as as in the proof of Proposition 5. As was the case there, since $\mathrm{F}^*\mathcal{M} = \mathrm{F}^*(\mathrm{E}^*\mathcal{M})$, we know that $\mathrm{Id}_{\mathrm{F}^*\mathcal{M}} \in \mathrm{Hom}(\mathrm{F}^*\mathcal{M}, \mathrm{F}^*(\mathrm{E}^*\mathcal{M}))$. If there was some $H : \mathcal{M} \to \mathrm{E}^*\mathcal{M}$ such that $\mathrm{F}^*H = \mathrm{Id}_{\mathrm{F}^*\mathcal{M}}$, then H would need to act as the identity on the underlying set $D_{\mathcal{M}}$. But there is no homomorphism from $\mathcal{M}$ to $\mathrm{E}^*\mathcal{M}$ which does this. So there is no such H; thus, the induced map is not surjective. □

Again, this proof is not fully decisive; but given the implausibility of constructing any equivalence from $\mathrm{Mod}(\mathbb{T}_H)$ to $\mathrm{Mod}(\mathbb{T}_C)$ other than F^*, we can at least be reasonably confident that these are not equivalent categories. So the interpretational commitments of $\mathbb{T}_H$ and $\mathbb{T}_C$ (under their literal interpretations) are distinct.

The theory $\mathbb{T}_C$ represents an alternative way in which one can seek to encode the implications of the lesson about symmetries: by constructing a theory whose models are *invariant* under the action of the symmetry. In such a theory, there is an exact, one-to-one correspondence between the putative models and the genuine models (i.e., the models obtained by applying the interpretative principle discussed above). This contrasts with $\widetilde{\mathbb{T}}_H$, in which there was a multiplicity of putative models for each genuine model; however, in $\widetilde{\mathbb{T}}_H$ all the putative models corresponding to one genuine model were isomorphic to one another. I remarked above that categorical equivalence only requires a pair of functors which are (in rough terms) "inverse up to isomorphism". So it should not be too surprising to learn that

Proposition 7. $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ and $\mathrm{Mod}(\mathbb{T}_C)$ are equivalent as categories.

Proof. We can regard the dictionary map F as inducing a functor from $\mathrm{Mod}(\widetilde{\mathbb{T}}_H) \to \mathrm{Mod}(\mathbb{T}_C)$; just to maintain notational hygiene, call this functor $\mathrm{F}^\dagger$. Explicitly, for any $m \in \mathrm{Mod}(\widetilde{\mathbb{T}}_H)$, let $\mathrm{F}^\dagger m$ be the Σ_C-picture such that

- $D_{\mathrm{F}^\dagger m} = D_m$
- For any $a, b \in D_{\mathrm{F}^\dagger m}$, $C_{\mathrm{F}^\dagger m}(a, b) = 1$ iff $\chi_m(a) = \chi_m(b)$

For any $h : m \to n$, let $\mathrm{F}^\dagger h$ be the map $H : D_{\mathrm{F}^\dagger m} \to D_{\mathrm{F}^\dagger n}$ such that $H = h_1$. It is straightforward to verify that $\mathrm{F}^\dagger m \in \mathrm{Mod}(\mathbb{T}_C)$, and that $\mathrm{F}^\dagger h$ is a Σ_C-homomorphism; that is, that $\mathrm{F}^\dagger$ really is a functor. We now show that $\mathrm{F}^\dagger$ is full, faithful, and essentially surjective.

First, consider any $m, n \in \mathrm{Mod}(\widetilde{\mathbb{T}}_H)$, and let H be any homomorphism from $\mathrm{F}^\dagger m$ to $\mathrm{F}^\dagger n$. Now define $h_1 : D_m \to D_n$ by the condition that $h_1 = H$ (as a map on sets). Then, letting a be some arbitrary element of D_m, define $h_2 : \mathbf{2}^m \to \mathbf{2}^n$ as the (unique) bijection such that $h_2(\chi_m(a)) = \chi_n(h_1(a))$; it is easily seen that this uniquely determines h_2, and that it does so independently of the choice of a. Moreover, it is clear that $h := (h_1, h_2)$ is a homomorphism $m \to n$, and that $H = \mathrm{F}^\dagger h$. So $\mathrm{F}^\dagger$ induces a surjective map on morphisms between any m and n, i.e. $\mathrm{F}^\dagger$ is full.

Second, for any $m, n \in \mathrm{Mod}(\widetilde{\mathbb{T}}_H)$, consider any $h, h' : m \to n$ such that $\mathrm{F}^\dagger h = \mathrm{F}^\dagger h'$. Clearly, $h_1 = h'_1$. Furthermore, since h and h' are homomorphisms, it follows that for

any $a \in D_m$, $h_2(\chi_m(a)) = \chi_n(h_1(a)) = \chi_n(h_1'(a)) = h_2'(\chi_m(a))$; hence, $h_2' = h_2$. So $h = h'$. So $\mathrm{F}^\dagger$ induces an injective map on morphisms between any m and n, i.e., $\mathrm{F}^\dagger$ is faithful.

Finally, let $\mathcal{M}$ by any model of $\mathbb{T}_H$. Define a hands-free picture m by picking some arbitrary $a \in D_\mathcal{M}$, then setting

- $D_m = D_\mathcal{M}$
- $\mathbf{2}^m = \{\varnothing, a\}$
- For any $b \in D_\mathcal{M}$, $\chi_m(b) = a$ iff $C_\mathcal{M}(a, b) = 1$, and $\varnothing$ otherwise

Clearly, $\mathrm{F}^\dagger m = M$. So $\mathrm{F}^\dagger$ is surjective, and therefore essentially surjective. □

Therefore, we have the following results. First, $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ and $\mathrm{Mod}(\mathbb{T}_C)$ are equivalent as categories; second, although we don't have a demonstration that $\mathrm{Mod}(\mathbb{T}_H)$ is inequivalent to either $\mathrm{Mod}(\widetilde{\mathbb{T}}_H)$ or $\mathrm{Mod}(\mathbb{T}_C)$ (since we have not ruled out there is *some* appropriate functor between them), we have at least shown that the obvious functors will not do the job. Thus, I make the following claim: $\mathbb{T}_C$ or $\widetilde{\mathbb{T}}_H$, interpreted using their putative semantics, are not equivalent to $\mathbb{T}_H$ interpreted using its putative semantics. However, they are equivalent to $\mathbb{T}_H$ interpreted by application of the symmetry-interpretation link. Any of these interpreted theories are in a position to avoid the triple challenge (indeterminacy of reference, insoluble ignorance, and indeterminism) that plague a symmetry-containing theory such as $\mathbb{T}_H$ under its literal interptation; and hence, are better theories to subscribe to.

4 Internal symmetries

> But does it matter if the place cannot be mapped, as long as I can still describe it?
>
> Jeanette Winterson, *Sexing the Cherry*

In this chapter, I apply the general metaphysical framework developed in chapters 2 and 3 to analysing symmetries in physics—and specifically, to the project of developing an account of the symmetry-interpretation link.

4.1 Internal transformations

The theories we consider have the following feature in common: their models consist of certain structures on fibre bundles.[1] More specifically, we will suppose that any such theory lays down a certain kind of *background base space* structure, and a certain kind of *bundle structure*. This defines a class of *background bundles* for the theory, namely, those fibre bundles which manifest the kind of structure desired by the theory. In general, I will denote a generic bundle by $E \xrightarrow{\pi} M$, where E is the total space, M is the base space, and π is the projection map.

Each model of the theory then consists of certain kinds of data on such a bundle, or on a class of associated bundles: such data may include sections of the bundles, connections on the bundles, or constructions on the base space of the bundle (e.g. curves through the base space or tensor fields on the base space). We will take the specification of such data to include syntactic provisions that specify (e.g.) what variables, or what kinds of variables, range over data of a certain kind. These syntactic provisions serve to bring about a specific formal language (namely, the language of differential geometry equipped with the appropriate variables). The sense in which

[1] See [Baez and Muniain, 1994], [Healey, 2007], or [Weatherall, 2015b] for an introduction to the fibre bundle formalism.

this is a formal language is simply that it consists of a circumscribed set of symbols, whose grammatically permissible combinations may be exactly specified (so that e.g. $\partial_\mu \phi = K^\mu$ could be a permissible sentence, but $a^\nu =^\sigma \partial$ typically would not).

Several of our examples will—at least at first—take the form of a product bundle: that is, a bundle of the form

$$F \times M \stackrel{\pi_M}{\to} M \tag{4.1}$$

where F is the standard fibre, M is the base space, and π_M is the projection map onto the second factor. Such bundles are especially simple and easy to work with. They come equipped with a natural flat connection (that according to which parallel-transporting $(f, p) \in F \times M$ to $q \in M$ simply yields (f, q)); and any section of such a bundle may be identified with a function $s : M \to F$.

To count as a model, the dynamical data must satisfy some appropriate set of differential equations, couched in the associated formal language. In general, I'll refer to the differential equations as (or as encoding) the *dynamics*; everything prior to the specification of the dynamics (what the background bundles are, and what kinds of data are needed for a model) will be referred to as the *kinematics*. In particular, a background bundle equipped with the right kind of data (but which may or may not satisfy the dynamical equations) will be referred to as a *kinematically possible model*; if the data does satisfy those equations, then we have a *dynamically possible model*.

Given a bundle $E \stackrel{\pi}{\to} M$, a *bundle automorphism* is a pair of diffeomorphisms $\alpha : E \to E$ and $\beta : M \to M$, such that $\pi \circ \alpha = \beta \circ \pi$; i.e., such that the diagram

$$\begin{array}{ccc} E & \xrightarrow{\alpha} & E \\ \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} \\ M & \xrightarrow{\beta} & M \end{array} \tag{4.2}$$

commutes. A *vertical* bundle automorphism is a bundle automorphism of the form (α, Id_M). Note that it is required only to be an automorphism on the bundle as a (mere) fibre bundle; if the bundle carries extra structure (e.g. if it is a vector bundle, or even a trivial product bundle) then that structure need not be preserved.

In the case of a product bundle, a bundle automorphism may be specified as a pair of diffeomorphisms $\beta : M \to M$ and $\gamma : F \times M \to F$: this determines a bundle automorphism in the above sense by taking $\alpha(f, p) = (\gamma(f, \beta(p)), \beta(p))$, and any bundle automorphism may be specified in this fashion. In other words, for product bundles any bundle automorphism may be decomposed into a vertical bundle automorphism

γ and a diffeomorphism β of the base space. The vertical bundle automorphism may be more perspicuously presented as a (smoothly cohering) family $\{\gamma_p : F \to F\}_{p \in M}$ of maps: that is, as a collection of fibrewise maps, allowed to vary from one point of the base space to another.

Given a kinematically possible model $\mathcal{S}$, an *internal transformation* of $\mathcal{S}$ is a transformation of $\mathcal{S}$ which is generated by a vertical bundle automorphism of the background bundle of $\mathcal{S}$: that is, an internal transformation takes $\mathcal{S}$ to $\alpha\mathcal{S}$, where $\alpha\mathcal{S}$ is obtained by pulling back the dynamical data on $\mathcal{S}$ by the bundle automorphism (α, Id_M). In general, applying an internal transformation to a dynamically possible model will not yield another dynamically possible model. If, however, an internal transformation maps all and only dynamically possible models to other dynamically possible models, then we say that it is an *internal symmetry*.

All this is rather abstract; let us consider some examples. The toy Newtonian theory discussed in Chapter 1 has this form. The background bundle is the product space $X \times T$, regarded as a (trivial) bundle of X over T, equipped with a scalar field V on the total space $T \times X$. The dynamical data is just a section of this bundle, better-known as a function from T to X. The only differential equation is the equation (1.1): sections satisfying (1.1) (i.e., in the case $V = 0$, those of the form (1.2)) are the models of the theory. This example is somewhat unusual, in that (physical) space is represented by the fibres, rather than by the base space, but it has the same formal character nevertheless.[2] We will denote this theory by $\mathbb{T}_N$. Of the symmetries of this theory that were discussed in Chapter 1, those of the form

$$x \mapsto x + vt + d \tag{4.3}$$

are internal symmetries.

For a second example, we will use the theory $\mathbb{T}_\phi$ of *instantaneous electrostatics*. This is also very much a toy theory: as the name suggests, it is just the portion of electrostatics concerning how things can stand at a specific time, without concern for how things evolve over time. Let $\mathfrak{X}$ be an affine space whose associated vector space, $V_\mathfrak{X}$, is equipped with a Euclidean inner product. The background bundle of this theory is then $\mathbb{R} \times \mathfrak{X}$, regarded as a bundle of $\mathbb{R}$ over $\mathfrak{X}$. The dynamical data for each kinematically possible model is a section ϕ of the background bundle, and a scalar field $\rho : \mathfrak{X} \to \mathbb{R}$. (There isn't an especially principled reason to regard ρ as a scalar field on the base space,

[2] cf. [Belot, 2007, Remark 12]

rather than enlarging the background bundle to $\mathbb{R}^2 \times \mathfrak{X}$ and taking (ρ, ϕ) as a section of it; it's just a bit more elegant.) Again, the dynamics are given by a single equation:

$$\nabla^2 \phi = -\rho \tag{4.4}$$

where ∇ is the affine derivative on $\mathfrak{X}$. In this theory, transformations of the form

$$\phi \mapsto \phi + k \tag{4.5}$$

where k is some real number, are internal symmetries.

We are now in a position to distinguish two kinds of transformation for theories in which the background bundle is a product bundle—that is, where it is of the form $F \times M$, for standard fibre F and base space M. As remarked earlier, any section of such a bundle may equally well be thought of as a function $s : M \to F$. Correlatively, we can think of such a section as implementing, in a reasonably straightforward fashion, the metaphysical concept of a *universal*: a property which distinct individuals can have in common.[3] The standard fibre F represents a determinable universal, with points in F corresponding to different determinates of it. For such a theory, we can distinguish two kinds of (internal) transformation one might apply. A *constant* transformation is one in which the transformation is the same at every point: it can be specified simply by a map of type $F \to F$. The transformation (4.5) is of this kind. Alternatively, a *variable* transformation is one in which we apply different transformations at different points of M. It must be specified by a map of type $M \times F \to F$. This was the case for the transformation (4.3).

Our third example will be that of *Maxwell electromagnetism in terms of potentials*, which theory will be denoted by $\mathbb{T}_A$. This time, let $\mathfrak{T}$ be a one-dimensional oriented affine space equipped with a Euclidean metric, and define *Newtonian spacetime* $\mathfrak{N} := \mathfrak{T} \times \mathfrak{X}$ (where $\mathfrak{X}$ is as above). We take the background bundle to be $T^*\mathfrak{N}$, the cotangent bundle over $\mathfrak{N}$. The dynamical data for any model consists of a section A_a of $T^*\mathfrak{N}$, and a vector field J^a over $\mathfrak{N}$. Since $\mathfrak{N} = \mathfrak{T} \times \mathfrak{X}$, we can decompose these as $A_a = (\phi, \mathbf{A})$ and

[3] cf. [Maudlin, 2007c].

$J^a = (\rho, \mathbf{J})$. The dynamics is given by the following equations:

$$\nabla^2\phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) = -\rho \tag{4.6a}$$

$$\frac{\partial^2 \mathbf{A}}{\partial t^2} - \nabla^2\mathbf{A} + \nabla\left(\nabla \cdot \mathbf{A} + \frac{\partial \phi}{\partial t}\right) = \mathbf{J} \tag{4.6b}$$

It should be emphasised that this theory is highly unphysical: it merely describes how the fields behave in the presence of a background charge current J^a, without giving any dynamics for the evolution of J^a itself. Indeed, it has no dynamics for the motion of matter at all (i.e., it has no equation of motion). A more adequate theory would include the Lorentz force law, indicating how charged matter ought to move in response to the electromagnetic fields. A properly adequate theory would model the charge current J^a as such matter, by including (say) a dust flow field obeying the Lorentz force law and coupled to J^a via continuity equations. However, such a theory would be rather more complicated; in the interests of simplicity, we will therefore stick with the basic theory outlined above—although we will see below some of the consequences of so doing. In $\mathbb{T}_A$, any transformation

$$A_a \mapsto A_a + \nabla_a\lambda \tag{4.7}$$

where $\lambda : \mathfrak{N} \to \mathbb{R}$ is some smooth scalar function, is an internal symmetry.

4.2 Interpreting internal symmetries

We have now defined what it is for a theory to manifest an internal symmetry. The question before us is then: what should we say about models related by an internal symmetry? In particular, should we take models related by such a symmetry to represent the same possible world or not? In order to answer this question, we need to say a little more about the analogy between theories specified in terms of first-order logic and theories framed (as in the previous section) in terms of fibre bundles. First, suppose that we are only thinking about *product* bundles, as is the case for the theories $\mathbb{T}_N$ and $\mathbb{T}_\phi$. (Indeed, given that $\mathfrak{N}$ is an affine space, the cotangent bundle $T^*\mathfrak{N}$ may be characterised as a product bundle as well: there is a canonical means of identifying $T^*_p\mathfrak{N}$ and $T^*_q\mathfrak{N}$, for any $p, q \in \mathfrak{N}$.) A section s of a product bundle $F \times M$ can equally well be thought of as a function from the base space M to the standard fibre F. As I noted in Chapter 1, this sets up a certain level of analogy to a Σ-picture; we can now

go into a little more detail about what that analogy looks like.

The base space plays the role of the base set: it offers a collection of individuals, of which we will be predicating attributes. (As mentioned in Chapter 1, the base space of a bundle carries much more structure than a mere set. We will see the consequences of this in the next chapter.) The fibre represents the range of possible ways for a member of the base set to be. We may think of it as representing a determinable property, with each point in the fibre corresponding to a single determinate of that determinable; the function s indicates what particular determinate each member of the base space has. In this sense, the fibre F plays a similar role to the set Ξ of properties that we discussed in §3.1, which similarly served to abstractly represent the range of possible ways for individuals to be. The analogy still requires a certain amount of judicious squinting, given that (for instance) a given individual will be assigned *multiple* elements of Ξ but only a single element of F. But let us see if it can bring us insight nevertheless.

Suppose first that the internal symmetry transformation in question is a constant transformation. The symmetry (4.5) is an example of such a transformation. In such a case, the relationship to translational equivalence and anti-quidditism is straightforward. A syntactic map (from a signature to itself), after all, is just a map from Ξ to itself; so a map $\eta : F \to F$ may be regarded as doing the same thing. In particular, the effect of such a map on a model is simply that any point $p \in M$ goes from being represented as having the property $f \in F$, to being represented as having the property $\eta(f)$. As such, the analysis which was carried out in Chapter 3 carries over in a reasonably straightforward fashion: f and $\eta(f)$ represent properties with identical nomological roles, and so we should not believe that the models $\mathcal{M}$ and $\eta^*\mathcal{M}$ represent distinct possible worlds. If we did, then we would face the semantic and epistemic problems canvassed in the last chapter.

The case for a variable transformation, or a vertical transformation of a non-trivial bundle, is a little more subtle. The relevant observation here is that the fibre above any given point represents a space of property-values *for that point*.[4] That is, we think of field theories as involving the assignment of properties to spacetime points. It is just that the properties come partitioned or indexed into distinct sets by the members of M: the points of $F_p := \pi^{-1}(p)$ represent the properties which p can bear. It is important to note that these properties are usually represented as having, as it were, "trans-individual" structure: that is, there are usually systematic relationships between the points of F_p and those of F_q. How much structure depends on the nature of the bundle,

[4] Again, cf. [Maudlin, 2007c]; see also [Arntzenius, 2012, chap. 6].

of course. For a product bundle, in which $F_p = \{p\} \times F$ and $F_q = \{q\} \times F$, there is the relationship of *corresponding to the same point of* F, borne by (p, f) to (q, f) for any $f \in F$. And since F will typically itself carry certain kinds of structure (e.g. that of a vector space), there are relationships like "having twice the F-value", borne by (p, f) to $(q, 2f)$. For more austere bundles, the relationships will be rather more thin on the ground: if the bundle has a connection, say, then it could be that $u \in F_q$ is the parallel transport of $w \in F_p$ along some specific path from p to q. And in more austere settings yet, the only relationships may be topological in character: that this set of points in the bundle form a continuous path, for instance. Note that I say only that the properties are *prima facie* represented as bearing these kinds of relationships to one another, not that the bundles *does* represent them thus. Whether it does so or not will depend, as ever, on what kind of interpretation we give to the formalism.

So when we apply an internal transformation to a model, we systematically alter what properties are borne by the various spacetime points. However, since we are considering a way in which the property-values can be mapped to one another whilst preserving dynamical possibility, this still preserves nomological structure. Whatever nomological role the points $u_1, \ldots, u_n \in E$ play, the fact that α is a symmetry means that the points $\alpha(u_1), \ldots, \alpha(u_n)$ play the same role; correlatively, even for variable transformations, the relationship between a pair of fibre-bundle models $\mathcal{M}$ and $\alpha^*\mathcal{M}$ is the same as the relationship between a pair of first-order models $\mathcal{M}$ and $\mathrm{D}^*\mathcal{M}$. In particular, insofar as we take our practices as speakers and knowers to be responsive only to features of natural laws, and not to extra-nomological "naturalness" structure, we obtain the same result as in Chapter 3: such a pair of models will be semantically and epistemically indistinguishable, and so we have good reason to apply the symmetry-interpretation link.

Are there problems with taking such a stance? That is, are there cases where adopting this attitude leads to implausible physical consequences? I do not believe so. We have to be careful about the deployment of such counterexamples, since we are working with highly idealised theories whose physical content is not always immediate. For example, [Belot, 2013] observes that any set of linear homogeneous partial differential equations has an extremely large symmetry group. For if u_1 and u_2 are solutions, then so is $u_1 + u_2$; hence, given any pair u_1 and u_2, they are related by the symmetry operation of adding $u_2 - u_1$. But the source-free Maxwell equations (i.e. the equations (4.6) with $J^a = 0$) are linear homogenous differential equations. So the symmetry-interpretation link enjoins us to not merely identify gauge-equivalent solutions, but to identify *all*

solutions. But that seems extremely implausible, as Belot observes:

> under any ordinary reading, [these equations] admit solutions that represent situations in which nothing is happening ([...] the field is in a ground state) and others that represent situations in which plenty is going on ([...] energy in the form of heat or waves is propagating). An approach to understanding physical theories that leaves us unable to see these distinctions is not something we can live with. So [the symmetry-interpretation link] is false if understood as a thesis concerning classical symmetries of differential equations.[5]

The problem, however, is that Belot is appealing to physical intuitions about a very unphysical theory. As soon as we add a little more physics, the problematic symmetry disappears: specifically, the symmetry is broken (in this example) by the introduction of test particles, coupled to the electromagnetic field by the Lorentz force law. That is, consider any solution $(\phi, \mathbf{A})$ to the homogeneous Maxwell equations. Now add to that model the trajectories $\mathbf{x}_i : \mathfrak{T} \to \mathfrak{X}$ of some particles, each of which has charge q_i and mass m_i, such that each particle satisfies the following equation at all times:

$$m_i \ddot{\mathbf{x}}_i = q_i \left(-\nabla(\phi - \dot{\mathbf{x}}_i) - \frac{d\mathbf{A}}{dt} \right) \tag{4.8}$$

If we were to superimpose an arbitrary second solution $(\mathbf{A}', \phi')$ on to this model, then we would still have a solution to the homogeneous Maxwell equations—but we would *not* typically have a solution to (4.8). (For carefully chosen $(\mathbf{A}', \phi')$ we might; but in that case, there would be other distributions of particle trajectories that would block the symmetry.) Hence, it is only by representing the electromagnetic field in abstraction from all matter that we obtain the result that all solutions are equivalent; and I, for one, have very few robust intuitions about what interpretations of a theory such as that are physically reasonable.

4.3 Sophistication

I turn now to the issue of how the symmetry-interpretation link is to be applied to these theories. As was the case in the handedness theory, to treat a pair of models as equivalent means regarding them as isomorphic; equivalently, one renounces commitment to

[5] [Belot, 2013, p. 330]

whatever structure is blocking that isomorphism. So implementing the link means that we are to interpret these theories by seeking semantics such that symmetry-related models are isomorphic. In the handedness theory, the hands-free theory $\widetilde{\mathbb{T}}_H$ met this job description; let us look at similar constructions for the theories $\mathbb{T}_\phi$ and $\mathbb{T}_A$.

First, consider the electrostatic theory. We retain the same set of equations, but change what objects are used to semantically interpret those equations. Rather than taking ϕ to range over $\mathbb{R}$, we instead take it to range over Φ, where Φ is a one-dimensional, oriented, metric affine space (such a space could be defined as a set equipped with a free, transitive of $\mathbb{R}$ as an additive group). Φ has sufficient structure to enable $\nabla^2\phi$ to be straightforwardly defined. We can therefore continue to use the equation (4.4), interpreted as equations governing models of this kind rather than the original kind. The transformation (4.5) also still makes sense, but is now an automorphism of Φ. As a result, if two Φ-valued fields are related by the application of such a transformation, they are isomorphic to one another.[6] Moreover, note that it's not just that the symmetry transformations of the form (4.5) are automorphisms of Φ: *every* automorphism of Φ is a transformation of the form (4.5). Let us denote this theory (with background bundle $\Phi \times \mathfrak{X}$ and dynamics (4.4)) by $\widetilde{\mathbb{T}}_\phi$.

Second, consider the electromagnetic theory. This time, models of the theory are to be connections on a principal $U(1)$-bundle over $\mathbb{R}^4$. Once more, we retain the equations (4.6), but now interpreted in a way that makes use only of the more minimalist structure available in the models: A_A is now interpreted as the vector potential of the target connection relative to some *arbitrarily chosen* flat connection on the principal bundle; it is straightforward to show that any two such flat connections will be related by a gauge transformation (a vertical automorphism of the bundle), and hence that it doesn't matter which flat connection we choose as a reference-point. And, since gauge transformations are vertical automorphisms of the bundle, the action of (4.7) on the target connection will yield a model isomorphic to the original. Let us denote this theory by $\widetilde{\mathbb{T}}_A$.

Hopefully, these examples make clear enough what is intended; let us now seek a general characterisation. Note that the proposal on the table—that we can do justice to a symmetry by seeking models on which the symmetry acts as isomorphism—is analogous to the "sophisticated substantivalist" method for dealing with spacetime

[6] Given two functions $f : U \to V$ and $f' : U' \to V'$, the appropriate definition of morphism is as follows: a pair of morphisms $\alpha : U \to U'$ and $\beta : V \to V'$ such that $\beta \circ f = f' \circ \alpha$. An isomorphism is then an invertible morphism.

symmetries.[7] With that in mind, let us refer to theories equipped with semantics of this sort as *sophisticated* theories.

In general, suppose that we have some theory $\mathbb{T}$, subject to a group G of symmetries; for any model $\mathcal{M}$ of $\mathbb{T}$ and symmetry action $g \in G$, let $g^*\mathcal{M}$ denote the result of acting on $\mathcal{M}$ by g. Then a *sophistication by* G of $\mathbb{T}$'s semantics is the "forgetting" of the G-variant structure (but *only* the G-variant structure) from each picture in $\mathbb{T}$'s original semantics, thereby obtaining a semantics which is adequate to assign truth-values to the G-invariant sentences of $\mathbb{T}$'s language—and which has the feature that if F is the forgetful map from the unsophisticated to the sophisticated pictures, then for any unsophisticated models $\mathcal{M}$ and $\mathcal{M}'$, $\mathcal{M}' = g^*\mathcal{M}$ (for some $g \in G$) iff $F(\mathcal{M}) \cong F(\mathcal{M}')$. Since all the syntactic conditions of $\mathbb{T}$ (sentences, differential equations, or what have you) will perforce be invariant under G, we obtain a well-defined "sophisticated theory" $\widetilde{\mathbb{T}}$: that comprising the same syntactic conditions, but applied to the sophisticated semantics.

However, this remains somewhat vague. Is there a way to precisify what is meant? Here is one way to do so. Rather than trying to define the objects of the new semantics "internally", as mathematical structures of such-and-such a kind (paradigmatically, as sets equipped with certain relations or operations), we instead define them "externally": as mathematical structures of a given kind, but with certain operations *stipulated* to be homomorphisms (even if they're not "really" homomorphisms of the given kind). For example, one way to define vector spaces is to define them as sets equipped with operations of addition and scalar multiplication, obeying appropriate axioms. This is the internal method. The alternative is to define them as spaces of the form $\mathbb{R}^k$, with the further feature that linear transformations are declared to be homomorphisms—and in particular, that invertible linear transformations are isomorphisms. This is the external method. It would also be apposite to refer to the internal method as a "synthetic" approach, and the external method as an "analytic" approach, following the terminology of synthetic and analytic geometry. Alternatively, one could see the external method as following in the tradition of Klein's Erlangen program for geometry, and the internal method as falling more under the Riemannian tradition.[8]

Hence, the proposal is that the pictures on the new semantics are simply what we obtain by taking the old objects, and *declaring*, by fiat, that the symmetry transformations

[7] [Pooley, 2006, Pooley, 2013b]. We will consider sophisticated substantivalism in more detail in the next chapter.

[8] See [Wallace, MS] for a detailed defence of using the external method for defining spacetime geometry, and for an expansion on the connection to Klein and Riemann.

are now going to "count" as isomorphisms.[9] If we consider our examples above, we can see that the method for introducing the new semantics for the handedness theory was very much in this vein: the introduction of hands-free models was essentially just a means of legitimating the new definition of homomorphism. The advantage of defining the new semantics externally is that it offers a relatively easy means of characterising the objects of the semantics, and of the means by which they accord truth-values to sentences of the formal language: simply (as we saw for the handedness case) use the old semantics, then construct a supervaluationist semantics over the members of each equivalence class of isomorphic new objects. So defined, it will certainly meet the conditions required to be a sophistication.

The main disadvantage of this method is that it might seem far *too* easy. In general, the external method of defining some kind of mathematical structure might be thought to offer less insight into the nature of that structure: it is one thing to know that a vector space consists of precisely those features of $\mathbb{R}^k$ which are invariant under linear transformations, but another to see that those features are exactly the operations of addition and scalar multiplication, as codified by the axioms for a vector space. More ecumenically, one might think merely that both kinds of construction are important for fully understanding the structure—in which case, one would desire an internal construction as well. And it is often very opaque what kind of internal construction will correspond to an external construction. Electromagnetism makes this fairly clear: it is not at all obvious (I contend) that the features of maps $\mathbb{R}^4 \to \mathbb{R}^4$ preserved under gauge transformations (4.7) are precisely the features of vector potentials between connections on a $U(1)$ principal bundle. Nevertheless, we could reason as follows. Assuming that one accepts the external method of definition as mathematically legitimate,[10] then its application gives us a way of defining a sophisticated semantics for the theory, by brute force. It then means that we do have a precise target for a sophisticated semantics which is internally defined: we are looking for some internal construction which delivers an equivalent class of structures.[11]

Sophistication raises some questions. The major issue is simply whether it really does succeed in implementing the idea that we should get rid of "surplus" (i.e., symmetry-variant) structure. After all (someone might say) surely the ontology postulated by

[9] In category-theoretic terms, this amounts to introducing arrows into the category of models corresponding exactly to the symmetry transformations—which is precisely what [Weatherall, 2015a] proposes to do for (gauge) symmetry transformations. I expand upon the relation to Weatherall's proposal in §§ 4.3–4.5 below.

[10] Which, to be clear, is in accord with standard mathematical practice.

[11] "Equivalent" here meaning that they are isomorphic (not just equivalent) as categories.

the sophisticated version is mostly the same as that of the original theory: a pair of properties in the handedness case, an electrical potential in the electrostatic case, and a vector potential (up to arbitrary choice of reference connection) in the electromagnetic case? So how on earth could it be the case that the sophisticated theory is more parsimonious than the original, in the manner required by the symmetry-interpretation link?

There are two components to the answer: one mathematical, and one more metaphysical. The mathematical observation is that the standard way to explicate the idea of mathematical structure is via isomorphism: what it is for a pair of mathematical objects to have the same structure is for them to be isomorphic to one another.[12] Thus, insofar as we want to defend sophistication's credentials as genuinely "expurgating structure", we can invoke standard mathematical usage in support. This doesn't mean that there is no alternative construal of "structure" that would not be so kind to the sophisticate; but the burden is on the opponent of sophistication to explain what that would be, and to justify their departure from its accepted mathematical meaning.

The metaphysical answer is to recognise that the structure we have abjured is not structure *within* a world, but rather structure across worlds. As I've stressed, the relevant metaphysical move here is anti-quidditism: renouncing primitive identities for properties, above and beyond their identification by nomological role.[13] Applying a sophisticated semantics as the genuine semantics for a theory is simply a way of enacting that, by recognising permutations of properties as isomorphisms (provided that those permutations preserve nomological role). Again, as with the simple example of the handedness theory, this should not be understood as the claim that symmetric properties are to be identified with one another *within* any given world.

Both of these ideas can be illustrated by the following observation: the natural mapping from the category of models of the unsophisticated theory, to the category of models of the sophisticated theory, is not typically a categorical equivalence. In the previous chapter, we saw that this was the case for the handedness theory, but the observation generalises.

For instance, consider the electrostatic theory. The category of models of the unsophisticated theory $\mathbb{T}_\phi$ has models of $\mathbb{T}_\phi$ as objects. A morphism between two such objects $\mathcal{M}$ and $\mathcal{N}$ is an isometry $h : \mathfrak{X} \to \mathfrak{X}$ such that $\phi^{\mathcal{M}} = \phi^{\mathcal{N}} \circ h$ (i.e., such that $\phi^{\mathcal{M}}$ is the pullback of $\phi^{\mathcal{N}}$ under h).[14] The sophisticated category, $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$, has sophisticated

[12] cf. [Barrett, 2015a]; [Swanson and Halvorson, 2012]; [Weatherall, MS].
[13] I'm grateful to Al Wilson for initially suggesting approaching this in terms of quiddities.
[14] $\phi^{\mathcal{M}}$ here denotes the potential ϕ in $\mathcal{M}$, analogously to the use of $P^{\mathcal{S}}$ (say) in first-order model theory

models as objects. In addition to the isometries, its morphisms also include vertical bundle automorphisms of the form (4.5). The natural functor to use in relating these two categories is defined as follows. First, let $K : \mathbb{R} \to \Phi$ be any bijection such that K^{-1} is a bijective embedding of Φ into $\mathbb{R}$. We can then define $\mathrm{K}^* : \mathrm{Mod}(\mathbb{T}_\phi) \to \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$ as the functor such that $\phi^{\mathrm{K}^*\mathcal{M}} = K \circ \phi^{\mathcal{M}}$, and which takes any morphism h in $\mathrm{Mod}(\mathbb{T}_\phi)$ to itself (*qua* isometry of $\mathfrak{X}$) in $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$. We then have the following result.

Proposition 8. $\mathrm{K}^* : \mathrm{Mod}(\mathbb{T}_\phi) \to \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$ is not full.

Proof. Let C^* be the functor $\mathrm{Mod}(\mathbb{T}_\phi) \to \mathrm{Mod}(\mathbb{T}_\phi)$ induced by the symmetry transformation (4.5): i.e. which maps any model to the model obtained by the application of (4.5), and maps all the isometries to themselves. Let $\mathcal{M}$ be any model of $\mathbb{T}_\phi$ such that there is no non-trivial isometry under which $\mathcal{M}$ is invariant. Given this, we know that $\mathrm{C}^*\mathcal{M} \neq \mathcal{M}$. Moreover, it is clear that there is no isometry from $\mathcal{M}$ to $\mathrm{C}^*\mathcal{M}$. Hence, $\mathrm{Hom}(M, \mathrm{C}^*M) = \varnothing$. Yet we also know that $\mathrm{K}^*(\mathrm{C}^*M)$ is related by a vertical bundle automorphism of the form (4.5) to K^*M, and hence that $\mathrm{Hom}(\mathrm{K}^*(\mathrm{C}^*M), \mathrm{K}^*M) \neq \varnothing$. So, the map on arrows induced by K^* is not surjective for the pair of objects M, C^*M; that is, K^* is not full. □

We can provide an analogous result for the case of electromagnetism, showing that the natural functor from $\mathrm{Mod}(\mathbb{T}_A)$ to $\mathrm{Mod}(\widetilde{\mathbb{T}}_A)$ is not full; I forbear discussing it, since it is very similar to the above.

4.4 Reduction

Thus, sophistication is a way of implementing the symmetry-interpretation link. However, its virtues have not always been fully appreciated: in many discussions about the proper way to implement the above interpretational principle for symmetries, it is taken for granted that what we seek is a theory which is the result of a *reduction* (not just a sophistication) by the relevant symmetry. In very general terms, the idea is that we (i) identify some collection of invariants of the original theory; (ii) specify a theory in terms of those invariants; and (iii) show that the new theory captures all the symmetry-invariant content of the old theory. Before getting more specific, it will be best to introduce examples.

We have already met one example of such a theory: the congruence theory $\mathbb{T}_C$ from the last chapter. The point I wish to make here is that the form it exhibits can be

to denote the extension of P in a first-order picture $\mathcal{S}$.

generalised to other cases. For example, consider the case of electrostatics. This time, the chosen invariant is the electric field $\mathbf{E}$, defined by

$$\mathbf{E} := \nabla\varphi \tag{4.9}$$

Again, the first thing we want is some kind of indication that the electric field suffices to capture all the invariant content of the electrostatic theory. So, let $\mathbb{T}_\phi^+$ be the definitional extension of $\mathbb{T}_\phi$ by (4.9), and suppose that $\mathcal{M}$ and $\mathcal{N}$ are two models of $\mathbb{T}_\phi^+$, such that $\mathbf{E}^\mathcal{M} = \mathbf{E}^\mathcal{N}$. Then by elementary integration, their potentials agree to within a symmetry transformation: that is, for some constant k,

$$\varphi^\mathcal{N} = \varphi^\mathcal{M} + k \tag{4.10}$$

So now consider the following theory, $\mathbb{T}_E$. The background bundle of this theory is $T\mathfrak{X}$, the tangent bundle of $\mathfrak{X}$. (Since $\mathfrak{X}$ is an affine space, $T\mathfrak{X}$ can be canonically identified with the trivial product bundle of $V_\mathfrak{X}$ over $\mathfrak{X}$.) The dynamical data for each kinematically possible model consists of a section $\mathbf{E}$ of $T\mathfrak{X}$ (i.e., a vector field over $\mathfrak{X}$) and a scalar field ρ on $\mathfrak{X}$. The equations of the theory are

$$\nabla \times \mathbf{E} = 0 \tag{4.11a}$$

$$\nabla \cdot \mathbf{E} = 4\pi\rho \tag{4.11b}$$

Again in analogy to the handedness case, we have the following pair of observations about how the models of $\mathbb{T}_E$ relate to those of $\mathbb{T}_\phi$. First, for any model M of $\mathbb{T}_\phi^+$, the electric field $\mathbf{E}^\mathcal{M}$ satisfies equations (4.11). This is obvious just from plugging the definition (4.9) into (4.11). Second, for any model $\mathcal{N}$ of $\mathbb{T}_E$, there is a model $\mathcal{N}^+$ of $\mathbb{T}_\phi^+$ such that $\mathbf{E}^{\mathcal{N}^+} = \mathbf{E}^\mathcal{N}$. This is also standard: an irrotational vector field over a simply connected base space admits some scalar field of which it is the gradient.

In fact, we can be more precise yet about the sense in which $\mathbb{T}_E$ captures the invariant content of $\mathbb{T}_\phi$: there is a categorical equivalence between $\mathrm{Mod}(\mathbb{T}_E)$ and $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$. The category $\mathrm{Mod}(\mathbb{T}_E)$ has the models as objects; the morphisms between $\mathcal{M}$ and $\mathcal{N}$ are just the isometries $h : \mathfrak{X} \to \mathfrak{X}$ such that $\mathbf{E}^\mathcal{M} = h^*(\mathbf{E}^\mathcal{N})$, i.e., such that $\mathbf{E}^\mathcal{M}$ is the pullback of $\mathbf{E}^\mathcal{N}$ under h. Let J be the functor $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi) \to \mathrm{Mod}(\mathbb{T}_E)$ whose action on models is given by (taking the dual of) the definition (4.9); its action on any isometry h is to carry it to itself; and for any global potential shift k (i.e. transformation of the form (4.5)) from $\mathcal{M}$ to $\mathcal{N}$, J takes k to $\mathrm{Id}_{\mathrm{J}\mathcal{M}}$ (equivalently, given that $\mathrm{J}\mathcal{M} = \mathrm{J}\mathcal{N}$, to $\mathrm{Id}_{\mathrm{J}\mathcal{N}}$). This fixes

its action on all morphisms in $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$, by functoriality. We then have:

Proposition 9. $\mathrm{J} : \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi) \to \mathrm{Mod}(\mathbb{T}_E)$ is full, faithful and surjective; i.e., it is an equivalence of categories.[15]

Proof. First, consider any $m, n \in \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$, and let h be any morphism from $\mathrm{J}m$ to $\mathrm{J}n$; thus, h is some isometry $\mathfrak{X} \to \mathfrak{X}$. So there are two possibilities: either $m = h^*n$, or m and h^*n are related by some global potential shift k. If the former, then $\mathrm{J}h = h$; if the latter, then $\mathrm{J}(h \circ k) = h$. Either way, therefore, J induces a surjective map on arrows between m and n; so J is full.

Second, consider any $m, n \in \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$, and any morphisms $f, f' : m \to n$ such that $\mathrm{J}f = \mathrm{J}f'$. Since f (or f') can be uniquely decomposed into an isometry $h : \mathfrak{X} \to \mathfrak{X}$ (or h') and a global potential shift $k : \Phi \to \Phi$ (or k'), we have that $\mathrm{J}h = \mathrm{J}h'$ and $\mathrm{J}k = \mathrm{J}k'$. But J takes isometries to themselves, so it follows that $h = h'$. And the global potential shift between any pair of models is unique if it exists; so $k = k'$. Hence, J induces an injective map on arrows between m and n; so J is faithful.

Third, consider any $\mathcal{M} \in \mathrm{Mod}(\mathbb{T}_E)$. As already discussed, for any such model there is some $m \in \mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$ such that $\mathrm{J}m = \mathcal{M}$. So J is surjective, and hence essentially surjective. □

Note that as a consequence, the natural functor from $\mathrm{Mod}(\mathbb{T}_\phi)$ to $\mathrm{Mod}(\mathbb{T}_E)$ is not an equivalence. More specifically, let J^* be the functor $\mathrm{Mod}(\mathbb{T}_\phi) \to \mathrm{Mod}(\mathbb{T}_E)$ which acts on models and morphisms in just the same manner as J does (which defines it fully, given that $\mathrm{Mod}(\mathbb{T}_\phi)$ is a subcategory of $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$). It follows from the above that J^* is not full.

Finally, let us look at the case of electromagnetism. The invariant we use here is the electromagnetic field

$$F_{ab} := \nabla_a A_b - \nabla_b A_a \tag{4.12}$$

Let $\mathbb{T}_A^+$ be the result of supplementing $\mathbb{T}_A$ with the definition (4.12). Once again, we observe first that the electromagnetic field determines all gauge-invariant quantities. That is, for any models (A_a, F_{ab}) and (A'_a, F'_{ab}) of $\mathbb{T}_A^+$, if $F_{ab} = F'_{ab}$ then for some scalar function λ, $A'_a = A_a + \nabla_a \lambda$. This is, again, a standard result.

Now consider the theory $\mathbb{T}_F$. The background bundle of $\mathbb{T}_F$ is J^a and F_{ab}, where $0 \leq a, b \leq 3$ (so $q = 20$), whilst the base-variables are the same as those of $\mathbb{T}_A$. The

[15] cf. [Weatherall, MS, Proposition 2].

equations are

$$F_{ab} = -F_{ba} \tag{4.13a}$$

$$\partial_{[a}F_{b\rho]} = 0 \tag{4.13b}$$

$$\nabla_a F^{ab} = J^b \tag{4.13c}$$

where, again, indices (all of which range from 0 to 3) are raised using the Minkowski matrix (and the square bracket $[\dots]$ indicates anti-symmetrisation). Then, once more, we find a certain kind of alignment between the models of $\mathbb{T}_F$ and the models of $\mathbb{T}_A^+$. That is, for any model $\mathcal{M}$ of $\mathbb{T}_A^+$, the field $F_{ab}^{\mathcal{M}}$ is a solution of (4.13); and for any model $\mathcal{N}$ of $\mathbb{T}_F$, there is a model $\mathcal{N}^+$ of $\mathbb{T}_A^+$ such that $F_{ab}^{\mathcal{N}^+} = F_{ab}^{\mathcal{N}}$.

These examples make fairly clear what is meant by a reduced theory; let us now offer a general definition. Suppose that $\mathbb{T}$ is the target theory, admitting some group G of symmetries (and let us denote the action of $g \in G$ on models by $\mathcal{M} \mapsto g^*\mathcal{M}$). Say that a collection Q of symmetry-invariant quantities/predicates (in $\mathbb{T}$, or in some definitional extension $\mathbb{T}^+$) is *complete* if agreement on Q guarantees agreement to within G: i.e., if it is the case that for any models $\mathcal{M}$ and $\mathcal{N}$ of $\mathbb{T}^{(+)}$, if $q^{\mathcal{M}} = q^{\mathcal{N}}$ for every $q \in Q$, then for some $g \in G$, $\mathcal{N} = g^*\mathcal{M}$. A *reduction* of $\mathbb{T}$ to Q is a theory $\mathbb{T}'$, with dynamical data Q, such that:

(i) for any model $\mathcal{M}$ of $\mathbb{T}'$, there exists some model $\mathcal{N}$ of $\mathbb{T}^{(+)}$ such that for every $q \in Q$, $q^{\mathcal{M}} = q^{\mathcal{N}}$; and

(ii) for any model $\mathcal{M}$ of $\mathbb{T}^{(+)}$, the reduct[16] of $\mathcal{M}$ to Q is a model of $\mathbb{T}'$

I'll refer to the pair of conditions (i) and (ii) as the *Goldilocks conditions* for symmetry reduction: they state that the class of models of the reduced theory must be neither too big nor too small. One expects (although it is not clear how to prove it) that the Goldilocks conditions could also be stated as follows: the reduced theory should be such that its category of models is the *internal skeleton* of the category of the sophisticated theory, i.e., should be the category that results from identifying models related by an internal isomorphism.

Many discussions of symmetry assume, implicitly or explicitly, that changing one's theory to incorporate the lessons of a symmetry—to get rid of the "surplus structure"

[16] In the model-theoretic context, the *reduct* of a Σ-picture $\mathcal{S}$ to $\Sigma' \subset \Sigma$ is the Σ'-picture $\mathcal{S}'$ such that $|\mathcal{S}'| = |\mathcal{S}|$, and for every $\Pi \in \Sigma'$, $\Pi^{\mathcal{S}'} = \Pi^{\mathcal{S}}$ (see e.g. [Hodges, 1997]). In other contexts, I intend the reduct to be what one gets by throwing away all of the structure other than that in Q (e.g. when one throws away the electric potential from a model of electrostatics, but keeps the electric field).

the symmetry reveals—means moving from the original theory to a reduced theory.[17] It is worth pointing out, however, that there are problems with making reduction the gold standard for expunging surplus structure. First, it is highly non-trivial to find such a reduced theory—or even to demonstrate with confidence that such a theory could exist. All the examples above were chosen as cases where we know how to specify the reduced theory. But doing so required that we could both find a complete set of invariant quantities Q, and then provide a theory in terms of Q whose class of models meets the Goldilocks conditions. Note that these tasks are somewhat in tension. Plausibly, the set of *all* invariant quantities will always be complete.[18] But the more invariant quantities one wants to use in Q, the harder it is going to be to build a finitely or recursively axiomatisable theory out of them (satisfying both the Goldilocks conditions).[19]

As an illustration of these perils, consider the theory $\mathbb{T}^{\tau}_{A}$. The equations of this theory are precisely the same as those of $\mathbb{T}_{A}$: the only difference is that background bundles of this theory may have base space U, where U is any open subset of $\mathfrak{N}$. So, in particular, models of this theory include cases where the base space is topologically non-trivial. It is now no longer the case that the set F_{ab} comprises a complete set of quantities: there are gauge-invariant quantities which are not determined by fixing the value of F_{ab} everywhere. To take the best-known example, define the *holonomy* of a loop γ to be

$$h(\gamma) = \exp\left(\oint_{\gamma} A_a \mathrm{d}x^a\right) \tag{4.14}$$

It is straightforward to verify that holonomies are gauge-invariant. Yet if U is not simply connected, the value of F_{ab} everywhere in U underdetermines the values of the holonomies: two models of $\mathbb{T}^{\tau}_{A}$ (both with base space U) might agree on the former, yet disagree on the latter.[20] Of course, this does not mean that there can be no reduced theory of $\mathbb{T}^{\tau}_{A}$. It certainly doesn't mean that there is no complete set of invariant quantities for $\mathbb{T}^{\tau}_{A}$: in fact, it can be shown that the set of all holonomies comprises just such a complete set. However, it remains very much an open question whether one

[17] e.g. [Earman, 2003], [Healey, 2007], [Baker, 2010], [Dasgupta, 2014b].

[18] Note that proving this will not be entirely straightforward: one could imagine certain global obstructions (e.g. topological issues) that might yield a pair of models agreeing on all invariants, yet lying on different symmetry orbits.

[19] The rider "finitely or recursively axiomatisable" is necessary to rule out theories consisting simply of all the logical consequences of $\mathbb{T}$ expressible in terms of Q. (My thanks to Teruji Thomas for highlighting this possibility.)

[20] This fact is the essential kernel of the Aharonov-Bohm effect [Aharonov and Bohm, 1959]; for further details, see [Healey, 2007].

can give some closed-form set of equations for holonomies, such that the solutions of those equations satisfy the Goldilocks conditions (relative to the definitional extension of $\mathbb{T}^{\tau}_{A}$ by (4.14)).[21] By contrast, forming a *sophisticated* version of the theory $\mathbb{T}^{\tau}_{A}$ is straightforward: we simply take models to be connections on principal $U(1)$-bundles over $U \subseteq \mathfrak{N}$. And these models do indeed contain all the same gauge-invariant quantities as the unsophisticated models: in particular, such a connection fixes the values of all the holonomies.

By contrast, we have seen that finding a sophisticated semantics will always be easy if we use the external method. And although we don't have any kind of general guarantee that we will thereby be able to find some kind of internal characterisation of those structures, we do—as a matter of fact—generally seem to have success in finding them. This isn't terribly mysterious when one appreciates the role that symmetry considerations play in the construction of theories. If we are demanding that the equations of the theory manifest certain symmetries, then the easiest way to ensure that they do is to construct them as equations governing objects upon which the sought-for symmetries act as isomorphisms. As a result, modern theories are typically *born* sophisticated. (The paradigm case is the construction of Yang-Mills theories as theories governing connections on a principal G-bundle, which then ensures a sophisticated semantics with respect to G acting as a local gauge group.)

The second problem with insisting that one must provide a reduced theory is that, even if such a theory can be found, that theory may well have explanatory deficits relative to the original theory. For the reduced theory treats the invariant quantities Q as primitives; this means that if some $q \in Q$ obeys some non-trivial condition as a result of its definition (in the unreduced theory), it must be asserted to obey that condition (in the reduced theory) as a simple posit. Let us consider some examples of this phenomenon.

For the handedness theory, note that the reduced theory T_C includes axioms to the effect that C is an equivalence relation. No such axioms are needed in the theory T^{+}_{H}, since—in that theory—the definition of C (3.17) entails that it is an equivalence relation. For example, the claim that C is symmetric becomes, when translated using (3.17), the tautology

$$\forall x \forall y (((Lx \wedge Ly) \vee (Rx \wedge Ry)) \rightarrow ((Ly \wedge Lx) \vee (Ry \wedge Rx))) \tag{4.15}$$

[21] See [Loll, 1994] for discussion.

In the case of electrostatics, one can see that the equation (4.11b) in $\mathbb{T}_E$ corresponds to the equation (4.4) of $\mathbb{T}_\phi$. Equation (4.11a) is a new addition, however; again, the reason it is not needed in $\mathbb{T}_\phi$ is because, translated using (4.9), it becomes the mathematical truth that

$$\nabla \times \nabla\varphi = 0 \tag{4.16}$$

This example also demonstrates that this phenomenon is part of what makes finding a reduced theory so hard. In trying to find the reduced version of $\mathbb{T}_\phi$, one might be encouraged by the observation that ϕ only ever appears in (4.4) in the form $\nabla\phi$—which is a complete invariant. Even then, though, one still has work to do. It's not enough to merely substitute $\mathbf{E}$ for $\nabla\phi$ in (4.4); one also has to add in further equations to recapture conditions such as (4.16).

For electromagnetism, it is the equations (4.13a) and (4.13b) which have no counterpart in the unreduced theory $\mathbb{T}_A$; for in that theory, they reduce to the mathematical trivialities that

$$\nabla_a A_b - \nabla_b A_a = -(\nabla_b A_a - \nabla_a A_b) \tag{4.17a}$$

$$\nabla_{[a}\nabla_b A_{\rho]} = 0 \tag{4.17b}$$

The list goes on. Any attempt to reduce $\mathbb{T}_A^\tau$ to holonomies must stipulate that the holonomies obey various identities; attempting to reduce a non-Abelian gauge theory to so-called "Wilson loops" (the relevant analogue of the holonomies for the non-Abelian case) requires positing an even more restrictive set of conditions still.[22] Or consider relationalist theories of space, which must posit constraints amongst the spatial relations (e.g. the Triangle Inequality) that merely follow from the definitions of those relations on substantivalist views.[23] Why is it bad for the reduced theory to introduce these extra conditions as primitive posits? Part of the issue is just that it adds to the complexity of those theories. More significantly, though, it seems to remove a certain *explanatory* virtue from the original formulation of the theory. In the unreduced theory, there is a good answer to the question of *why* the invariant quantities obey these conditions: they obey these conditions because of how they are built up out of other kinds of structure in the theory. In the unreduced theory, it seems, we get some kind of insight into these conditions—an insight that risks being lost, or occluded, if we insist that the reduced theory is the be-all and end-all.

[22] See [Arntzenius, 2012, chap. 6].
[23] See [Maudlin, 2007a, chap. 3].

And we should note that the invariants remain definable, even using the sophisticated semantics: the fact that a sophisticated semantics determines unambiguous truth-values for invariant sentences of the language guarantees that the definitions will remain well-posed. As a result, the explanation of why the invariants manifest such-and-such features are also preserved. In the handedness theory, for example, it remains the case that congruence is a matter of possessing the same handedness property—and, hence, that congruence is an equivalence relation. The electric field is still definable as the gradient of the potential, even if the latter is taking values in Φ rather than $\mathbb{R}$; so its irrotationality is still explicable as a consequence of its being a gradient. In the case of electromagnetism, one can still understand the definition (4.12) of F_{ab} as the antisymmetric part of the four-gradient of the vector potential (of the target connection relative to an arbitrarily chosen flat reference connection); however, it is more insightful to appreciate that this is precisely the definition of the curvature of the connection. Either way, however, the fact that F_{ab} is antisymmetric (4.13a) and governed by the homogeneous Maxwell equation (4.13b) receives a satisfying explanation.

4.5 Equivalence and explanation

We've now seen three forms a theory can take (or more carefully, which a formally interpreted theory can take): an unreduced and unsophisticated form (let's call it the *vulgar* form), in which there are symmetries relating non-isomorphic models; a reduced form, in which there are no symmetries; and a sophisticated form, in which symmetries relate isomorphic models. Along the way, I've noted some results concerning the extent to which these forms ought to be considered equivalent to one another. In particular, I noted that the reduced and sophisticated forms are plausibly equivalent to one another, given that their categories of models are equivalent; and that neither seems to be equivalent to the unsophisticated form, at least not under the natural or obvious translation.

This does not prove that there are *no* categorical equivalences between $\mathrm{Mod}(\mathbb{T}_\phi)$ and either $\mathrm{Mod}(\mathbb{T}_E)$ or $\mathrm{Mod}(\widetilde{\mathbb{T}}_\phi)$. However, it seems very plausible that any functor which is describable in appropriately systematic terms (i.e. which meshes appropriately with respect to the non-categorical characterisation of the models) will not be an equivalence. (Proving this formally would have to await a precisification of "appropriately systematic" or "meshes appropriately".) And we do unambiguously have the result that the categories of sophisticated models come out equivalent to the relevant category

of reduced models.

All of this suggests some general (if vague) conjectures. Suppose that a theory $\mathbb{T}$ admits some group G of symmetries (and that $\mathbb{T}$ is unsophisticated with respect to G). Let $\mathbb{T}'$ be a reduction of $\mathbb{T}$ to some complete set of G-invariants, and let $\widetilde{\mathbb{T}}$ be the sophistication of $\mathbb{T}$ by G. Finally, let's say that a "reasonable" functor is one which meshes appropriately with the architecture of the models (whatever exactly that gets made out to mean).[24] Then the following conjectures seem plausible:

- There is a reasonable functor $F : \mathrm{Mod}(\widetilde{\mathbb{T}}) \to \mathrm{Mod}(\mathbb{T}')$ which is full, faithful, and essentially surjective.
- There are no reasonable functors from $\mathrm{Mod}(\mathbb{T})$ to either $\mathrm{Mod}(\mathbb{T}')$ or $\mathrm{Mod}(\widetilde{\mathbb{T}})$ which are full, faithful and essentially surjective (or perhaps the stronger claim: there are no such functors which are full).

Making these conjectures precise would require (a) a more thorough treatment of how to characterise reduction and sophistication in category-theoretic terms, and (b) a clarification of the notion of "reasonableness". I defer doing so to future work; instead, let us consider the philosophical implications of these technical observations.

Begin with the inequivalence between the reduced and unreduced theories. Prima facie, this may seem in tension with Weatherall's claim that categorical equivalence (of categories of models) is "a criterion of equivalence that does capture the sense in which [electromagnetism in terms of fields] and [electromagnetism in terms of potentials] are synonymous."[25] However, there is no serious disagreement here. The equivalence that Weatherall describes is between electromagnetism formulated in terms of fields—what we have been calling $\mathbb{T}_F$—and electromagnetism formulated in terms of potentials, *when gauge transformations are counted as morphisms in its category of models*. In other words, the equivalence described by Weatherall is precisely the equivalence between the reduced theory on the one hand, and the unreduced theory *under the sophisticated semantics* on the other.

However, this does highlight a reason why one has to be careful in the use of categorical equivalence as a criterion for theory equivalence. Categorical equivalence does not straightforwardly pronounce on the equivalence of theories if they are conceived

[24] At least in our examples, reasonableness seems to be a matter of being definable in terms of the *syntactic* content (e.g. being generated by a translation between two theories). Hence, my emphasis on reasonableness accords with recent work on the sometimes-neglected virtues of the syntactic view of theories ([Halvorson, 2012], [Halvorson, 2013], [Lutz, 2014a], [Lutz, 2014b]).

[25] [Weatherall, 2015a, p. 15]

of syntactically, as sets of sentences: rather, it passes judgment on the equivalence of theories relative to a certain way of characterising the models of a theory as a category. In other words, categorical equivalence is a criterion that applies to theories *together with* a choice of semantics: change the semantics (from a vulgar to a sophisticated semantics, for example) and one will, in general, change the category of models. To be clear, all of this is present in Weatherall's discussion, albeit in a slightly different form. Whereas I have emphasised the need to specify (not just a theory, but also) the semantic structures one intends to use in formally interpreting the theory, Weatherall speaks of constructing the category of models of a theory in such a way that we appropriately privilege "maps that preserve the "physical structure" of a model, in the sense that two models related by such a map are physically equivalent."[26] I take these to be two ways of getting at the same idea. If one intends to renounce commitment to a certain amount of structure in one's models as "unphysical", then one had better also think that the role such structure plays in determining the semantic content of the theory is inessential and/or the product of arbitrary convention.

With these clarifications to hand, it does seem right to say that the reduced and unreduced theories are not equivalent. Electromagnetism with fields and electromagnetism with potentials can only feasibly be regarded as equivalent if gauge symmetries are regarded as relating physically equivalent models; but to judge that they do so is precisely to affirm a commitment to sophisticated rather than vulgar semantics as embodying the true commitments of the theory.

However, what of the relationship between the reduced and sophisticated categories of models? In what sense are sophistication and reduction equivalent? In particular, one might be worried by the fact that I suggested that sophistication about theories might possess superior explanatory powers compared to reduction. So if there is indeed something to choose between them, surely they can't be equivalent after all?

Here is what seems to me like the right thing to say: the two theories are equivalent in terms of their *ontology*, in terms of the kinds of structures that they postulate as present in any world aptly described by them; but they differ in their *explanatory* structure. Electrostatics in terms of sophisticated potentials, and electrostatics in terms of fields, both agree that there is a physically significant irrotational vector field; and both agree that this field (as with any such field) is representable as the gradient of a scalar field—provided that that scalar field is defined only up to potential shifts, or (equivalently) that it take values in Φ rather than $\mathbb{R}$. However, they disagree over what

[26] [Weatherall, 2015a, p. 17]

kind of explanation can be given of why this vector field is irrotational. For the theory in terms of fields, its irrotationality is simply a brute fact—a fact which usefully permits the field's representation as a certain kind of gradient, but not arising from anything else. For the theory in terms of sophisticated potentials, the field is the derivative object, and so admits of an explanation in terms of what is fundamental (i.e., the potential): it is irrotational *because* it is a gradient, and gradients always have vanishing curl.

As a result, whether the two theories are "really" equivalent will turn on what one wants to say about the role of explanation in theory equivalence. On some accounts,[27] two theories cannot be equivalent if they offer different explanations of the phenomena. This will be particularly true if one is inclined to view explanations of this sort as arising from some kind of ontological structure out there in the world, such as if one is committed to some notion of grounding—conceived of as a genuine part of the world's architecture, and responsible for answering in-virtue-of questions (e.g. "in virtue of what is the electric field irrotational?").[28] If, however, one is sceptical of grounding (and cognate notions), then there is space for some more quietist or deflationary attitude towards the relevant explanations. Such a view meshes naturally with the scepticism I sounded earlier about fundamentality or naturalness.

On this kind of view, there need not always be some fact of the matter about what kind of explanatory architecture is correct. It is certainly illuminating to see that some feature in a theory *can* be explained by another, if the theory is set up a particular way; but (in general) there is no compulsion towards setting the theory up one way rather than another, or towards accepting one pattern of explanation amongst its parts as uniquely privileged.[29]

On either account, though, the case can be made for valuing sophistication over reduction. On some more realist account of explanation (e.g. the grounding account), the explanatory virtues of sophistication make it more likely to be the correct account of the (objective) grounding structure of the world. On a more deflationary picture, those virtues make it a more helpful or convenient way of characterising the structure of the world; even if a reduced theory is picking out the same structure, it will generally do so in a less tractable way. And of course, both accounts will appreciate the fact that sophistication is typically easier to come by than reduction.

[27] e.g. [Putnam, 1983]

[28] See e.g. the essays in [Correia and Schnieder, 2012].

[29] I read Weatherall's "puzzleball" account of the foundations of physical theories [Weatherall, 2012] as expressing this kind of picture; it is also closely related to Cartwright's "dappled-world" conception of inter-theoretic relationships [Cartwright, 1999].

Overall, the main aim of this chapter has simply been to convince you that fixating on reduction as the only acceptable means of dealing with symmetries is a mistake.[30] If, as I've argued, sophistication rather than reduction is a legitimate way to seek to expurgate symmetry-variant structure, then a number of interesting consequences follow. One is that carrying out that expurgation becomes (in general) somewhat more straightforward: if all we are required to do is provide a sophisticated understanding of the theory (especially if we do so using the external method), then our lives are made substantially easier than if we need to find a reduced theory. Moreover, with more expurgatory options on the table, we can open up new approaches to classic problems concerning symmetry. The debate on the Aharonov-Bohm effect, for example, is often characterised as requiring us to choose between a trilemma of unpalatable ontologies: a locally acting[31] and separable (but not gauge-invariant) ontology of potentials; a locally acting and gauge-invariant (but non-separable) ontology of holonomies; or a separable and gauge invariant (but non-locally acting) ontology of fields. But the argument here suggests another option: adopting the "sophisticated" ontology of connections of a principal bundle (or, more carefully, of whatever the metaphysical correlate of such a connection is). I don't claim that doing so will magically resolve these problems;[32] but it at least enlivens the conceptual geography.

[30] In this regard, cf. [Pooley, 2013b], [Weatherall, 2016], and [Weatherall, MS].

[31] In the sense of having no "action at a distance".

[32] In the Aharonov-Bohm case, for instance, there will be significant subtleties about the sense in which connections are separable: note that specifying a connection on a region U, and a connection on an overlapping region V, generally underdetermines the connection on $U \cup V$ (absent information about how things stand in $U \cup V$).

5 External symmetries

> But I've got a blank space, baby
> And I'll write your name.
>
> Taylor Swift, *Blank Space*

5.1 External transformations

We now turn to *external* transformations, which we define as follows.

Definition 7. Let $E \xrightarrow{\pi} M$ be a fibre bundle. An *external transformation* on E is a bundle automorphism $(\alpha : E \to E, \beta : M \to M)$ which is only vertical if it is trivial (i.e., is such that unless $\alpha = \mathrm{Id}_E$, $\beta \neq \mathrm{Id}_M$).

So external transformations are mostly those bundle automorphisms which are not internal; we "count" the identity transformation as an external transformation just in order that external transformations will form groups. As with any bundle automorphism, data defined on the bundle can be transformed by an external transformation: a section s gets pulled back to $s^* := \alpha^{-1} \circ s \circ \beta$ (or pushed forward to $s_* := \alpha \circ s \circ \beta^{-1}$), as is a connection D. Hence, any external transformation α for the background bundle of some model $\mathcal{M}$ naturally yields a pullback model $\alpha^*\mathcal{M}$ (or a pushforward model $\alpha_*\mathcal{M}$).

At first glance, this definition may look a little unusual: it is more common to see external transformations defined simply as transformations of the spatiotemporal variables—not as transformations of the spatiotemporal variables, *conjoined* with transformations of the non-spatiotemporal variables! There are three reasons why this definition is preferable. First, it means that the categories of "internal transformation" and "external transformation" are jointly exhaustive. Second, this is the only definition that is apt for application to generic fibre bundles: for, in a generic fibre bundle, there is no sense to be made of the idea of transforming *only* the spatiotemporal degrees of

freedom. For instance, consider the principal bundle electromagnetic theory discussed in the previous chapter. If we merely specify a transformation β on the base space M, that does not tell us where to send the members of one fibre: we know that $f \in F_p$ must be sent to some point of $F_{\beta(p)}$, but beyond that we are in the dark. Of course, if there was a canonical identification between the points of F_p and those of $F_{\beta(p)}$, then there would be a natural candidate—but the characteristic feature of fibre bundles, recall, is the absence of such canonical trans-fibre identities.[1] Hence, the only way to specify an external transformation in a bundle such as this is to specify a transformation $\alpha : E \to E$ which "drops" to a transformation $\beta : M \to M$: that is, is such that for any $e, f \in F_p$, $\alpha(e) \in F_{\beta(p)}$ iff $\alpha(f) \in F_{\beta(p)}$.

That said, many theories are set on bundles for which this problem does not arise: that is, on bundles in which specifying a diffeomorphism on the base space *does* canonically single out some privileged transformation on the bundle as a whole. We describe this by saying that, in such a bundle, diffeomorphisms on the base space "lift" to the total space. One way this can happen is if the bundle in question is a product bundle, for then we do have trans-fibre identities which serve to fix the lift. So, for instance, if our model takes the form of a function $f : M \to N$ (which may be regarded as a section of the bundle $N \times M \stackrel{\pi_M}{\to} M$), then it is natural to think of the map $\beta : M \to M$ as transforming f to its pullback $f \circ \beta$.

But the phenomenon is more generic than this. For instance, tensor bundles also exhibit lifts: this is just the observation that diffeomorphisms on a manifold induce pullbacks and pushforwards on any tensor. In general, say that a bundle E, with base space M, is a *basic bundle* if for any diffeomorphism $\beta : M \to M$ there exists a uniquely natural diffeomorphism $\lambda_\beta : E \to E$ (the "lift" of β), such that (λ_β, β) is a bundle automorphism (i.e. $\pi \circ \lambda_\beta = \beta \circ \pi$) and such that $\lambda_* : \mathrm{Diff}(M) \to \mathrm{Diff}(E)$ is a group homomorphism (i.e. $\lambda_{\beta' \circ \beta} = \lambda_{\beta'} \circ \lambda_\beta$). The terminology is to indicate that for such bundles, the base space is more important than is the case for a generic vector bundle. I stress that this criterion is vague and heuristic, since it depends on the hand-wavey notion of "uniquely natural". In fact, the only examples of basic bundles that we will deal with in this chapter are tensor bundles; I offer this general definition only to indicate that tensor bundles are not the only such case (spinor bundles, for instance, manifest the same phenomenon). Data on a basic bundle (e.g. sections and connections of tensor bundles) will be referred to as *geometric object fields*.

[1] Note that not even having a connection around would help: unless the connection is flat, the identification of points in one fibre with those in any fibre (that isn't infinitesimally nearby) will not be canonical.

Corresponding to this, a theory whose background bundles are basic will be referred to as a *basic theory*. (This terminology should *not* be taken to mean that such theories are always simple!) It will often be helpful to write the models of a basic theory in the form $\langle M, \{O_i\}\rangle$, where M is the base space and the O_i are the geometric object fields. For basic theories, one can specify an external transformation *merely* by specifying a certain diffeomorphism on the base space; if no further information is provided, then it can simply be understood that the intended bundle automorphism is the lift of the specified base-space diffeomorphism. For instance, in the theory of instantaneous electrostatics, it suffices to merely give a diffeomorphism $h : \mathfrak{X} \to \mathfrak{X}$; we can then transform any model $\langle \mathfrak{X}, \phi, \rho\rangle$ into the pullback model $\langle \mathfrak{X}, h^*\phi, h^*\rho\rangle$ or the pushforward model $\langle \mathfrak{X}, h_*\phi, h_*\rho\rangle$. Similarly, in Maxwell electromagnetism, any diffeomorphism $d : \mathfrak{N} \to \mathfrak{N}$ induces a map taking any model $\langle \mathfrak{N}, A_a, J^a\rangle$ to its pullback $\langle \mathfrak{N}, d^*A_a, d^*J^a\rangle$ (and one taking any model to its pushforward). However, although the lift is the uniquely natural bundle automorphism to associate to a given base-space diffeomorphism, it will not typically be the *only* bundle automorphism which "projects" to that base-space diffeomorphism. We will refer to a bundle automorphism which is the lift of some base-space diffeomorphism as inducing a *pure* external transformation; otherwise, we are dealing with an *impure* external transformation.

In this chapter, I will confine my attention to basic theories, just in order to keep things reasonably simple. As per usual, however, making these kinds of simplifications comes at a cost. One is simply that we do not immediately have an account of external transformations for non-basic theories, such as the principal $U(1)$-bundle formulation of electromagnetism; although the extension of what I say here to non-basic theories should be reasonably straightforward, it would require a lot of careful work to keep everything hygienic. More subtly, however, working with basic bundles encourages us to blur the distinction between transformations on the base space and automorphisms of the bundle as a whole. Although such blurring is convenient, and will be used below, it is not without its risks: in §5.6, I will clear up some confusions that it can generate.

5.2 External symmetries

The definition of an external symmetry of a theory is what one would expect: it is an external transformation of the theory's background bundle which maps solutions to solutions. Let us look at some examples. In instantaneous electrostatics, the external symmetries consist of the isometries of the base space $\mathfrak{X}$: translations, rotations, and

parity transformations (or, more precisely, the natural lifts of those isometries). In an adapted coordinate system $x^i : \mathfrak{X} \to \mathbb{R}^3$ (i.e., a coordinate system which preserves $\mathfrak{X}$'s structure as a Euclidean space), these transformations take the form

$$x^i \mapsto R^i_j x^j + a^i \tag{5.1}$$

where $a^i \in \mathbb{R}$, and the coefficients R^i_j form an orthogonal matrix. One way to see that any such transformation $h : \mathfrak{X} \to \mathfrak{X}$ is a symmetry of T_{IE} is to observe that, since h is an automorphism of $\mathfrak{X}$, the kinematically possible models $\langle \mathfrak{X}, \rho, \phi \rangle$ and $\langle \mathfrak{X}, h^*\rho, h^*\phi \rangle$ are isomorphic to one another (with h itself as the isomorphism): hence, the one is dynamically possible iff the other is. We will discuss cases of this kind in section 5.3.

In Maxwellian electromagnetism, the external symmetry transformations are the *Poincaré transformations*. Again, let $x^\mu : \mathfrak{N} \to \mathbb{R}^4$ be an adapted coordinate system. First, define the collection of coefficients $\eta_{\mu\nu}$ by

$$\eta_{\mu\nu} := \begin{cases} 1 & \text{if } \mu = \nu = 0 \\ -1 & \text{if } \mu = \nu = 1, 2, \text{ or } 3 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

The *Lorentz transformations* are then all and only those transformations of the form

$$x^\mu \mapsto \Lambda^\mu{}_\nu x^\nu \tag{Lor}$$

for some collection of coefficients $\Lambda^\mu{}_\nu$ such that

$$\Lambda^\alpha{}_\mu \eta_{\alpha\beta} \Lambda^\beta{}_\nu = \eta_{\mu\nu} \tag{5.3}$$

Finally, the full set of Poincaré transformations is the semidirect product of the Lorentz transformations with spatiotemporal translations, which means that it consists of all transformations of the form

$$x^\mu \mapsto \Lambda^\mu{}_\nu x^\nu + c^\mu \tag{Poin}$$

where $\Lambda^\mu{}_\nu$ is as above, and $c^\mu \in \mathbb{R}^4$.

In this case, it is *not* true that symmetry-related models are (always) isomorphic. The automorphisms of Newtonian spacetime are given by the *Newton group*, which is a subgroup of the Poincaré group: writing the coordinates x^μ as (t, x^i), the Newton

group is given by transformations of the form

$$\begin{aligned} t &\mapsto t + c^0 \\ x^i &\mapsto R^i_j x^j + c^i \end{aligned} \tag{Newt}$$

where (as before) $c^\mu \in \mathbb{R}^4$ and R^i_j forms an orthogonal matrix. Models related by a member of the Newton group are isomorphic to one another, but the Newton group is a *strict* subgroup of the Poincaré group: it does not include *Lorentz boosts*.

Since a Lorentz boost l is not an automorphism of $\mathfrak{N}$, $\langle \mathfrak{N}, A_a, J^a \rangle$ will not be isomorphic to $\langle \mathfrak{N}, l^* A_a, l^* J^a \rangle$. To verify that the group of Poincaré transformations is nevertheless a symmetry group of the theory, we must instead verify the invariance of the equations (4.6) under (Poin); this is straightforwardly done by using their expression in the adapted coordinates x^μ, which (using the Einstein summation convention) may be given as

$$\eta^{\nu\rho}\eta^{\mu\sigma}\partial_\mu(\partial_\rho A_\sigma - \partial_\sigma A_\rho) = J^\nu \tag{5.4}$$

where A_μ and J^ν are the x^μ-components of A_a and J^a, and the coefficients $\eta^{\mu\nu} = \eta_{\mu\nu}$. We will discuss cases of this kind in section 5.4.

It will be helpful to contrast $\mathbb{T}_A$ with a theory that resembles it as closely as possible, except with respect to the structure of its external symmetries. As such a theory, I will use *Galilean electrodynamics*, $\mathbb{T}_G$. Galilean electrodynamics has the same kinematics as Maxwellian electromagnetism; however, its dynamics are specified by

$$\nabla^2 \phi = -\rho \tag{5.5a}$$

$$-\nabla^2 \mathbf{A} + \nabla\left(\nabla \cdot \mathbf{A} + \frac{\partial \phi}{\partial t}\right) = \mathbf{J} \tag{5.5b}$$

These equations form the so-called *electric limit* of Maxwellian electrodynamics, expressed in terms of potentials (they are obtained from (4.6) simply setting $\partial\mathbf{A}/\partial t = 0$).[2] Like (4.6), the equations (5.5) are covariant under the action of the Newton group (Newt). However, they do *not* have the same boost symmetries.

Rather than Lorentz boosts, the boost symmetries of (5.5) are given by *Galilean boosts*:

[2]The electric limit is one of two non-relativistic limits of Maxwell's equations; the other is referred to as the *magnetic limit*. See [Le Bellac and Lévy-Leblond, 1973], [Holland and Brown, 2003], and [Manfredi, 2013] for discussion .

$$t \mapsto t \tag{5.6a}$$

$$\mathbf{x} \mapsto \mathbf{x} - \mathbf{v}t \tag{5.6b}$$

The group of combinations of the Newton group with transformations of the form (5.6) is known as the *Galilean group*. That is, the Galilean group consists of all transformations of the form

$$\begin{aligned} \mathbf{x} &\mapsto \mathbf{R}\mathbf{x} + \mathbf{v}t + \mathbf{c} \\ t &\mapsto t + c^0 \end{aligned} \tag{Gal}$$

Since the Galilean group strictly contains the Newton subgroup, not all external symmetries of Galilean electrodynamics are automorphisms of the base space $\mathfrak{N}$.

5.3 External isomorphisms

We saw, in the previous chapter, that there was a close connection between internal symmetry transformations and anti-quidditism. I will now argue that to understand the import of external symmetry transformations, we must draw upon the resources of Chapters 2 *and* 3. In this section, we consider those cases in which the external symmetry is an automorphism of the base space. As indicated above, all the external symmetries of instantaneous electrostatics are of this character, as are the members of the Newton subgroup of the external symmetries of Maxwellian electromagnetism or Galilean electrodynamics. And as discussed there, the reason why such an automorphism is guaranteed to be (or rather, to lift to) an external symmetry is that it is an isomorphism between models related by its lift. For, suppose that $\beta : M \to M$ is an automorphism of M. Then $\beta M = M$, and hence

$$\begin{aligned} \langle M, \{\beta_* O_i\}\rangle &= \langle \beta M, \{\beta_* O_i\}\rangle \\ &\cong \langle M, \{O_i\}\rangle \end{aligned}$$

Thus, we know that such automorphisms will account for at least some of the external symmetry transformations.

The relevant observation to make about such a transformation is that it falls within the compass of the anti-haecceitism discussed in Chapter 2. We saw there that anti-haecceitism is most perspicuously characterised as the view that isomorphic possibility-

representations should be understood to represent the same possible world. But this is precisely the situation we find ourselves in: confronted by various isomorphic models, and confronted by the question of how they relate to one another. Hence, the considerations adduced in Chapter 2 in favour of identifying these representatives hold. Such a claim is not original: it is the view, most associated with Pooley, known as *sophisticated substantivalism*.[3]

Let us look directly at the consequences of recognising this as an instance of the haecceitist/anti-haecceitist debate. We observed that there was no way of attaining a determinate referential relationship between the distinct (though isomorphic) Kripke-worlds and the haecceitistic possibilities, and that haecceitism gave rise to a certain kind of in-principle ignorance; moreover, we saw that the former means that the latter is inexpressible. It is easy to see that these features are reproduced. For instance, consider the application of a spatial translation $a : \mathfrak{N} \to \mathfrak{N}$ to a model $\mathcal{M} = \langle \mathfrak{N}, Y \rangle$ of Galilean electrodynamics (letting Y abbreviate the dynamical data (A_a, J^a)). We obtain another model $a_*\mathcal{M} = \langle \mathfrak{N}, a_*Y \rangle$, differing from the first only in that whereas (say) $\rho(p) = 0$, $a_*\rho(p) = 3$—and instead, $a_*\rho(a(p)) = 0$. The notion that such a pair of models are epistemically indistinguishable is widely accepted, going back (at least!) to the correspondence between Leibniz and Clarke.[4] The inexpressibility of the ignorance thereby engendered was first discussed explicitly by [Maudlin, 1993]—indeed, this case is a motivating example for the discussion of inexpressible ignorance in [Dasgupta, 2015].

The issues we raised in Chapter 2 about *de re* modality also arise in this case. For example, consider a Newtonian world W which consists of nothing but a single particle at absolute rest: it occupies some point p of absolute space throughout all time. If there are no worlds which differ from this world by a static shift, then aren't we left with a rather strong essentialist claim about this point: that *necessarily*, had there been only one particle in the world, then it would have occupied this point and no other. But how can that be? What is so special about p that it is the only possible place for a lone particle to be?[5]

The solution to this problem should not be terribly surprising, given the discussion of Chapter 2: we should not analyse *de re* modal claims such as this by looking at variation across worlds, but rather at variation across counterpart relations. What counterpart relations will be available between the space(time) points of one model

[3] See [Pooley, 2006], [Pooley, 2013b].
[4] [Alexander, 1956]
[5] This formulation of the example is that of [Arntzenius, 2012, p. 178, n. 10], but the point is due originally to [Skow, 2005, pp. 32–33].

and those of another? If we take counterpart relations to supervene upon qualitative character, then the answer is uniquely determined by considering the structure of each model. It does not follow, however, that the only available counterpart relations are those which preserve as many as possible of those qualitative properties: that would be one way for the supervenience to work, but isn't required. And indeed, as this example illustrates, that would be a rather implausible standard, since it would mean that the only counterpart functions available between a model and some isomorphic companions would be the isomorphisms between them. A better standard, in this context, is to admit that *at least* any map between the spacetime points of two models which preserves spacetime structure ought to be available as a counterpart function; this would preserve at least some sense in which the spatial relations in which a point stands are essential to it.[6]

If this line is taken, then the example here is disarmed. Let h be any non-trivial spatial translation, mapping each spacetime point of W to some uniformly displaced point. This is an appropriate counterpart function: hence, the claim that p is necessarily occupied is true only if the point $h(p)$ is occupied. Since $h(p)$ is not occupied, the claim is false; it is not a necessary or essential fact about p that it be occupied by a lone particle.

Conversely, the haecceitist's natural definition of determinism (Definition 2) will yield the somewhat unpalatable result that General Relativity is indeterministic: this is the celebrated (or notorious) "Hole Argument"[7] for General Relativity (although, in fact, analogous cases can be constructed for other spacetime theories). Models of GR are of the form $\langle M, g, T\rangle$, where M is a four-dimensional differentiable manifold, g a metric field on M, and T a tensor field on M, satisfying Einstein's field equations. Take any such model $\mathcal{M} = \langle M, g, T\rangle$. For any diffeomorphism (smooth bijection) $d : M \to M$, $\mathcal{M}' = \langle M, d^*g, d^*T\rangle$ is isomorphic to $\mathcal{M}$, and (hence) $\mathcal{M}'$ is a model of GR. This means that the external symmetry group of General Relativity set on M is the full diffeomorphism group $\mathrm{Diff}(M)$. We will refer to this property as *general covariance*, although there is a long history of dispute over the best way to use that label (and over whether general covariance—however defined—is a robust or interesting property of theories). Some of this controversy will be sampled in section 5.6.

For now, the important point is that in general, $\mathcal{M} \neq \mathcal{M}'$: in $\mathcal{M}$, the tensor fields g and T are distributed over the manifold M in one way, whilst in $\mathcal{M}'$, they are dis-

[6] cf. [Newton, 2004], [Maudlin, 1988].
[7] [Earman and Norton, 1987]

tributed across M in a different way. So in general, for the haecceitist, $\mathcal{M}$ and $\mathcal{M}'$ represent distinct possible (general-relativistic) worlds. So now suppose that d is a *hole diffeomorphism* on M: a diffeomorphism which is the identity everywhere outside some bounded neighbourhood H of M (the "hole"), but diverges smoothly from the identity within H. Then $\mathcal{M}$ and $\mathcal{M}'$ agree on everything outside H, despite representing distinct possible worlds. However, we can make H arbitrarily small; we can certainly arrange it so that there is some Cauchy surface S which does not intersect H. So, in particular, $\mathcal{M}$ and $\mathcal{M}'$ agree on S. So GR is not deterministic after all—at least, not according to Definition 2. According to the anti-haecceitist's natural definition of determinism, however (Definition 3), we get the "right" result: since these diffeomorphic models are isomorphic (and since a model *is* fixed up to isomorphism by its state on a Cauchy surface), General Relativity comes out as deterministic after all.

As discussed in Chapter 2, we should not get too worked up about this issue. The haecceitist can admit that the most immediate or natural definition of determinism is one which (combined with their metaphysics) renders GR indeterministic; but, they can insist, this only goes to show that scientists are interested in the alternative, derivative notion of determinism captured by Definition 3. That said, as with the particle-decay world, the anti-haecceitist does have the rejoinder that on *their* metaphysics, there is *no* sense in which GR is indeterministic—not even a "metaphysical" rather than "physical" sense. In truth, however, it is hard to see this dialectic really advancing: both sides, I suspect, will find themselves more than able to massage their intuitions about determinism to fit their preferred thesis about possibilities.

A more substantial concern is that this kind of anti-haecceitist view is not available to anyone who is a spacetime substantivalist. The claim that substantivalism obliges one to affirm the distinctness of permuted models goes right back to Earman and Norton's 1987 presentation of the Hole Argument; after discussing various ways in which we might try to pin down the distinction between substantivalists and relationalists, they remark that

> Whatever reformulation a substantivalist may adopt, they must all agree concerning an acid test of substantivalism, drawn from Leibniz. If everything in the world were reflected East to West (or better, translated 3 feet East), retaining all the relations between bodies, would we have a different world? The substantivalist must answer yes [...] The diffeomorphism is the counterpart of Leibniz' replacement of all bodies in space in such a way that their relative relations are preserved. [...] In sum, substantivalists,

whatever their precise flavour, will deny:

Leibniz equivalence Diffeomorphic models represent the same physical situation.[8]

The idea, then, is that substantivalists are barred from making use of the kind of apparatus laid out above; if one wishes to do so, then one has to be either a relationalist, or else (perhaps) some kind of "structuralist" about space-time points.[9]

One response, at least in the context of the hole argument, is to challenge whether diffeomorphisms are indeed the relevant analogue of the Leibniz shift: surely it is highly contestable whether applying an arbitrary diffeomorphism to all the structure (save the manifold) of a model of GR is indeed analogous to applying a translation (a very particular kind of diffeomorphism) to just the material structure of a model of FNG. However, if our interest is in defending anti-haecceitism, then this response will not suffice; for in both cases, we are dealing with models isomorphic to one another, and so in both cases the anti-haecceitist is committed to the co-representationality of said models. So—if substantivalism is to remain a live option for us—we should say a little about whether it really is committed to the existence of possible worlds that differ only by a static shift or a hole diffeomorphism.

To start, it just really isn't clear why a commitment to the existence of space-time thereby commits one to certain modal theses about how many kinds of possible world might exist, of either the static-shifted or hole-diffeomorphed varieties. The idea, perhaps, is supposed to go something like this. Subtantivalists, it is claimed, should believe that spacetime points are substantial in the sense of having a kind of modal robustness: more precisely, they should hold that spacetime points have at least some non-essential properties, so that the substance can support non-trivial *de re* modal claims.[10] Unless it is essential to each spacetime point that it have a particular set of metrical properties or occupational properties, then each point could have had different such properties; in fact, it seems that each such point could have had precisely the metrical/occupational properties of some other point, and that they could *all* have changed their properties in this way together; and, hence, that there are possible worlds which differ only with regards to which space-time point is which.

This suggests that substantivalism is being understood as individualism (see §2.4) for spacetime points, where individualism is read as the strong view that there are

[8] [Earman and Norton, 1987, pp. 521-522]

[9] [Mundy, 1992], [Dorato, 2000] or [Rickles, 2008] are all self-declared structuralist approaches; see also [Greaves, 2011] for a (somewhat sceptical) overview of spacetime structuralism.

[10] This criterion of substantiality is discussed by [Healey, 1995].

fundamental facts about *particular* spacetime points. As discussed there, it does appear to be true that individualism (in that form) entails haecceitism. But nevertheless, the substance of this criticism is presumably that only generalists are entitled to identify isomorphic models as representing the same possible world. And as we've seen, there doesn't seem to be good evidence for that claim: at least on the face of it, there is nothing incoherent about an ontology of fundamental individuals, which nevertheless fail to have modally robust identities.

That said, it may be worth noting that generalism about spacetime points looks to be somewhat more plausible than generalism about (general) individuals. The relevant technical observation here is that manifolds are individuated (up to diffeomorphism) by their "smooth algebras", where the smooth algebra of a manifold M is the algebra $C^\infty(M)$ of smooth functions on M: given manifolds M and N, there is an $\mathbb{R}$-algebra isomorphism between $C^\infty(M)$ and $C^\infty(N)$ iff there is a diffeomorphism from M to N.[11] So, intuitively, once we know what the smooth algebra looks like, we know what manifold we are dealing with. If we are dealing with a basic bundle, then all the structures we need are constructible from the manifold. Thus, as [Geroch, 1972] pointed out, they are constructible from the smooth algebra. As a result, one is able to reformulate any basic theory in terms of smooth algebras rather than manifolds: for instance, one can follow Geroch in defining an "Einstein algebra" to be a smooth algebra equipped with a metric and stress-energy tensor satisfying the Einstein field equations.

This line of reasoning is fairly well-known to philosophers of physics, as a result of Earman's claim that one could use Einstein algebras as a way to finesse the Hole Argument: the idea, roughly, is that by taking the algebras as fundamental, one can view distinct but isomorphic models of GR as merely alternative representations of that numerically identical algebra.[12] That claim is rather dubious, given that one can multiply isomorphic algebras just as easily as one can multiply isomorphic spacetime models—and in particular, in such a way that two algebras coincide to within (the algebraic translation of) a "hole".[13] The observation I wish to make here is that taking the smooth algebras as primitive could, instead, be thought of as the appropriate mathematical correlate of taking properties as fundamental. For, after all, the elements of the smooth algebra just are the smooth scalar fields on the manifold, i.e., smooth

[11] [Milnor and Stasheff, 1974]

[12] [Earman, 1989, §9.9]; see also [Earman, 1977], for the application of the same idea to classical spacetimes.

[13] [Rynasiewicz, 1992]

distributions of (scalar) properties over spacetime points. In particular, note that given a smooth algebra, one can explicitly reconstruct the manifold: the points of the manifold are identified with the real maximal ideals of the algebra.[14] This could be interpreted as showing how the derivative ontology of individuals is obtained from the fundamental ontology of general facts.

That said, two quick notes of caution to the aspiring generalist. One is to note that the general facts to which they must commit are modally rich. The algebra of smooth functions comprises *all* such functions, not just those which happen to be realised in some particular model (assuming, that is, that the model contains any scalar fields at all!). So this version of the generalist program reconstructs individuals in terms of all possible general facts, not just the actual general facts. The second is that there are question marks over how, for a generalist, the notion of a smooth algebra is to be defined. A smooth algebra is a ring of a certain kind, but not all rings are smooth algebras. The easiest way to characterise which rings are smooth algebras is by saying "a smooth algebra is a ring which is isomorphic to the algebra of smooth functions over some manifold"—but this is presumably not acceptable to the generalist, since it requires an antecedent notion of manifold. What is required is some more "intrinsic" definition of smooth algebra, but doing so is highly nontrivial.[15] All of this is to say, then, that although there is interesting territory here for the generalist to explore, it is not going to be entirely plain sailing—nor does it seem that only generalists are entitled to take isomorphic models as equivalent.

5.4 External symmetries which are not isomorphisms

What about other external symmetries, however? That is, what about cases where the base-space transformation $\beta : M \to M$ is *not* an automorphism? I claim that in such cases, we should look to a combination of qualitativism *and* anti-quidditism to resolve the issue.

Let us look at a specific example: the fact that Galilean boosts are *not* automorphisms of Newtonian spacetime, but are external symmetries of Galilean electrodynamics. Consequently, applying a (non-trivial) boost to a model $\mathcal{M} = \langle \mathfrak{N}, Y \rangle$ of $\mathbb{T}_G$ yields the non-isomorphic model $\mathcal{M}' = \langle \mathfrak{N}, b_* Y \rangle$. Hence, mere qualitativism will not be enough to obtain the desired conclusion. Nor will anti-quidditism alone suffice. As we saw

[14] See [Rynasiewicz, 1992], [Bain, 2003] or (especially) [Rosenstock et al., 2015] for details.
[15] For overviews, see [Schreiber, 2015b], [Schreiber, 2015a].

in Chapter 3, anti-quidditism enjoins us to reckon as equivalent those models which are intertranslatable with one another: in intuitive terms, in which the individuals in one world instantiate some clutch of nomologically equivalent properties to those they instantiate in the other. A given individual in $\mathcal{M}'$ does not instantiate nomologically equivalent properties to those it instantiates in $\mathcal{M}$. In general, for instance, it might be that $\rho(p) = 0$ but $b_*\rho(p) \neq 0$.

However, the crucial thing to observe is that we can *combine* the pair of doctrines to get something stronger than either taken alone. As the above makes clear, the anti-quidditist injunction that we developed in Chapter 3 still assumes a certain kind of haecceitism. But as anti-haecceitists, we should not be bound by that. We can apply our anti-quidditism provided only that there is *some* means of identifying the individuals represented in $\mathcal{M}$ with those represented in $\mathcal{M}'$—i.e. some counterpart-function κ—such that every individual instantiates a nomologically equivalent set of properties to those instantiated by its counterpart (and the same for tuples of individuals).

And of course, there is a natural candidate for such a counterpart-function: just let $\kappa = b$. Since b is a diffeomorphism (and so, amongst other things, a bijection), then we can use the model $\langle b^*\mathfrak{N}, Y\rangle$ to represent the boosted possibility; i.e., we can use the identity relation as a convenient formal means of representing the counterpart relation under consideration—and as discussed in Chapter 2, we can do so without giving up our anti-haecceitism. Intuitively, rather than pushing the dynamical structure forwards under the diffeomorphism β, we instead pull the base-space structure back. Mathematically, we do not generally bother to distinguish these operations, since they are equivalent: equivalent, here, in the sense that the models$\langle \mathfrak{N}, b_*Y\rangle$ and $\langle b^*\mathfrak{N}, Y\rangle$ are isomorphic. Taking the qualitativism argued for in Chapter 2 in hand, we are perfectly able to follow the mathematicians in regarding these two models as equivalent (regarding the choice between them as merely a choice over which counterpart function from $\langle \mathfrak{N}, Y\rangle$ to use). But the use of the latter model makes vivid that—under this counterpart-function—the difference between $\langle \mathfrak{N}, Y\rangle$ and $\langle b^*\mathfrak{N}, Y\rangle$ is that, for any points $p_1, \ldots, p_n$, they are represented as standing in different spatiotemporal relations in the latter compared to the former.

For example, consider a stationary timelike curve through $\mathfrak{N}$: that is, a curve γ such that the spatial projection of $\gamma(s)$ is the same as that of $\gamma(s')$ for all $s, s' \in I$. The image of this curve is a set of points of $D_{\mathfrak{N}}$, which in $\mathfrak{N}$ is entirely "vertical" but in $b^*\mathfrak{N}$ as a certain "slope". We could say that $\mathfrak{N}$ represents the points as being at absolute rest, whilst $b^*\mathfrak{N}$ represents them as having absolute velocity $\mathbf{v}$. Then the observation is

that these two properties are nomologically equivalent: since they are related by a symmetry, they play the same nomological role in $\mathbb{T}_G$. As a consequence, we are able to bring the apparatus of Chapter 3 to bear, and conclude that whatever differences one might seek between the two models will be subject to our now-familiar set of concerns. Let's look at those in detail.

First, we have the semantic problem. Recall that this took two guises, depending on what kind of account of possible worlds is presupposed. If the possible worlds are something metaphysically above and beyond the models, then the problem is that there is no way of ensuring that verticality in a model really does represent the property of being at absolute rest—rather than, say, the property of having absolute velocity $\mathbf{v}$.[16] (Observe that part of what is doing the work here is the recognition of "having absolute velocity $\mathbf{v}$" as an entirely legitimate property, even though in this context we are supposing it to be defined from the property of absolute rest rather than taken as primitive. This mirrors the fact that, in our argument in Chapter 3, syntactic maps were allowed to map atomic formulae to *complex* formulae, not just to other atomic formulae.)

However, the proponent of Newtonian spacetime might protest that we are just *ignoring* the extra structure in which they believe. After all (they continue), their ontology is one consisting of an absolute space whose points persist through time. Exactly how to cash that out will depend on one's preferred metaphysics of persistence. For perdurantists, there are facts about which spacetime points are temporal parts of one and the same spatial point; for stage theorists, there are facts about which spacetime points are (temporal) counterparts of one another; and for endurantists, there are just facts about what the points of absolute space are like at different times. In general, we can say that there are supposed to be facts to the effect that *this* spacetime point is a latter temporal stage of *that* spacetime point. So (our interlocutor continues) a given model is clearly most suited to represent some particular world: namely, one in which two temporal stages of spatial points are represented by some $(\mathbf{x}, t)$ and $(\mathbf{x}, t')$ if and only if they are temporal stages of the very same spatial point. Or, perhaps we can just *stipulate* that the relationship of "being stages of the same persisting object"—the kind of relationship with which one is acquainted from ordinary cases of persisting physical objects—is to be represented by the mathematical relationship of "sharing a first member." In other words, surely the facts regarding the persistence of the points of absolute space (which we cannot deny without begging the question) provide the

[16] cf. [Healey, 2006], [Peacocke, 2014].

means by which a model could be yoked to one world in particular.[17]

In response, I contend that the problem is not that we are ignoring structure: rather, it is the advocate of Newtonian spacetime who is ignoring structure. Grant the facts about persistence of points. The point is that these facts necessarily bring in their wake facts about "uniform scrolling":[18] facts to the effect that such-and-such a sequence of spatial points, over time, corresponds to some uniform absolute velocity (in the sense that were a point particle to have occupied those points at those times, it would have had that absolute velocity). And the point is precisely that uniformly scrolling points are, *so far as the laws are concerned*, precisely as natural, eligible, etc., as persisting points. Hence, in order to acclaim the persistence structure as more natural than the scrolling structure, one has to appeal to a notion of naturalness that goes above and beyond what is codified in laws of nature.

But by the same token, if one is willing to make such an appeal, then this concern can be circumvented: that enables one to maintain that the property of being at absolute rest stands out as the uniquely natural candidate for the referent of the term "absolute rest", since it is the most fundamental, or the most natural, of all the absolute-velocity properties. Of course, in principle some other absolute velocity could be picked as the fundamental absolute speed property. But it is hard to imagine anyone defending such a view, except as a reductio—especially since absolute rest is, as Maudlin puts it, "the unique absolute velocity compatible with the isotropy of space, the one, so to say, isotropic velocity"[19] (though cf. note 20 below).

Alternatively, suppose that we are working with a conception of possible worlds in which they are abstractions (under an appropriate equivalence relation) from models, in the sense discussed in Chapter 1. Then there is no problem of how one model and its boosted counterpart come to represent distinct worlds: they represent themselves—or rather, they represent their equivalence classes under isomorphism. But again, the concern is that any such model carries implicit commitment not only to absolute rest, but to absolute velocities, since (say) a vector field corresponding to absolute velocity $\mathbf{v}$ can be defined.[20] Once we recognise such properties—and if, contra the proponent of

[17] I owe this objection to Oliver Pooley.

[18] The terminology is meant to evoke the display of scrolling text on LED displays, via sequential lighting.

[19] [Maudlin, 1993, pp. 192–193]

[20] A wrinkle: surely, given the isotropy of Euclidean space, there is no way of defining these other vector fields uniquely? However, bear in mind that we are already working in the context of having selected some particular direction in which to apply the boost b. That lets us break the isotropy to define the vector field corresponding to absolute velocity $\mathbf{v}$ (indeed, recall that $\mathbf{v}$ is just the velocity of the applied boost).

extra-nomological naturalness, we take them as being on a par with absolute rest—then the only feature distinguishing the two models is that what one labels with a 0, the other labels with a $\mathbf{v}$. If we are abstracting worlds from models, then we ought (at least) to abstract away from arbitrary choices of labelling; and hence, we ought to abstract the distinction between these models away, insofar as we are regarding them as representing worlds.

Second, there is the epistemic concern. Again, this closely parallels what we have seen before—and in particular, the travails of attempts to attain knowledge of quiddities through the use of a noumenometer. A putative knower, armed with what is supposed to be an absolute speedometer, finds herself facing the following dilemma.[21] If the absolute speedometer's output is invariant under boosts, then it is unreliable: if it gives the right answer in world W, then it gives the wrong answer in any world obtained from W by a boost. But if the absolute speedometer's output is not invariant under boosts, then it is inscrutable: only if the knower has direct perceptual access to the boost-variant facts (i.e. the absolute velocities) would she be able to read the speedometer's output. Given the obscurities about how perceptual qualia arise from the physical substrate of brain activity, one can consistently postulate that experiences of different absolute velocities issue in different phenomenological experiences—but only at the cost of making those experiences entirely internal, and incapable of manifesting in external behaviour.[22]

Finally, we turn to the case of determinism. We have already seen that the Hole Argument provides us with an illustration of how external symmetries can generate indeterminism unless we commit to regarding isomorphic models as equivalent. Now, we show that unless we regard models related by an external symmetry as equivalent, the Hole Argument can be straightforwardly resurrected. The trick is just to do the Hole Argument in coordinate-based terms. So, suppose that we were doing GR on some local region U, small enough that we can cover it by a single coordinate chart $x^\mu : U \to \mathbb{R}^4$: this means that we can represent any metric by a function $g_{\mu\nu} : x^\mu[U] \to \mathbb{R}^2$, and any stress-energy tensor by a function $T_{\mu\nu} : x^\mu[U] \to \mathbb{R}^2$ (where $x^\mu[U] \subset \mathbb{R}^4$ is the image of U under x^μ). So suppose that $g_{\mu\nu}$ and $T_{\mu\nu}$ are a pair of such functions solving the Einstein field equations, as expressed in x^μ; for the sake of simplicity, let's suppose

[21] The below is based directly on [Roberts, 2008]; see also [Dasgupta, 2014b].

[22] Note that even if you, dear reader, have never experienced such qualia, you don't know that no-one else does. Certainly, the fact that no-one else has ever described such experiences, and will deny having had them if asked, is no evidence: for *ex hypothesi*, these experiences result in behaviour that is no different to that resulting from the absence of those experiences. (In effect, of course, this is just to stress the utter bizarreness of what is being postulated here.)

that $T_{\mu\nu} = 0$, so that $g_{\mu\nu}$ is a solution to the vacuum field equations.

Now, we consider a different coordinate chart $\widetilde{x}^\mu$ over U, with the following property: except for some compact $H \subset U$, $\widetilde{x}^\mu = x^\mu$. We know, of course, that the metric formerly represented by $g_{\mu\nu}$ will now be represented by a *new* function $\widetilde{g}_{\mu\nu} : \widetilde{x}^\mu[U] \to \mathbb{R}^2$. Next, observe that $x^\mu[U] = \widetilde{x}^\mu[U]$ (since the two coordinate systems agree on the boundary of H). So now fix on the coordinate system x^μ, and consider the function $\widetilde{g}_{\mu\nu}$ *as a function* $x^\mu[U] \to \mathbb{R}^2$ (in more intuitive terms, ask what gravitational field is represented by that functions in the coordinate system x^μ). The relevant observation is just that $\widetilde{g}_{\mu\nu}$ is a different *and non-isomorphic* function to $g_{\mu\nu}$—but at all points outside of H, $g_{\mu\nu} = \widetilde{g}_{\mu\nu}$. Now, we know that $\widetilde{g}_{\mu\nu}$ solves the vacuum field equations when those equations are expressed in $\widetilde{x}^\mu$. But given that the equations are generally covariant, that means that $\widetilde{g}_{\mu\nu}$ solves the vacuum field equations when those equations are expressed in x^μ. Thus, we have two non-isomorphic solutions to the field equations, $g_{\mu\nu}$ and $\widetilde{g}_{\mu\nu}$, which agree everywhere except the region H (which can, as in the regular Hole Argument, be made arbitrarily small). So unless $g_{\mu\nu}$ and $\widetilde{g}_{\mu\nu}$ in fact represent the same gravitational field, we are confronted with radical indeterminism. This gives us good reason to seek to interpret them in this way—i.e., to reject the structural differences between them as being surplus to what is required.

I actually think the Hole Argument is more illuminating when given in this form, since it looks considerably more compelling than in more modern formulations; certainly, it cannot be dissolved merely by appeal to the dubiousness of distinguishing isomorphic models.[23] It is also historically interesting, since it was in this form that Einstein first encountered it.[24] These two observations are connected, since seeing this presentation goes some way to explaining how Einstein could have been led astray by the Hole Argument. The modern, coordinate-free presentation of General Relativity is valuable precisely because it attributes no more structure to the base space than is appropriate—which in this case, means attributing to it no more structure than the structure of a (mere) differential manifold. But that presentation is only permissible insofar as we think that this *is* the appropriate amount of structure to attribute to the base space, i.e., insofar as we are confident that the theory may be interpreted as not committed to any of the structure coded up in the base space $x^\mu[U] \subset \mathbb{R}^4$ (except its smoothness structure). In other words, presenting the theory in a coordinate-free way

[23] This serves to refute [Weatherall, 2016]'s claim that the Hole Argument can be dealt with *merely* by attention to the fact that mathematical representation is only as fine-grained as isomorphism—for, not all versions of the Hole Argument require a more fine-grained level of representation than that.

[24] My presentation above essentially follows [Norton, 1993, §3.3].

is not a means of avoiding the interpretational problem raised by the (coordinate-based) Hole Argument; it is, rather, a means of implementing the solution to it.

5.5 Determining spacetime structure

Thus, as with internal symmetries, external symmetries should be taken to relate representations of the same possibility. Now, in the previous chapter I defended the coherence of doing so externally—i.e., of simply declaring that the relevant external symmetries are to "count" as isomorphisms. However, as discussed there, there are also useful insights to be had from an internal presentation of the structure to which one is committed, following such an interpretational stipulation. To do so is to build up the model structure from more primitive ingredients (rather than obtaining it by stripping down some richer structure), so as to make manifest that the symmetries in question are isomorphisms.

Clearly, we need not worry about doing so for the first kind of case discussed—those in which the symmetry is an automorphism of the base space. Moreover, this makes it clear what an internal (i.e. Riemannian rather than Kleinian) presentation of the structure of our sophisticated theory will involve: a necessary and sufficient condition for internally presenting such a theory, at least in the case of a basic theory, is that we set it on a base space which is invariant under the relevant external symmetries.

For instance, in the case of $\mathbb{T}_G$, we are looking for a base space whose structure is invariant under Galilean boosts; or, slightly more accurately, whose structure comprises the substructure of Newtonian spacetime which is invariant under Galilean boosts (and hence, under the Galilei group as a whole). It is well-known what kind of a base space this is: it is *Galilean spacetime*. Define a Galilean spacetime $\mathfrak{G}$ to be a four-dimensional affine space, equipped with the following data:[25]

- A privileged three-dimensional *spatial subspace* $V_{\mathfrak{X}}$ of $V_{\mathfrak{G}}$
- A Euclidean inner product on $V_{\mathfrak{X}}$
- An inner product on the quotient space $V_{\mathfrak{G}}/V_{\mathfrak{X}}$

Galilean spacetime provides the minimal structure required for the theory (5.5).

[25] This follows [Saunders, 2013].

Similarly, in the case of $\mathbb{T}_A$, we want the base space to have the Lorentz-invariant substructure of Newtonian spacetime. Again, it is well-known what kind of mathematical structure fits the bill: *Minkowski spacetime*, $\mathfrak{M}$, which consists of a four-dimensional affine space equipped with a Lorentzian inner product. To see that this is indeed a substructure of Newtonian spacetime, observe that given a Newtonian spacetime $\mathfrak{N}$, a Lorentzian inner product $\eta : V_{\mathfrak{N}} \times V_{\mathfrak{N}} \to \mathbb{R}$ is definable by

$$\eta\left(\langle \boldsymbol{\tau}, \boldsymbol{\xi} \rangle, \langle \boldsymbol{\tau}', \boldsymbol{\xi}' \rangle\right) = \delta_{\mathfrak{T}}(\boldsymbol{\tau}, \boldsymbol{\tau}') - \delta_{\mathfrak{X}}(\boldsymbol{\xi}, \boldsymbol{\xi}') \tag{5.7}$$

where $\delta_{\mathfrak{T}}$ and $\delta_{\mathfrak{X}}$ are the Euclidean inner products on $V_{\mathfrak{T}}$ and $V_{\mathfrak{X}}$ respectively. (Note that the fact that we can define a Lorentzian inner product on $V_{\mathfrak{N}}$ is not very surprising, given how richly structured Newtonian spacetime is. It would equally be possible to define a Euclidean inner product on it, by

$$\delta_{\mathfrak{N}}\left(\langle \boldsymbol{\tau}, \boldsymbol{\xi} \rangle, \langle \boldsymbol{\tau}', \boldsymbol{\xi}' \rangle\right) = \delta_{\mathfrak{T}}(\boldsymbol{\tau}, \boldsymbol{\tau}') + \delta_{\mathfrak{X}}(\boldsymbol{\xi}, \boldsymbol{\xi}') \tag{5.8}$$

The difference is merely that the Lorentzian inner product has the interesting feature of being invariant under Lorentz transformations, whilst the Euclidean inner product is not.)

It is often remarked upon that the transition to Minkowski spacetime is more naturally done from Newtonian spacetime (which permits the definition of the Lorentzian inner product) than Galilean spacetime (which does not). Sometimes, moreover, this is presented as a striking or surprising fact: Earman, for example, comments that "[The definability of η within $\mathfrak{N}$] suggests, paradoxically, that absolute space provides a stepping stone from classical to relativistic space-times".[26] One aim of the above analysis is to draw out why this should be so (and diffuse any air of paradox) by making clear how Newtonian spacetime is, in a certain sense, common heritage for both Galilean and Minkowski spacetime. Both of the latter result from stripping away a certain amount of structure from the former, although in different ways. (Sufficiently different, indeed, that to ask whether pre-relativistic or relativistic spacetime has "more structure" does not seem like a well-posed question: their levels of structure are more or less incommensurable.) [27] The reason why they have this common heritage,

[26] [Earman, 1989, p. 34]

[27] [Barrett, 2015b] also makes these observations, and provides a rigorous proof that $\mathfrak{G}$ and $\mathfrak{R}$ are incommensurable. The essential point is that the Newton group is a subgroup of both the Galilei and Poincaré groups, but neither of the latter is a subgroup of the other; hence, the Galilei-invariant spacetime $\mathfrak{G}$ and the Poincaré-invariant spacetime $\mathfrak{R}$ can each be embedded in the Newton-invariant

moreover, is that both Newtonian and relativistic dynamics have the Newton group as a dynamical symmetry group. (It was, in part, to stress this dependence on dynamics that I mentioned the fact that a Euclidean inner product may be defined on $\mathfrak{N}$ just as easily as a Lorentzian one. The reason we retain the latter rather than the former is nothing to do with the mathematics of $\mathfrak{N}$, but is simply because the latter but not the former is invariant under the dynamical symmetries of a theory that we take seriously.)

However, one might think that the above analysis engenders a paradox of its own. For it can seem a little as though we have simply pulled a rabbit out of a hat here: surely it doesn't follow *just* from Maxwell's equations that spacetime has a Minkowskian structure, rather than a Newtonian (or Galilean) one? After all, the view that spacetime had a non-Minkowskian structure certainly persisted after the acceptance of Maxwell's equations—that is what made Einstein's postulation of special relativity such a profound scientific achievement! Furthermore, this view led to concrete empirical predictions: predictions which were refuted by (*inter alia*) the Michelson-Morley experiment, but need not have been. Surely if Michelson and Morley had not got a null result, we would have learned that spacetime was not Minkowskian, the validity of Maxwell's equations notwithstanding?

As a first step towards untangling this issue, we should start by asking what model of electromagnetism is an appropriate way of representing the theoretical commitments of the pre-relativistic understanding of Maxwell's equations. The consensus is that the right answer to this question is so-called "classical electromagnetism":[28] electromagnetism, as governed by Maxwell's equations, but set on Newtonian spacetime.[29] This, of course, is precisely the formal structure I introduced as *Maxwellian electromagnetism*. Now, I presented this as a theory which, merely by reflection on general matters of anti-quidditism, could be recognised as empirically equivalent to Maxwellian electromagnetism set on Minkowski spacetime. But then how on *earth* is it the case that there is apparently empirical evidence which tells in favour of relativistic electrodynamics against Newtonian electrodynamics? What, on this view, was the content of the Michelson-Morley null result?

The answer is that, as per usual, what was empirically tested were not the bare theories themselves, but the conjunctions of those theories with certain auxiliary hypotheses. In this example, the auxiliary hypotheses concerned the relationship

spacetime $\mathfrak{N}$, but not in one another.

[28] See [Friedman, 1983, §III.5], or [Earman, 1989, §3.5].

[29] Note that this theory could *not* be set on Galilean spacetime, since Galilean spacetime lacks the Minkowski structure needed for the formulation of the dynamics (again, cf. [Barrett, 2015b]).

between the spacetime structure and the behaviour of rigid *mechanical* objects. In particular, pre-relativistic physics assumed that if a rigid rod's equilibrium state when at rest with respect to the ether is such as to occupy a region of length L, then its equilibrium state when in motion with respect to the ether is still to occupy a region of length L. In other words, what was refuted by the Michelson-Morley experiment was not merely electrodynamics set upon Newtonian spacetime, but that theory *plus* a number of assumptions about how Newtonian spacetime bore on the mechanics of rigid bodies. In the terms of this discussion, this amounts to assuming that the dynamics of rigid bodies are governed by dynamics exhibiting Galilean symmetry, not Lorentz symmetry. As a consequence, the *combined* theory, including electromechanical coupling, exhibited Newtonian symmetry (since the Newton group is the common subgroup of the Lorentz and Galilei groups).

This analysis gives us the resources to explain the profundity of Einstein's contribution. As various historical treatments make plain, other researchers besides Einstein came, in many ways, extremely close to articulating their own versions of relativity.[30] What separated Einstein from these precursors? The standard view is that although many relativistic phenomena (length contraction, the phenomenon of local time, and even the variation of mass with energy)[31] were anticipated, Einstein was the first to characterise these as "kinematical rather than dynamical"—in the sense of arising from the nature of space and time, rather than merely as curious features of electromagnetic dynamics. (This usage of "kinematical/dynamical" is only loosely related to our terminology of kinematics and dynamics.) There is clearly a sense in which this is correct, but the discussion so far suggests a way in which we can flesh out what is meant by this. The essential core of Einstein's contribution was to postulate the *universality* of these effects: to propose, that is, that *all* physical phenomena were subject to Poincaré symmetries. What had persisted after the acceptance of Maxwell's equations (even in the thinking of Lorentz, Poincaré, et al.) was Newtonian mechanics, in the specific sense of a belief that mechanics was governed by Newton's equations. Once electromechanical coupling is taken into account, this view entailed that mechanics was governed by equations whose symmetry group was the Newton group: if mass is held constant, then the Lorentz force law (4.8) is only invariant under transformations of the form (Newt), not under transformations of the form (Gal) or (Poin).

[30] See, in particular, [Pais, 1982, chaps. 6–8] and [Brown, 2005, chap. 4].

[31] Length contraction was proposed, as is well known, independently by FitzGerald and Lorentz. The notion of "local time" was discussed by Lorentz and Poincaré. Finally, so-called "electromagnetic mass", which varied with the electromagnetic energy of its bearer, was a widely used notion.

By contrast, the first of Einstein's postulates (the relativity principle) required that whatever dynamics was taken to govern electromechanical phenomena, it must manifest some form of boost symmetry; and the second postulate (the light postulate) served to ensure that the specific *variety* of boost symmetry would be that of Lorentz boosts rather than Galilean boosts. Hence, we can understand Einstein's role as being to assert that the Lorentz symmetry of electromagnetism extends to all physical phenomena—and in particular, to mechanical phenomena. Why is it apt to characterise this as an assertion about spacetime? One answer is that paradigmatic chronogeometric phenomena, i.e. rods and clocks, are mechanical entities. This kind of view is suggested by the following remark of Einstein:

> The theory to be developed is based—like all electrodynamics—on the kinematics of the rigid body, since the assertions of any such theory have to do with the relationships between rigid bodies (systems of co-ordinates), clocks, and electromagnetic processes. Insufficient consideration of this circumstance lies at the root of the difficulties which the electrodynamics of moving bodies at present encounters.[32]

I would prefer a slightly different way of putting things: rather than pre-theoretically privileging some class of entities as being rods or clocks, we should instead say that the characteristic feature of something's being *spacetime* structure—as opposed to material structure—is its pervasiveness, the fact that all dynamical processes take notice of it.[33] So asserting the Lorentz transformations as a spacetime symmetry group, as Einstein did, serves as a means of asserting Lorentz symmetry as a universal symmetry. Of course, if all phenomena manifest Lorentz symmetry then rods and clocks (whatever they might be) will do so. Conversely, it is plausible that something only gets to count as a rod or a clock if it manifests an appropriate kind of universality. That is, clocks are deemed good insofar as they synchronise diverse phenomena; and more generally, using one phenomenon to measure diverse others will only be helpful to the extent that the measuring phenomenon remains invariant under transformations that leave the measured phenomena invariant. So we can see these two ways of singling out spacetime structure as being complementary, rather than competitive.

[32] [Einstein, 1905, p. 892]; p. 2 of the translation.

[33] This claim—that spacetime structure is picked out by being the universal mediator of dynamical interaction—is due to Knox ([Knox, 2011], [Knox, 2013], [Knox, 2014]). It offers one attractive way of unpacking the "dynamical approach" to spacetime theories of [Brown, 2005].

5.6 Against purity

There is one final wrinkle that could do with being cleared up. This concern is that there is an alternative way of producing a theory in which the external symmetry in question is rendered as an isomorphism: rather than *eliminating* the surplus base-space structure, instead "promote" it to lie amongst the dynamical structure. For example, one could replace $\mathbb{T}_G$ by a theory $\mathbb{T}'_G$ whose models are of the form $\langle \mathfrak{G}, \sigma^a, Y \rangle$, where σ^a is a constant timelike vector field (i.e., one such that $\nabla_b \sigma^a = 0$, where ∇ is the derivative operator associated with $\mathfrak{G}$). In effect, we have "split" Newtonian spacetime into its boost-invariant part $\mathfrak{G}$, and its boost-variant part σ^a (which represents the tangents to the trajectories of the points of absolute space). Now, it is straightforwardly the case that the models of this new theory are sophisticated under boosts. Yet intuitively, we have precisely as much structure here as we did before: so we want to resist the idea that this "counts" as a means of taking on board the lesson of the symmetry.

An especially good reason to do so is that it isn't even necessary that the external transformation in question be a symmetry at all! That is, suppose that $\gamma : M \to M$ is some diffeomorphism under which M is *not* invariant, but that M admits of being "factorised" in the manner above: that is, we have some pair $\langle N, H \rangle \cong M$, where N is invariant under γ. Then by treating H as bundle-data rather than as a component of the base space (i.e., by "promoting" it in the same way we considered promoting σ^a above), we obtain an intuitively equivalent theory such that for any model $\mathcal{M}$, $\gamma^* \mathcal{M} \cong \mathcal{M}$. In other words, we seem to be typically able to reformulate the theory so as to ensure that γ is an external symmetry—and, indeed, that it is a symmetry relating isomorphic models. The most general way of doing such a split is to have N be a *mere* manifold, and hence invariant under arbitrary diffeomorphisms. By doing so, we can produce (for any given theory) an equivalent theory which is generally covariant: that is, the Kretschmann objection is a special case of this problem.

However, we should think carefully about why this might be considered a problem at all. One concern is that in the passage from $\mathbb{T}_G$ to $\mathbb{T}'_G$, some important information is lost: namely, the fact that boosts are an external symmetry of the theory. A closely related thought is that promotion shouldn't count as elimination: i.e., we shouldn't think that changing to $\mathbb{T}'_G$ is sufficient to deal with the problems raised by taking symmetry-related models to represent distinct possibilities. In a nutshell: given that $\mathbb{T}'_G$ is equivalent to $\mathbb{T}_G$, where has the boost-symmetry got to in $\mathbb{T}'_G$? The answer to this question is to recognise that the boost symmetry is entirely present, and indeed is

present as an external symmetry: it's just that it is no longer a *pure* external symmetry. Recall that a pure external symmetry is when we lift a base-space diffeomorphism to the most natural bundle automorphism on every background bundle relevant to the theory. But as I was keen to emphasise above, what is really important here is the bundle automorphisms themselves; it is of secondary importance that we can (when dealing with basic theories) "abbreviate" a collection of basically agreeing bundle automorphisms by their common base-space action (and conversely, can use any base-space diffeomorphism to generate a specific collection of bundle automorphisms).

So the external symmetry has not been lost, provided that we can find some bundle automorphism which manifests it. And indeed, such will be the case. For suppose that $H \xrightarrow{\pi} M$ is the bundle we have used to encode what was, in the original theory, base-space data: so in the context of $\mathbb{T}'_G$, H is a copy of the tangent bundle, a timelike section σ^a of which is used to represent absolute space. Then consider the following collection of bundle automorphisms: we act on H as the identity, and on all background bundles other than H with the lift of some chosen diffeomorphism β. We then observe that this transformation is a symmetry transformation iff β is a member of the Galilei group. From this, the analysis outlined above can proceed as before. Indeed, note that what we have effectively done is to treat the bundle H as though it were part of the base space. The fact that it is not "officially" part of the base space means that this involves a bit more rigmarole, but without—so far as I can see—causing any insurmountable obstacles.

This gives rise to a worry that it is underdetermined what the external symmetry group of a theory is: or rather, that there is no way of identifying an external symmetry group for a theory which is invariant under reformulating the theory. Although I sympathise with the worry, I think it is just true that theories lack unique external symmetry groups (at least in the robust sense desired by the worrier). In particular, Kretschmann was simply *right* that given any theory, we may find an equivalent theory manifesting general covariance—so having a formulation manifesting general covariance is not an interesting property of a theory. This is not to say that there are not interesting properties in the neighbourhood! For instance, it may be important that a theory admits of a formulation in which it does *not* manifest general covariance. Or we could consider a theory's external symmetry group in a formulation which has a particularly nice base space: e.g. one in which all "absolute" dynamical structures are codified into the base space.[34] Or we could try to explicate the notion of background-independence,

[34]I have in mind here the Anderson-Friedman conception of an absolute object ([Anderson, 1967],

as a more robust property of theories than general covariance.[35]

It does mean, however, that identifying interesting external symmetry groups requires more subtlety than simply looking at which base-space diffeomorphisms (of a basic theory) lift to symmetries of the theory as a whole. In particular, just because a theory employs a mere manifold M as its base space, we should not immediately conclude that it merely has $\mathrm{Diff}(M)$ as its only interesting external symmetry group. Indeed, it seems plausible that this lesson may hold good for General Relativity. If the theory is formulated in terms of tetrads, then the local Poincaré group emerges as an interesting group of symmetries, in a way that is not true on the more standard manifold-plus formulation.[36] I lack both the space and the expertise to discuss this fully here; but it does seem to me helpful to insist that we look at impure external symmetry groups just as much as pure external symmetry groups. In fact, the next chapter is about a particularly important impure external symmetry; let us turn to it.

[Friedman, 1983]; however, see [Pitts, 2006] for an important critique of that program.

[35] See [Belot, 2011], [Pooley, 2015].

[36] [Wallace, 2015a]

6 Maxwell-Cartan gravitation

> It's not the fall that kills you; it's the sudden stop at the end.
>
> Attributed to Douglas Adams

The following two observations are well-known to philosophers of physics:

1. Newtonian gravitation admits, in addition to the well-known velocity-boost and potential-shift symmetries, a "gravitational gauge symmetry" in which the gravitational field is altered.

2. Newtonian gravitation may be presented in a "geometrised" form,[1] in which the dynamically allowed trajectories are the geodesics of a non-flat connection.

Moreover, it is widely held that these two observations are intimately related. However, aspects of this relationship remain somewhat obscure. In particular, there is widespread disagreement over the sense in which the symmetry of observation 1 motivates the move from a non-geometrised formulation to the geometrised formulation of observation 2; and over the extent to which such motivation ought to be regarded as analogous to the use of the velocity-boost symmetry to motivate the move from Newtonian to Galilean spacetime, or to the use of the potential-shift symmetry to motivate the move from a formulation in terms of gravitational potentials to a formulation in terms of gravitational fields.

In this chapter, I seek to clarify this relationship. In §6.1, I introduce some preliminary mathematical notions. In §6.2 I introduce Newtonian gravitation, set on Galilean spacetime and in terms of gravitational fields, and present the gravitational gauge symmetry referred to in observation 1. In Section §6.3 I present Newton-Cartan theory, the geometrised formulation of Newtonian gravitation referred to by observation 2, and discuss how it relates to the gravitational gauge symmetry. In §6.4, I consider

[1] Due originally to [Trautman, 1965].

some reasons (highlighted in [Saunders, 2013]) for being confused about the way in which Newton-Cartan theory relates to the gravitational gauge symmetry, and the desirability of a gravitational theory set on Maxwellian spacetime. I then discuss [Weatherall, 2015c]'s suggestion for how to construct such a theory, and indicate some ways in which it is not as perspicuous as we might wish. §6.5 contains the main contribution of this chapter: the specification of a gravitational dynamics, set upon Maxwellian spacetime, which bears a particularly perspicuous relationship to Newton-Cartan theory and to the gravitational gauge symmetry. §6.6 concludes.[2]

6.1 Leibnizian spacetime

All of the spacetime structures we will be considering contain at least as much structure as *Leibnizian spacetime*.[3] Such a spacetime may be defined as a structure comprising the following data:[4]

- A differential manifold M, which we take to be diffeomorphic to $\mathbb{R}^4$
- A smooth, curl-free 1-form t_a on M
- A smooth, symmetric $(0,2)$-rank tensor h^{ab} on M; this is required to be *flat*

subject to the *orthogonality condition*

$$t_a h^{ab} = 0 \tag{6.1}$$

We will denote such a structure by either $\mathfrak{L}$ or $\langle M, t_a, h^{ab}\rangle$. A Leibnizian spacetime contains enough structure to permit judgments regarding the continuity and smoothness of spatiotemporal paths or regions; the temporal distance between any two events; and the spatial distance between any two simultaneous events. However, it does not contain much structure beyond that.

In particular, it does not permit judgments regarding the straightness of spatiotemporal paths. That kind of structure is represented by an *affine connection* ∇ on the base

[2] See also [Wallace, 2015b] for an alternative analysis of these issues.

[3] [Earman, 1989, chap. 2]

[4] Note that I'm using "a spacetime" here to mean a mathematical structure which is apt to represent some kind of physical spacetime. Although this terminology is a little unfortunate, it is sufficiently convenient and standard to be worth cooperating with. We just have to be careful to avoid assuming that all and only the structure in a spacetime is best interpreted as representing spatiotemporal structure (this will come up at p. 142 below).

manifold M. Such a connection provides (roughly speaking) a way of differentiating tensor fields on the manifold, taking any tensor field $T^{a_1\ldots a_m}_{b_1\ldots b_n}$ to a tensor field $\nabla_c T^{a_1\ldots a_m}_{b_1\ldots b_n}$. The main result we will need from the general theory of affine connections is the following:

Proposition 10. [Malament, 2012, Proposition 1.7.3] Let ∇ and ∇' be derivative operators on the manifold M. Then there exists a smooth symmetric tensor field C^a_{bc} on M that satisfies the following condition for all smooth tensor fields $T^{a_1\ldots a_r}_{b_1\ldots b_s}$ on M:

$$\begin{aligned}\nabla'_c T^{a_1\ldots a_r}_{b_1\ldots b_s} =& \nabla_c T^{a_1\ldots a_r}_{b_1\ldots b_s} \\ &- C^{a_1}_{cn} T^{na_2\ldots a_r}_{b_1\ldots b_s} - \cdots - C^{a_r}_{cn} T^{a_1\ldots a_{n-1}n}_{b_1\ldots b_s} \\ &+ C^{n}_{cb_1} T^{a_1\ldots a_r}_{nb_2\ldots b_s} + \cdots + C^{n}_{cb_s} T^{a_1\ldots a_r}_{b_1\ldots b_{s-1}n}\end{aligned} \tag{6.2}$$

Conversely, given any derivative operator ∇ on M and any smooth symmetric tensor field C^a_{bc} on M, if ∇' is defined by equation (6.2), then ∇' is also a derivative operator on M.

Proof. See [Malament, 2012, pp. 51–52]. □

The field C^a_{bc} plays a role precisely analogous to that of the vector potential A_a in the theory of connections on fibre bundles. For this reason, we will refer to it as the *affine vector potential* (of ∇' relative to ∇). It is closely related to the Christoffel symbols. Given a coordinate chart $\phi : U \subseteq M \to \mathbb{R}^4$, let ∂ be the affine connection naturally induced on U by ϕ (i.e., the pullback to U, by ϕ, of the canonical affine connection on $\mathbb{R}^4$). In the coordinate chart ϕ, the Christoffel symbols for an arbitrary connection ∇ on U are then the components, in ϕ, of the affine vector potential of ∇ relative to ∂.[5] The advantage of working with affine vector potentials, rather than Christoffel symbols, is that they are coordinate-independent. When C^a_{bc} is the affine vector potential of ∇' relative to ∇, we will write $\nabla' = (\nabla, C^a_{bc})$.

[5] This may sound puzzling: how can they be the components of the affine vector potential (a tensor field), when every schoolchild knows that the Christoffel symbols are not the components of a tensor field? The answer is that the Christoffel symbols of ∇ in a different chart, ϕ', are the components in ϕ' of the affine vector potential of ∇ relative to ∂' (*not* ∂), where ∂' is the affine connection induced on U by ϕ'. In other words, the Christoffel symbols are (as it were) coordinate-dependent twice over: the choice of chart affects both which tensor field is the affine vector potential of interest, and (as ever) what the components of that field are.

We will only consider affine connections on $\mathfrak{L}$ which are *compatible* with the temporal and spatial metric fields, i.e., which satisfy the compatibility conditions

$$\nabla_a t_b = 0 \tag{6.3a}$$

$$\nabla_a h^{bc} = 0 \tag{6.3b}$$

Note that these equations parallel the defining equation $\nabla_a g_{bc} = 0$ for the Levi-Civita connection on a metric space $\langle M, g_{ab} \rangle$; however, unlike the Levi-Civita connection, there is not a *unique* connection satisfying (6.3) for a given Leibnizian spacetime $\mathfrak{L}$. Given our earlier condition that h^{ab} be flat, it turns out that every compatible connection is *spatially flat*: that is, for any compatible connection ∇ with curvature tensor $R^a{}_{bcd}$, $R^{abcd} = 0$.

A connection not only characterises which curves are straight and which are not: it precisely quantifies the deviation from straightness. More precisely, if θ^a is a unit timelike vector field, then the *acceleration field* of θ^a (representing the acceleration of particles whose worldlines are the integral curves of θ^a) is

$$\theta^n \nabla_n \theta^a \tag{6.4}$$

θ^a is a *geodesic field* just if its acceleration field vanishes. A *geodesic* is an integral curves of some geodesic field.

Furthermore, a connection characterises the rotation of the integral curves of a given vector field. Without going into full details, I will remark only that a unit timelike field θ^a is said to be *non-rotating* or *twist-free*, relative to a (compatible) connection ∇, if

$$\nabla^{[a} \theta^{b]} = 0 \tag{6.5}$$

6.2 Galilean gravitation

As indicated above, Leibnizian spacetime lacks sufficient spacetime structure to be the backdrop for a gravitational dynamics. The first dynamics we consider supplements Leibnizian spacetime with an affine structure, to differentiate between straight and curved spacetime paths. This is precisely the structure of *Galilean spacetime*, which we met in the previous chapter; it may equivalently be defined as comprising[6]

[6] [Earman, 1989, chap. 2]

- A Leibnizian spacetime $\mathfrak{L}$
- A flat affine connection $\overline{\nabla}$[7] which is compatible with $\mathfrak{L}$

We can now define our first gravitational theory.

Definition 8. A model of *Galilean gravitation* comprises a Galilean spacetime $\langle \mathfrak{L}, \overline{\nabla} \rangle$, equipped with

- A scalar field $\mu : \mathfrak{L} \to \mathbb{R}$
- A spacelike vector field G^a
- A maximal class Ξ of unit timelike vector fields

satisfying the following equations:

$$\overline{\nabla}_a G^a = -4\pi\mu \tag{6.6a}$$

$$\overline{\nabla}^{[c} G^{a]} = 0 \tag{6.6b}$$

$$\xi^n \overline{\nabla}_n \xi^a = G^a \tag{6.6c}$$

for every $\xi^a \in \Xi$.

The scalar field μ represents the mass density, the vector field G^a represents the gravitational field, and the set Ξ represents all possible tangents of possible test particles. Given a model of Galilean gravitation, we will refer to curves through $\mathfrak{L}$ which are integral curves of some member of Ξ as *dynamically allowed trajectories*. When I say that the class Ξ is *maximal*, I mean that it contains every field satisfying condition (6.6c) (so that we get all the possible curves that a test particle could take). The gravitational field is related to the mass density by equation (6.6a) (the source equation for this theory), whilst the dynamically allowed trajectories are fixed by equation (6.6c). Note that I have chosen to work with a gravitational field, rather than the gravitational potential. This is simply in order to remove the gauge symmetries of the potential, so that we can focus on those symmetries that alter the field itself. The condition (6.6b) ensures that this decision is harmless: it holds of G^a if and only if there is a scalar field φ such that $G^a = \overline{\nabla}^a \varphi$.[8]

[7] As a notational mnemonic, I will use an overline to indicate a flat connection.

[8] See [Malament, 2012, Proposition 4.1.6]. Note that this is precisely analogous to the role played by the equation $\nabla \times \mathbf{E} = 0$ in the theory (4.11).

The first remark to make about this theory is that it is somewhat unphysical: less so than the electromagnetic theory (4.6) (since it at least contains an equation of motion, (6.6c)), but still fairly bad. The mass density μ is just represented as a phenomenological background, in the sense that there is nothing constraining the motion of the matter whose density μ allegedly represents—in particular, nothing requiring that that it follow a dynamically allowed trajectory. (In fact, there isn't even anything in the models which can be identified as representing the motion of the matter comprising μ.) This has a number of counter-intuitive consequences. One is that if a model contains some region with vanishing mass density, it will nevertheless be threaded by dynamically allowed trajectories—the test particles do not contribute to the local mass density. Indeed, note that (because gravitational coupling is universal) we have not even had to assign the test particles any mass at all!

Moreover, it means that there is a substantial amount of underdetermination in the dynamics. In particular, fixing a mass density μ on a Galilean spacetime $\langle \mathfrak{L}, \overline{\nabla} \rangle$ does not uniquely determine the allowed trajectories, as the following proposition demonstrates.

Proposition 11. Let $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a, \Xi \rangle$ be a model of Galilean gravitation, and consider any spacelike field η^a such that $\nabla^a \eta^b = 0$. Then $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a + \eta^a, \Xi' \rangle$ is a model of Galilean gravitation, where Ξ' is defined by the condition that for any unit timelike field ξ'^a on $\mathfrak{L}$,

$$\xi'^a \in \Xi' \text{ iff } (\xi'^n \overline{\nabla}_n \xi'^a - \eta^a) \in \Xi \tag{6.7}$$

Proof. Since $\nabla^a \eta^b = 0$, $\nabla_a \eta^b = t_a \theta^n \nabla_n \eta^b$, where θ^n is any future-directed unit timelike field; it follows that $\nabla_a \eta^a = 0$.[9] The proposition immediately follows. □

This also means that the theory is indeterministic: by letting $\eta^a = 0$ prior to some arbitrary time, and then smoothly increasing thereafter, we can make the two models agree up to that time but disagree thereafter. (Such an η^a is permissible, for the condition $\nabla^a \eta^b =$ requires only that η^a be spatially constant at each time, not that it be constant over time.) One small remark, on why this should be considered genuine underdetermination (and hence, genuine indeterminism): note that the change from $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a, \Xi \rangle$ to $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a + \eta^a, \Xi' \rangle$ is *not* a symmetry transformation. A symmetry transformation ought to be specificiable in terms of systematic transformations of the structures in the theory, and there is no way—at least, no obvious way—to characterise

[9] This observation is adapted from [Malament, 2012, p. 277].

Ξ' as arising from the application of a systematic transformation to the members of Ξ.[10] This also points up the benefits of representing the test-particle trajectories explicitly in the model, even though they play a very minimal dynamical role: had they not been there, then we would have a (formal) symmetry transformation to hand, and might have been tempted to interpret the underdetermination as merely apparent.[11] However, although the underdetermination is a genuine feature of the theory, it is important to recognise that it is an artefact of decoupling the mass density μ from the dynamically allowed trajectories. We can block it if we require that the mass density "flows" along some particular dynamically allowed tangent field, ζ^a: if $\zeta^a \in \Xi$, then $\zeta^a \notin \Xi'$, and so the above move no longer lets us construct a second model consistent with the same background data.

Clearly, working with a theory in which we do subject the mass density to appropriate dynamics would have significant conceptual advantages. However, I do not intend to do so in this chapter. In my defence, I offer two considerations. One is just that this more realistic dynamics involves more complicated mathematics; so there is some advantage to working in the simpler case, bearing its defects in mind. The other is that formal presentations of Newtonian gravitation in the literature standardly do not relate the mass density to dynamically allowed trajectories (indeed, many presentations omit the dynamically allowed trajectories altogether).[12] By not doing so either, I make comparison easier—and in particular, can highlight some consequences of making this choice, which might otherwise go unremarked.

Let us now turn to our main topic. The jumping-off point is the presence of a certain kind of symmetry in the theory of Galilean gravitation. However, we have to be a little careful here, for one finds two presentations of the symmetry in the literature. Ultimately this is harmless, since the two ways of presenting it turn out to be more or less equivalent—but showing that to be the case is somewhat non-trivial, and illuminating to work through.

First, some discussions[13] present the relevant symmetry as involving a transformation of the *connection* and the *gravitational field*. More specifically, the symmetry is presented

[10] For example, one might be tempted to try the map $\xi^a \mapsto \sigma^a$, for some vector field σ^a with appropriate properties, and hope to identify η^a with some construction out of σ^a. But the accelerations transform as $\xi^n\nabla_n \mapsto \xi^n\nabla_n\xi^a + \sigma^n\nabla_n\xi^a + \xi^n\nabla_n\sigma^a + \sigma^n\nabla_n\sigma^a$: so this will only work if $\sigma^n\nabla_n\xi^a + \xi^n\nabla_n\sigma^a = 0$, and I am sceptical that there are conditions that could be imposed on σ^a that will make this so for all $\xi^a \in \Xi$.

[11] This is the same point that I made when criticising [Belot, 2013] in §4.2.

[12] e.g. [Friedman, 1983], [Knox, 2014], [Weatherall, 2015c].

[13] e.g. [Knox, 2014].

as follows:

$$\overline{\nabla} \mapsto (\overline{\nabla}, \eta^a t_b t_c) \tag{6.8a}$$

$$G^a \mapsto G^a + \eta^a \tag{6.8b}$$

where η^a is any spacelike vector field such that $\overline{\nabla}^a \eta^b = 0$. It is straightforward to show that this is, indeed, a symmetry of the above: if $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$ and $G'^a = G^a + \eta^a$ are substituted into the above equations, we get the same equations out again. It will be helpful to introduce some terminology for when two connections are related in the manner (6.8a) above:

Definition 9. Let $\mathfrak{L}$ be a Leibnizian spacetime structure, and suppose that ∇ and ∇' are two connections compatible with $\mathfrak{L}$. We say that ∇' is *rigidly corotational* relative to ∇ if $\nabla' = (\nabla, \eta^a t_b t_c)$, for some spacelike field η^a such that $\nabla^a \eta^b = 0$.

Sometimes, however,[14] the symmetry is instead presented as acting upon the gravitational field, the mass density, and the test-particle tangents. More precisely, let a *swerve* be a diffeomorphism $s : \mathfrak{L} \to \mathfrak{L}$ which, in any coordinate system adapted to $\mathfrak{L}$, takes the form

$$t \mapsto t \tag{6.9a}$$

$$\mathbf{x} \mapsto \mathbf{x} + \mathbf{a}(t) \tag{6.9b}$$

where $\mathbf{a}(t)$ is an arbitrary time-dependent translation. Furthermore, this spacetime transformation is conjoined with a certain kind of internal transformation. Again in terms of coordinates, we characterise it as follows: if the components of G^a in the same adapted coordinates as above are $\mathbf{g}$,[15] then define $\widetilde{G}^a$ as the vector field with components (in that same coordinate system)

$$\widetilde{\mathbf{g}} = \mathbf{g} + \ddot{\mathbf{a}}(t) \tag{6.10}$$

where $\ddot{\mathbf{a}}(t)$ is the second temporal derivative of $\mathbf{a}(t)$.

[14] e.g. [Saunders, 2013], [Pooley, 2013b].
[15] Note that there are only three components, since $G^0 = 0$ in any adapted coordinate system.

As discussed in the previous chapter, this transformation naturally lifts to a certain way of transforming the dynamical data: namely,

$$G^a \mapsto s^*(\widetilde{G}^a) \tag{6.11a}$$

$$\mu \mapsto s^*\mu \tag{6.11b}$$

$$\Xi \mapsto s^*\Xi \tag{6.11c}$$

where $s^*\Xi = \{s^*\xi^a : \xi^a \in \Xi\}$. Note that G^a undergoes a "double transformation": it is both acted on by the diffeomorphism, and by the internal transformation (6.10). Showing that this is a symmetry is not very easy in coordinate-free terms—but by translating the dynamical equations (6.6) into coordinates, we can do so.

Thus, it might look at first glance as though we have two symmetry transformations on the table: one of which turns a model $\langle \mathfrak{L}, \overline{\nabla}, G^a, \mu, \Xi\rangle$ into $\langle \mathfrak{L}, \overline{\nabla}', G'^a, \mu, \Xi\rangle$, and the other of which turns that same model into $\langle \mathfrak{L}, \overline{\nabla}, s^*(\widetilde{G}^a), s^*\mu, \{s^*\xi^a\}\rangle$. However, appearances can be deceptive: in an important sense, the effects of these two transformations are equivalent to one another. First, observe that $s^*\mathfrak{L} = \mathfrak{L}$ (that is, $s^*t_a = t_a$ and $s^*h^{ab} = h^{ab}$). As a result,

$$\langle \mathfrak{L}, (s^{-1})^*\overline{\nabla}, \widetilde{G}^a, \mu, \Xi\rangle \cong \langle \mathfrak{L}, \overline{\nabla}, s^*(\widetilde{G}^a), s^*\mu, \{s^*\xi^a\}\rangle \tag{6.12}$$

since the latter results from the former by applying the diffeomorphism s to all the structures comprising it. We can then show that for any swerve s, there is a spacelike field η^a satisfying $\overline{\nabla}^a\eta^b = 0$, such that

$$(s^{-1})^*(\overline{\nabla}) = (\overline{\nabla}, \eta^a t_b t_c) \tag{6.13}$$

and

$$\widetilde{G}^a = G^a + \eta^a = G'^a \tag{6.14}$$

Thus, the models $\langle \mathfrak{L}, \overline{\nabla}', G'^a, \mu, \Xi\rangle$ and $\langle \mathfrak{L}, \overline{\nabla}, s^*(\widetilde{G}^a), s^*\mu, \{s^*\xi^a\}\rangle$ are isomorphic, so we can indeed think of both of these transformations as capturing the same symmetry. Moreover, it is a bona fide symmetry (unlike the case above): we are engaged in systematically transforming the constituent fields.

So, using either characterisation, we can conclude that the models of Galilean gravitation are not invariant under symmetry transformations: more precisely, there is a symmetry transformation which relates non-isomorphic models to one another. Hence,

for the reasons discussed at length already, we have good grounds for thinking that there is something defective about the formalism of Galilean gravitation: it fails to most perspicuously represent the structure to which the best interpretation of the theory is committed. Fortunately, however, there is (or was) a reasonably broad consensus about the formalism which this symmetry motivates us to move to: that of *Newton-Cartan theory*.

6.3 Newton-Cartan gravitation

In Newton-Cartan theory, we still enrich a Leibnizian spacetime structure with an affine connection. However, this connection is no longer required to be flat. To that end, define a *Newton-Cartan spacetime* to comprise

- A Leibnizian spacetime, $\mathfrak{L}$
- An affine connection $\widetilde{\nabla}$ which is compatible with $\mathfrak{L}$, and which satisfies the *homogeneous Trautman conditions*:

$$\widetilde{R}^{ab}{}_{cd} = 0 \tag{6.15a}$$

$$\widetilde{R}^{a}{}_{b}{}^{c}{}_{d} = \widetilde{R}^{c}{}_{d}{}^{a}{}_{b} \tag{6.15b}$$

The conditions (6.15) can be given geometrical interpretations. Equation (6.15a) holds iff parallel transport of spacelike vectors is path independent.[16] As a result, any Newton-Cartan spacetime comes naturally equipped with a standard of rotation (a notion about which more will be said below): to find out if a pair of objects are rotating, for example, just parallel-transport their displacement vector along the path of one of them, and see if it coincides with the displacement vector at a second time. Equation (6.15b) holds iff it is possible to find a smooth, unit timelike field θ^a which is both geodesic ($\theta^n \widetilde{\nabla}_n \theta^a = 0$) and twist-free ($\widetilde{\nabla}^{[a}\theta^{b]} = 0$).[17] It follows from Equation (6.15b) that a vector field χ^a which is geodesic relative to $\widetilde{\nabla}$ will not start to "spontaneously rotate": along any integral curve of χ^a, if χ^a is twist-free at some point on the curve then it is twist-free at all points on the curve.[18] Thus, Equation (6.15a) ensures that the connection admits the comparison of spatial directions at different times (in a path-independent fashion),

[16] [Malament, 2012, Proposition 4.3.1]
[17] [Malament, 2012, Propositions 4.3.3 and 4.3.7]
[18] [Malament, 2012, Proposition 4.3.6]

whilst Equation (6.15b) ensures that geodesics do not exhibit abberrant rotational behaviour.

A dynamically possible model of *Newton-Cartan gravitation* then comprises a Newton-Cartan spacetime $\langle \mathfrak{L}, \widetilde{\nabla} \rangle$, equipped with

- A scalar field $\mu : \mathfrak{L} \to \mathbb{R}$
- A maximal class Ξ of unit timelike vector fields

satisfying the following equations:

$$\widetilde{R}_{bd} = 4\pi\mu t_b t_d \tag{6.16a}$$

$$\xi^n \widetilde{\nabla}_n \xi^a = 0 \tag{6.16b}$$

What is the relationship between Galilean and Newton-Cartan gravitation? Mathematically, their relationship is given by the following pair of theorems:[19]

Theorem 1. Geometrisation Theorem. Let $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a, \Xi \rangle$ be a model of Galilean gravitation. Then there is a unique derivative operator $\widetilde{\nabla}$ on $\mathfrak{L}$ given by

$$\widetilde{\nabla} = (\overline{\nabla}, G^a t_b t_c) \tag{6.17}$$

such that $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ is a model of Newton-Cartan gravitation.

Theorem 2. Recovery Theorem. Let $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ be a model of Newton-Cartan gravitation. Then any flat derivative operator $\overline{\nabla}$ on $\mathfrak{L}$ which is related to $\widetilde{\nabla}$ by equation (6.17), is such that $\langle \mathfrak{L}, \overline{\nabla}, \mu, G^a, \Xi \rangle$ is a model of Galilean gravitation (and, there exists at least one such flat derivative operator).

For proofs, see [Malament, 2012, §4.2]. Theorems 1 and 2 indicate the sense in which Newton-Cartan theory may be considered to include only those structures which are invariant under the gravitational gauge symmetry discussed above. First, consider the case where we characterise that symmetry in terms of a transformation of the connection and the gravitational field: so suppose that $\langle \mathfrak{L}, \overline{\nabla}, G^a, \mu, \Xi \rangle$ and $\langle \mathfrak{L}, \overline{\nabla}', G'^a, \mu, \Xi \rangle$ are models related by the application of the symmetry transformation (6.8). It is straightforward to show that

$$(\overline{\nabla}', G'^a t_b t_c) = (\overline{\nabla}, G^a t_b t_c) \tag{6.18}$$

[19] Due to [Trautman, 1965].

and hence that both models of Galilean gravitation give rise to the same model of Newton-Cartan gravitation. This bears a natural comparison to the shift from a potentials-based formulation of electromagnetism to a fields-based formulation: the sense in which the fields-based formulation contains only the invariants of the other formulation is that two models of the latter, related by a local potential symmetry, give rise to the same fields-based model. In other words, this amounts to a *reduction* of the theory, in the sense of Chapter 4.

Second, consider the case where we characterise it in terms of a "swerve" diffeomorphism, combined with a certain internal transformation; suppose that $\langle \mathfrak{L}, \overline{\nabla}, G^a, \mu, \Xi \rangle$ and $\langle \mathfrak{L}, \overline{\nabla}, s^* G'^a, s^*\mu, s^*\Xi \rangle$ are indeed related by such a diffeomorphism s accompanied by the internal transformation (6.10). Let $\widetilde{\nabla} = (\overline{\nabla}, G^a t_b t_c)$, so that the former model gets turned into $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ by (6.17). Then observe that

$$\begin{aligned}(\overline{\nabla}, s^* G'^a) &= s^*((s^{-1})^* \overline{\nabla}, G'^a) \\ &= s^*(\overline{\nabla}', G'^a) \\ &= s^* \widetilde{\nabla}\end{aligned}$$

So the latter model gets turned into $\langle \mathfrak{L}, s^*\widetilde{\nabla}, s^*\mu, s^*\Xi \rangle$, which is isomorphic to $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$. Note that symmetry-related models are rendered isomorphic rather than identical (i.e. we have a sophistication rather than a reduction); this is in keeping with other external symmetries, as we saw in Chapter 5. Given that we have concluded that the reduced and sophisticated versions of a theory ought to be considered equivalent, this difference does not undermine the claim that these should be considered two presentations of one symmetry, rather than as two distinct symmetries.

As such, we have a somewhat familiar pattern. If the two theories (Galilean gravitation and Newton-Cartan gravitation) are both interpreted literally, then they come out inequivalent. This reflects the different ontological commitments that such interpretations would bring about: for in Galilean gravitation, spacetime is flat, whilst in Newton-Cartan gravitation, spacetime is (typically) curved. However, if they are interpreted so that their commitments about the structure of spacetime are not to be so straightforwardly read off, then one could deem them equivalent. In particular, if one interprets Galilean gravitation in a sophisticated manner (i.e., in such a way that symmetry-related models are deemed equivalent), then it would be expected that it should be equivalent to Newton-Cartan gravitation, literally interpreted. And in fact, this expectation turns out to be correct: one can demonstrate that the sophisticated

category of models of Galilean gravitation is equivalent to the category of models of Newton-Cartan gravitation.[20]

6.4 Maxwell-Weatherall gravitation

Recently, however, Saunders[21] has queried whether we really should regard Newton-Cartan theory as the spacetime theory that properly encodes the lessons of the symmetry canvassed above. Roughly speaking, Saunders' concern might be paraphrased as follows: the symmetry above, at least on the characterisation in terms of swerves, looked like it ought to lead us to the repudiation of absolute accelerations, analogously to the way that boost symmetries lead us to repudiate absolute velocities. Yet any Newton-Cartan spacetime comes with a perfectly well-defined notion of acceleration: after all, each such spacetime carries a connection, albeit one which is not flat. So what gives?

Here is another way of getting at the same kind of worry. Look again at the sense in which Newton-Cartan theory is invariant under swerves: above, I suggested that it was because swerves relate the models $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ and $\langle \mathfrak{L}, s^*\widetilde{\nabla}, s^*\mu, s^*\Xi \rangle$, which we can see to be isomorphic to one another. This suggests that we have gotten this invariance by taking the Leibnizian spacetime $\mathfrak{L}$ as the base space, and everything else (i.e., $\widetilde{\nabla}$, μ and Ξ) as the dynamical data. But if so, then we lose the notion that there is something distinctive about swerves. For Leibnizian spacetime is invariant under a *much* larger class of transformations than swerves: for example, any transformation which (in adapated coordinates) takes the form

$$\mathbf{x} \mapsto \mathbf{R}(t)\mathbf{x} \tag{6.19a}$$

$$t \mapsto t \tag{6.19b}$$

will also be an automorphism of $\mathfrak{L}$, where $\mathbf{R}(t)$ is any time-dependent rotation matrix. So if r is a diffeomorphism of this kind (call it a "twist"), then $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle \cong \langle \mathfrak{L}, r^*\widetilde{\nabla}, r^*\mu, r^*\Xi \rangle$, and it looks as though swerves and twists are on an equal footing. But this isn't true! For twists are not symmetries of Galilean gravitation, no matter what kind of modifications we try to make to the gravitational field. This is exactly the tangle we discussed, in general terms, in §5.6.

[20] [Weatherall, 2015c]
[21] [Saunders, 2013]

However, we are no better off by taking Newton-Cartan spacetime $\langle \mathfrak{L}, \widetilde{\nabla} \rangle$ to be the base space, for a Newton-Cartan spacetime is not invariant under swerves: $\langle \mathfrak{L}, \widetilde{\nabla} \rangle \ncong \langle \mathfrak{L}, s^* \widetilde{\nabla} \rangle$, in general. Hence (in general), $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle \ncong \langle \mathfrak{L}, \widetilde{\nabla}, s^*\mu, s^*\Xi \rangle$. This isn't an idle concern, either: it is precisely this observation that has led various authors[22] to argue that—even though Galilean gravitation admits a symmetry involving arbitrary linear accelerations—it is nevertheless folly to claim that acceleration is relative according to that theory.

The solution is that neither $\mathfrak{L}$ nor $\langle \mathfrak{L}, \widetilde{\nabla} \rangle$ are appropriate candidates for being regarded as the "base space" in Newtonian gravitational theory. For under the application of a swerve s to a Galilean spacetime $\langle \mathfrak{L}, \overline{\nabla} \rangle$, more structure than just $\mathfrak{L}$ is invariant. The structure that is invariant goes by the moniker of *Maxwell spacetime*.[23] Intuitively, the idea is that a Maxwell spacetime contains a "standard of rotation", but no "standard of acceleration". First, I define what it is for two connections to agree on their standard of rotation,[24] and then define a Maxwell spacetime in terms of that.

Definition 10. Let $\mathfrak{L}$ be a Leibnizian spacetime, and let ∇ and ∇' be two connections compatible with $\mathfrak{L}$. ∇ and ∇' are *rotationally equivalent* if, for any unit timelike field θ^a on $\mathfrak{L}$, $\nabla^{[a}\theta^{b]} = 0$ iff $\nabla'^{[a}\theta^{b]} = 0$.

Definition 11. A *Maxwell spacetime* comprises

- A Leibnizian spacetime $\mathfrak{L}$
- A *standard of rotation* W: an equivalence class of rotationally equivalent flat affine connections (compatible with $\mathfrak{L}$)

The different connections in a given standard of rotation are related to one another in a fairly straightforward way, as the following proposition demonstrates.

Proposition 12. Let $\langle \mathfrak{L}, W \rangle$ be a Maxwell spacetime, and consider any $\overline{\nabla} \in W$. For any other connection $\overline{\nabla}'$, $\overline{\nabla}' \in W$ iff $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$, for some spacelike field η^a such that $\overline{\nabla}^a \eta^b = 0$.

Proof. See Appendix 6.A. □

In other words, all of the connections in a Maxwell spacetime's standard of rotation are rigidly corotational relative to one another. This proposition demonstrates the

[22] e.g. [Friedman, 1983, §V.4]
[23] [Earman, 1989, chap. 2]; N.B. that Saunders refers to this structure as "Newton-Huygens spacetime".
[24] This definition follows [Weatherall, 2015c].

invariance of Maxwell spacetime under swerves: defining $s^*W = \{s^*\overline{\nabla} : \overline{\nabla} \in W\}$, we get that for any Maxwell spacetime $\langle \mathfrak{L}, W\rangle$, $\langle s^*\mathfrak{L}, s^*W\rangle = \langle \mathfrak{L}, W\rangle$.

Maxwell spacetime lacks absolute acceleration. However, the sense in which we are led to a theory repudiating absolute acceleration—or in which the role of swerve symmetries is made manifest—depends on our being able to construct a gravitational dynamics that is, in some appropriate sense, "set" upon Maxwell spacetime. The challenge, as [Weatherall, 2015c] observes, is that without a connection, it is not clear how to do so: without a connection, one cannot characterise the dynamically allowed trajectories as those which deviate from inertial motion just insofar as they are acted on by forces. However, he argues that there is a natural indirect solution. In the terminology and notation used here, his core claim is that given a Maxwell spacetime $\langle \mathfrak{L}, W\rangle$, for any $\overline{\nabla} \in W$, there exists some twist-free vector field G^a such that (1) $\overline{\nabla}_a G^a = -4\pi\mu$, where μ is the mass density distribution of spacetime, and (2) the allowed trajectories of bodies are curves, with tangents ξ^a, whose acceleration (relative to $\overline{\nabla}$) is given by $\xi^n\nabla_n\xi^a = G^a$.[25] Here is a way of cashing this out in precise terms.

The idea is that we relativise the gravitational field to an arbitrary choice of derivative operator from W. Of course, the way in which gravitational fields are assigned to operators will have to be constrained, so that (roughly speaking) two fields are related to one another in the same way the operators are. More precisely:

Definition 12. Given a Maxwell spacetime $\langle \mathfrak{L}, W\rangle$, a *relative gravitational field* on $\langle \mathfrak{L}, W\rangle$ is a map $G^a_* : \overline{\nabla} \in W \mapsto G^a_{\overline{\nabla}} \in \mathfrak{X}(\mathfrak{L})$,[26] such that if $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$, then

$$G^a_{\overline{\nabla}'} = G^a_{\overline{\nabla}} - \eta^a \tag{6.20}$$

Definition 13. A model of *Maxwell-Weatherall gravitation* comprises a Maxwell spacetime $\langle \mathfrak{L}, W\rangle$, equipped with

- A relative gravitational field $G^a_* : W \to \mathfrak{X}(\mathfrak{L})$
- A scalar field $\mu : \mathfrak{L} \to \mathbb{R}$
- A maximal class Ξ of unit timelike vector fields

[25] The above is intended as a verbatim transcription of the following: "given a Maxwell-Huygens spacetime $(M, t_a, h^{ab}, [\nabla])$, for any $\nabla \in [\nabla]$, there exists some scalar field φ such that (1) $\nabla_a\nabla^a\varphi = 4\pi\rho$, where ρ is the mass density distribution of spacetime, and (2) the allowed trajectories of bodies are curves γ whose acceleration (relative to ∇) is given by $\xi^n\nabla_n\xi^a = \nabla^a\varphi$." [Weatherall, 2015c, p. 8].

[26] Where $\mathfrak{X}(\mathfrak{L})$ is the space of vector fields on $\mathfrak{L}$.

satisfying the following equations:

$$\overline{\nabla}_a G^a_{\overline{\nabla}} = -4\pi\mu \tag{6.21a}$$

$$\overline{\nabla}^{[a} G^{c]}_{\overline{\nabla}} = 0 \tag{6.21b}$$

$$\xi^n \overline{\nabla}_n \xi^a = G^a_{\overline{\nabla}} \tag{6.21c}$$

for any $\overline{\nabla} \in W$ and every $\xi^a \in \Xi$.

One might object: in what sense is this a dynamics set upon Maxwell spacetime? After all, the equations (6.21) clearly make use of a particular derivative operator from W! However, that need not express a commitment to the structure of that derivative operator in particular, *provided* that whether or not $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ satisfies the equations (6.21) is independent of our choice of $\overline{\nabla}$ from W. This is, indeed, the case.

Proposition 13. Let $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ be a possible model of Maxwell-Weatherall gravitation, and consider any $\overline{\nabla}, \overline{\nabla}' \in W$. Then (for any $\xi^a \in \Xi$) the equations (6.21) hold with respect to $\overline{\nabla}$ iff they hold with respect to $\overline{\nabla}'$.

Proof. This follows straightforwardly from the Trautman Recovery Theorem (Theorem 2). □

What is the relationship between Maxwell-Weatherall gravitation and the theories of gravitation we saw above? Simply this: Maxwell-Weatherall gravitation is (plausibly) equivalent to Newton-Cartan gravitation. There are generic translations which, given any model of Maxwell-Weatherall-gravitation, let us construct a model of Newton-Cartan gravitation; and which, given any model of Newton-Cartan gravitation, let us construct a model of Maxwell-Weatherall gravitation.[27]

Proposition 14. Let $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ be a model of Newton-Cartan gravitation. Consider all pairs $\langle \overline{\nabla}, G^a_{\overline{\nabla}} \rangle$ where $\overline{\nabla}$ is a flat affine connection, $G^a_{\overline{\nabla}}$ is a spacelike twist-free vector field, and $\widetilde{\nabla} = (\overline{\nabla}, G^a_{\overline{\nabla}} t_b t_c)$. Define W as consisting of all the flat connections $\overline{\nabla}$ that feature in some such pair; and define G^a_* as the map $W \to \mathfrak{X}(\mathfrak{L})$ which assigns $\overline{\nabla}$ to $G^a_{\overline{\nabla}}$. Then $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ is a model of Maxwell-Weatherall gravitation.

[27] Determining the categories of models for these theories is rather tricky, given some of the subtleties over how to define transformations for them, which makes applying the notion of categorical equivalence a somewhat delicate business. But the result here gives us reason for thinking that on any plausible way of doing so, the categories will come out equivalent. Indeed, given that we are able to set up a one-to-one correspondence, we anticipate that they would come out isomorphic: that is, that there would exist a pair of functors between them which were *genuinely* inverse to one another.

Proof. See Appendix 6.A. □

Proposition 15. Let $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ be a model of Maxwell-Weatherall gravitation. Let $\overline{\nabla}$ be an arbitrary element of W. Define $\widetilde{\nabla} = (\overline{\nabla}, G^a_{\overline{\nabla}} t_b t_c)$. So defined, $\widetilde{\nabla}$ is independent of the choice of $\overline{\nabla}$; and $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ is a model of Newton-Cartan gravitation.

Proof. See Appendix 6.A. □

Saunders asks the question, "What is the relation between a theory of gravity (and other forces) formulated in Maxwell space-time and one based on Newton-Cartan space-time?"[28] We now have a partial answer: at least for Maxwell-Weatherall gravitation (which could presumably be supplemented with dynamics for other forces as appropriate), the relation is one of equivalence. However, Maxwell-Weatherall gravitation is a rather inelegant theory. It would be nice to have a dynamics which did not have to traffic in "relative gravitational fields", whatever those are. Moreover, the relationship to Newton-Cartan theory is—as Weatherall acknowledges—rather indirect. Apart from anything else, there is still a mismatch between the explicitly geometrical structures of Maxwell-Weatherall theory and those of Newton-Cartan theory: the Newton-Cartan connection defined in Proposition 15 is not a member of W. Finally, one might feel a little unhappy about the prominence that standards of acceleration still play in this theory. It remains the case that the equation of motion (6.21c) is presented as a requirement that the acceleration of a test particle be such-and-such—it's just that the such-and-such is relativised to the standard used to measure the acceleration. It would be preferable to have a condition that is phrased, so far as possible, in terms of relative acceleration.

6.5 Maxwell-Cartan gravitation

In this section, I present a theory that meets these desiderata. First, I define the kind of spacetime structure which we will be using, which will let us make the relationship between our theory and Newton-Cartan theory more explicit. The starting observation is that the connections which figure in a given model of Newton-Cartan gravitation and its various "de-geometrisations" (i.e. the models of Galilean gravitation related to it by equation 6.17) are all rotationally equivalent. (Both in intuitive terms, and in the precise sense canvassed above.) This suggests an obvious way to liberalise the notion of a

[28] [Saunders, 2013, p. 46]

"standard of rotation" that was used in Maxwell spacetime: simply expand it to include non-flat connections. However, we don't want to expand it too far. After all, as we saw above, we know that whatever connection we wind up with after the geometrisation will still obey the homogeneous Trautman conditions (6.15). This observation motivates the following definition.

Definition 14. A *Maxwell-Cartan* spacetime, $\mathfrak{M}$, consists of data $\langle \mathfrak{L}, J \rangle$, where

- $\mathfrak{L}$ is a Leibnizian spacetime
- J is a non-empty equivalence class of Newton-Cartan connections (i.e. $\mathfrak{L}$-compatible connections satisfying the homogeneous Trautman conditions), under the equivalence relation of rotational equivalence.

We saw above that all the connections in a Maxwell spacetime's standard of rotation are related by the relation of rigid corotationality. An analogous result holds for the connections in a Maxwell-Cartan spacetime's standard of rotation.

Proposition 16. Let $\langle \mathfrak{L}, J \rangle$ be a Maxwell-Cartan spacetime, and consider any $\nabla \in J$. For any other connection ∇', $\nabla' \in J$ iff $\nabla' = (\nabla, \eta^a t_b t_c)$, for some spacelike field η^a such that $\nabla^{[a}\eta^{b]} = 0$.

Proof. See Appendix 6.A. □

Again, it is helpful to have some specific terminology for the kind of relationship that holds between the members of J.

Definition 15. Let $\mathfrak{L}$ be a Leibnizian spacetime, and suppose that ∇ and ∇' are two connections compatible with $\mathfrak{L}$. We say that ∇' is *corotational* relative to ∇ if $\nabla' = (\nabla, \eta^a t_b t_c)$, for some spacelike field η^a such that $\nabla^{[a}\eta^{b]} = 0$.

As one would expect, this is a strictly weaker condition that that of being rigidly corotational: that requires that $\nabla^a \eta^b = 0$.

How does Maxwell-Cartan spacetime relate to Maxwell spacetime? Rather closely, as it turns out: indeed, there is good reason to regard them as equivalent. This is because given any Maxwell spacetime, one can construct a unique Maxwell-Cartan spacetime, and vice versa. The following two propositions make this precise.

Proposition 17. Let $\langle \mathfrak{L}, W \rangle$ be a Maxwell spacetime. Define the class of connections J by taking the closure of W under the corotationality relation (so $\nabla \in J$ iff $\nabla = (\overline{\nabla}, \eta^a t_b t_c)$, for some $\overline{\nabla} \in W$ and some spacelike twist-free η^a). Then $\langle \mathfrak{L}, J \rangle$ is a Maxwell-Cartan spacetime.

Proof. See Appendix 6.A. □

Proposition 18. Let $\langle \mathfrak{L}, J \rangle$ be a Maxwell-Cartan spacetime. Define the class of connections W to consist of just those members of J which are flat. Then $\langle \mathfrak{L}, W \rangle$ is a Maxwell spacetime.

Proof. See Appendix 6.A. □

This suggests a sense in which Maxwell-Cartan and Maxwell spacetime are equivalent: a structure $\langle \mathfrak{L}, W, J \rangle$, where $W \subset J$, could be regarded as a "definitional expansion" of both the Maxwell spacetime $\langle \mathfrak{L}, W \rangle$ and the Maxwell-Cartan spacetime $\langle \mathfrak{L}, J \rangle$ (since each of W and J can be defined in terms of the other). However, I will continue to speak of Maxwell-Cartan spacetime to mean the construction above, for two reasons.[29] One is that it can be useful to maintain distinctions of nomenclature, even between what looks like the same thing, differently constructed. The other is that it means I can speak of operators which are "compatible with a given Maxwell-Cartan spacetime", which I stipulate to mean those operators which are members of J; and also to speak of those operators which are "compatible with a given Maxwell spacetime", which I stipulate to mean those operators which are members of W. Clearly, these are two different criteria, so the terminological distinction is being put to work. (Of course, using terminology this way is entirely optional: an alternative would be to speak of a connection being "Maxwell-compatible" rather than "Maxwell-Cartan compatible" with a given Maxwell/Maxwell-Cartan spacetime. But I find that to be a clunkier mode of expression.)

I now specify a gravitational dynamics, against the backdrop of Maxwell-Cartan spacetime, which makes no use of gravitational fields (even relative ones), nor of absolute accelerations. As with Maxwell-Weatherall gravitation, the equations will be expressed in terms of a derivative operator; but we will show that the equations' holding is independent of the choice of derivative operator from the background standard of rotation, J, and therefore presupposes no more structure than that specified by J.

We begin with the following preliminary definition.

Definition 16. Let ξ^a and ξ'^a be unit timelike fields on a Maxwell-Cartan spacetime $\langle \mathfrak{L}, J \rangle$. ξ^a and ξ'^a are *acceleratively equivalent* just in case they have the same acceleration at every point in $\mathfrak{L}$:

$$\xi^n \nabla_n \xi^a = \xi'^n \nabla_n \xi'^a \tag{6.22}$$

[29] See [Saunders, 2013] for yet another way of constructing essentially the same structure (which Saunders refers to as "Newton-Huygens spacetime".

Note that this definition is independent of the choice of derivative operator (from J) used to measure the acceleration of ξ^a and ξ'^a.

We can now state the dynamics. A dynamically possible model of *Maxwell-Cartan gravitation* comprises a Maxwell-Cartan spacetime $\langle \mathfrak{L}, J \rangle$, equipped with

- A scalar field $\mu : \mathfrak{L} \to \mathbb{R}$
- An equivalence class Ξ of acceleratively equivalent unit timelike fields

obeying the following equations:

$$R_{bd}\xi^b\xi^d - \nabla_a(\xi^n\nabla_n\xi^a) = 4\pi\mu \tag{6.23a}$$

$$\nabla^c(\xi^n\nabla_n\xi^a) - \nabla^a(\xi^n\nabla_n\xi^c) = 0 \tag{6.23b}$$

for any $\nabla \in J$ (with Ricci tensor R_{ab}) and $\xi^a \in \Xi$.

As with Maxwell-Weatherall gravitation, in order to justify the claim that we are presupposing only the structure of Maxwell-Cartan spacetime, we need to verify that these equations are invariant under choice of $\nabla \in J$. The following proposition does so.

Proposition 19. Let $\langle \mathfrak{L}, J, \mu, \Xi \rangle$ be a possible model of Maxwell-Cartan gravitation, and consider any $\nabla, \nabla' \in J$. Then (for any $\xi^a \in \Xi$) the equations (6.23) hold with respect to ∇ iff they hold with respect to ∇'.

Proof. See Appendix 6.A. □

As a result, the equations (6.23) are (despite appearances) independent of ∇ and R_{ab}; what is doing the work is the spatial derivative and rotation standard that are part of Maxwell-Cartan spacetime. Note that, as promised, neither of the equations (6.23) involve the assertion that the absolute acceleration of the test-particle trajectories takes a particular value—not even where that value is relativised to the standard used to measure acceleration. The second equation, equation (6.23b), asserts that all dynamically allowed accelerations are twist-free; i.e. that the four-acceleration field associated with any field in Ξ is non-rotating. The first equation, equation (6.23a), asserts that the "average radial relative acceleration" of any field in Ξ is given by $-4/3\pi\mu$.[30] That is, let λ^a be a *connecting field* for ξ^a: a spacelike vector field such that $\mathcal{L}_\xi\lambda^a = 0$ (where $\mathcal{L}_\xi$ denotes the Lie derivative along ξ^a). Intuitively, we think of λ^a as

[30] The below is modelled on [Malament, 2012, Propositions 2.7.2 and 4.3.2].

joining integral curves of ξ^a to "neighbouring" integral curves. The relative acceleration of such neighbouring curves is then given by

$$\xi^n \nabla_n (\xi^m \nabla_m \lambda^a) \tag{6.24}$$

and has radial component (magnitude in the direction of λ^a)

$$\lambda_a \xi^n \nabla_n (\xi^m \nabla_m \lambda^a) \tag{6.25}$$

where $\lambda_a = \hat{h}_{ab}\lambda^b$, for $\hat{h}_{ab}$ the spatial metric associated to ξ^a.[31] This depends on λ^a. But if we introduce three connecting fields $\overset{1}{\lambda}{}^a, \overset{2}{\lambda}{}^a, \overset{3}{\lambda}{}^a$ which are orthonormal to one another, then we can introduce the *average radial acceleration* of ξ^a as the average of the three radial components,

$$\frac{1}{3}\sum_{i=1}^{3} \overset{i}{\lambda}_a \xi^n \nabla_n (\xi^m \nabla_m \overset{i}{\lambda}{}^a) \tag{6.26}$$

It can then be shown that the average radial acceleration is independent of the choice of connecting fields $\overset{i}{\lambda}{}^a$; indeed, we have

Proposition 20. Let ξ^a be a unit timelike field, and suppose that $\{\overset{i}{\lambda}{}^a\}_i$ are three orthonormal spacelike fields such that $\mathcal{L}_\xi \overset{i}{\lambda}{}^a = 0$. Then

$$\frac{1}{3}\sum_{i=1}^{3} \overset{i}{\lambda}_a \xi^n \nabla_n (\xi^m \nabla_m \overset{i}{\lambda}{}^a) = \frac{1}{3}\big(\nabla_a (\xi^n \nabla_n \xi^a) - R_{bd}\xi^b \xi^d\big) \tag{6.27}$$

Proof. See Appendix 6.A. □

Thus, the average radial acceleration of ξ^a is $-4/3\pi\mu$ iff ξ^a obeys equation (6.23a).

Now, I claim that Maxwell-Cartan gravitation is equivalent to Newton-Cartan gravitation. The following two propositions support this claim.

Proposition 21. Let $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ be a model of Newton-Cartan gravitation. Define

$$J = \{\nabla : \nabla = (\widetilde{\nabla}, \eta^a t_b t_c)\} \tag{6.28}$$

for any spacelike η^a such that $\widetilde{\nabla}^{[a}\eta^{b]} = 0$. Then $\langle \mathfrak{L}, J, \mu, \Xi \rangle$ is a model of Maxwell-Cartan gravitation.

[31] In fact, given that λ^a is spacelike, we could have used the spatial metric associated to any unit timelike field; but since we have a particular such field knocking around, it is helpful to fix on it.

Proof. See Appendix 6.A. □

Proposition 22. Let $\langle \mathfrak{L}, J, \mu, \Xi \rangle$ be a model of Maxwell-Cartan gravitation. Let ∇ be an arbitrary element of J, and let ξ^a be an arbitrary element of Ξ. Define a derivative operator $\widetilde{\nabla}$ by

$$\widetilde{\nabla} = (\nabla, t_b t_c \xi^n \nabla_n \xi^a) \tag{6.29}$$

So defined, $\widetilde{\nabla}$ is independent of the choice of ξ^a and ∇; and $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ is a model of Newton-Cartan gravitation.

Proof. See Appendix 6.A. □

There are two comments to make about the above—both of them regarding choices I have made in constructing the theory, and some of the consequences of those choices. The first regards the role of requiring accelerative equivalence among the dynamically allowed trajectories. One might be concerned that this sneaks in a little extra structure "under the rug", as it were. Note, in particular, that the reason why no such requirement was involved in our earlier theories is that there was some object which coordinated the four-accelerations with one another: the gravitational field in Galilean gravitation, the Newton-Cartan connection in Newton-Cartan gravitation, and the relative gravitational field in Maxwell-Weatherall gravitation. So it might be thought that Maxwell-Cartan gravitation is secretly committed to something like a (relative) gravitational field, even though it doesn't appear explicitly on the list of ingredients of a model.

I think that this claim is essentially right. Indeed, it had better be right, if Maxwell-Cartan gravitation is meant to be equivalent to Newton-Cartan or Maxwell-Weatherall gravitation: that equivalence holds precisely because one can construct a (unique) Newton-Cartan connection, or a (unique) relative gravitational field, within any model of Maxwell-Cartan gravitation. Now, if the worry is just that this commitment isn't obvious from the formalism of the theory, then it doesn't strike me as compelling: that is, I don't think we should be worried about the fact that this commitment is less "visible" in Maxwell-Cartan theory than in these other cases. After all, plenty of theories are committed to structure other than that which shows up on the list of ingredients: as we have already seen, for instance, models of Galilean and Newton-Cartan gravitation are committed to a standard of rotation (in the sense that they pick out such a standard), even though no such object shows up in between the angle-brackets of any model.

Alternatively, the worry might be that there is something objectionably "arbitrary" about the fact that it is one equivalence class of trajectories that is picked out as the

class of dynamically allowed trajectories, rather than any other. For, note that there will exist multiple models of Maxwell-Cartan gravitation with the same Maxwell-Cartan spacetime and mass density, but with distinct classes of dynamically allowed trajectories. I think that it is correct to recognise this as a problem, but incorrect to diagnose it as a problem attending specifically to Maxwell-Cartan gravitation: rather, it is just the underdetermination we first saw in section 6.2, illustrated by Proposition 11. And just as was the case there, this underdetermination is (one hopes) an artefact of the decision to treat the mass density as merely phenomenological, rather than as comprised of matter flowing along a dynamically allowed trajectory. It seems plausible to conjecture that if we reversed that decision, then the problem would go away: there would only be one accelerative equivalence class of trajectories consistent with the flow field ζ^a. It would be of value to show this explicitly; but unfortunately, I lack the space to do so here.

The second remark concerns how the above relates to the work of Weatherall and Saunders. Weatherall makes the following claim, regarding the theory that was presented above as "Maxwell-Weatherall gravitation":

> What is the invariant physical structure in this theory? For one, as we have seen, there is the standard of rotation shared between the derivative operators. This gives the sense in which this is a theory in Maxwell-Huygens [[i.e., Maxwell]] spacetime. The other invariant structure, however, is the collection of allowed trajectories for bodies. These are calculated in different ways depending on which representative one chooses from $[[W]]$, and the accelerations associated with each such curve varies similarly. So we do not have the structure to say that these curves are accelerating or not. But however they are described, i.e., whatever acceleration (if any) is attributed to them, the curves themselves are fixed. Indeed, given some distribution of matter in spacetime, it is these curves that form the empirical content of Newtonian gravitational theory.
>
> Now suppose that we are given some such collection of curves, $\{\gamma\}_\rho$, relativized to a matter distribution ρ, in Maxwell-Huygens spacetime. Suppose, too, that however these curves are determined—whether by the calculational procedure just mentioned or some other method—they agree with the possible trajectories allowed by ordinary Newtonian gravitation. It turns out that with this information, one can uniquely reconstruct a [[model of

Newton-Cartan gravitation]].[32]

We are now in a position to clarify some points about this. First, as just discussed, Weatherall is not quite correct to say that the curves are fixed: the underdetermination means that there are several available equivalences classes of dynamically allowed curves. Moreover, if we are working in the context of Maxwell-Weatherall gravitation, then we do not simply get dynamically allowed collections of curves on Maxwell spacetime: each collection is accompanied by a relative gravitational field, pairing the connections in the standard of rotation to gravitational fields.

However, it is natural to see Maxwell-Cartan gravitation as capturing the intuitions expressed in the above passage. In Maxwell-Cartan gravitation, we can indeed make sense of just taking a Maxwell-Cartan spacetime, laying down a mass density upon it, and then considering some particular equivalence class of dynamically allowed trajectories consistent with that mass density. And having done so, we can indeed then construct a unique model of Newton-Cartan gravitation. This construction—i.e., Proposition 22—provides a precise analogue, in the context of Maxwell-Cartan gravitation, to Weatherall's proposition 4, which (translated into the notation used here) states that

> Let $\{\gamma\}_\mu$ be the collection of allowed trajectories for a given mass distribution μ in Maxwell-Huygen spacetime $\langle \mathfrak{L}, W\rangle$, as described above [i.e. in the passage just quoted]. Then there exists a unique derivative operator $\widetilde{\nabla}$ such that (1) $\{\gamma\}_\rho$ consists in the timelike geodesics of $\widetilde{\nabla}$ and (2) $\langle \mathfrak{L}, \widetilde{\nabla}\rangle$ is a model of Newton-Cartan theory for mass density μ.[33]

Weatherall goes on to conclude that

> this result—at least as I interpret it here—reveals a certain inadequacy in Saunders' account. Saunders insists that there is no privileged standard of acceleration in Maxwell-Huygens spacetime. And there are a few senses in which that is right: (1) before accounting for gravitational influences, Maxwell-Huygens spacetime does not have enough structure to make sense of acceleration; and (2) even in the presence of dynamical considerations, there is in general no privileged flat derivative operator, and thus no privileged collection of inertial frames in the standard sense, relative to which acceleration may be defined. Nonetheless, it turns out that once one takes

[32] [Weatherall, 2015c, p. 9]
[33] [Weatherall, 2015c, p. 9]

> the dynamically allowed trajectories into account, one can define a standard of acceleration, namely, the unique one relative to which the allowed trajectories are geodesics.[34]

Understood as a claim about Maxwell-Weatherall or Maxwell-Cartan gravitation, this is quite right. However, it is not clear that either of these really captures what Saunders had in mind. Consider the following passage:

> Take possible worlds each with only a single structureless particle. Depending on the connection, there will be infinitely many distinct trajectories, infinitely many distinct worlds of this kind. But in Newton-Huygens terms, as in Barbour-Bertotti theory, there is only one such world—a trivial one in which there are no meaningful predications of the motion of the particle at all. Only for worlds with two or more particles can distinctions among motions be drawn.[35]

This suggests that Saunders would object to the fact that models of Maxwell-Cartan theory cheerfully include all the dynamically allowed trajectories: for it is precisely by keeping track of all such trajectories that we are able to reconstruct the Newton-Cartan connection from any model of Maxwell-Cartan theory. (This is made clear in the proof of Proposition 22.)

Now, when working with a theory of the kinds discussed so far—in which the mass density is not subject to any kind of non-trivial dynamics—keeping those trajectories in play is pretty crucial (since otherwise, as Proposition 11 illustrates, we will overestimate the symmetry group of the theory). But in the context of a theory in which the mass distribution is *not* treated phenomenologically, Saunders' remark suggests a natural alternative: a version of Maxwell-Cartan theory whose models include only the dynamically *realised* trajectories, i.e., include only the matter flow field ζ^a—and *not* the full collection Ξ of dynamically allowed trajectories. Again, although examining such a theory would be interesting, I lack the space to do so here. I do want to note, however, that the empiricist warrant for such a theory is a little more strained than might at first appear. Empiricist scruples classically require us to permit only structure in our theory which is empirically accessible. A natural way to cash out the idea of empirical accessibility is as " that which can be directly observed by experiment". But the trajectories of possible test particles *can* be directly observed: *if* a particle were to

[34] [Weatherall, 2015c, p. 10]
[35] [Saunders, 2013, pp. 46–47]

be released at a given point, then observing its path would reveal the trajectory! In other words, the fact that these trajectories do not *in fact* contain any matter (in a given model), and so are not in fact *observed*, does not mean that they are *unobservable*—had they contained matter, they would have been observed.

6.6 Conclusion

I wish to conclude with two remarks; the first concerns how this issue relates to the relativity of acceleration. One venerable way of understanding the import of the relativity of velocity is that it demonstrates that the right spacetime structure to take as background is one which is invariant under arbitrary boosts—such as Galilean spacetime (rather than Newtonian spacetime). We can now see that, contrary to what has been claimed in the literature, there is indeed a precise analogue of this lesson to be had in the case of acceleration. The relativity of acceleration (i.e., the fact that Newtonian gravitation contains swerves as symmetries) warrants taking as background a spacetime which is invariant under arbitrary accelerations—such as Maxwell or Maxwell-Cartan spacetime.

This may sound like a reductio, however. For surely it is well-known that absolute accelerations are detectable, unlike absolute velocities? Indeed, the claim that such accelerations should produce measurable effects goes back to Clarke, responding to Leibniz's assertion that *no* absolute motions have physical significance:

> Neither is it sufficient barely to repeat his assertion, that the motion of a finite material universe would be nothing, and (for want of other bodies to compare it with) would [...] produce no discoverable change: unless he could disprove the instance which I gave of a very great change that would happen; viz. that the parts would be sensibly shocked by a sudden acceleration, or stopping of the motion of the whole: to which instance, he has not attempted to give any answer.[36]

However, there is no such problem; Clarke's assertion here is incorrect (at least, if we take a "sudden acceleration, or stopping of the motion of the whole" to be a swerve of the kind discussed above). In picturesque terms, we are inclined to think of the example of picking up a snowglobe and shaking it. It will be obvious to anyone familiar with a snowglobe that this will produce effects discernible within the globe. However,

[36] [Alexander, 1956, pp104–105].

this scenario is *not* what is being described by swerves. Rather, those transformations describe a scenario in which (as it were) God picks up the snowglobe *along with every snowflake and drop of water within it*, and shakes them all *in a precisely co-ordinated fashion*. A little thought makes clear that this process would not produce any measurable effects within the snowglobe: all the relative distances between snowflakes would be entirely unaffected, so that (from within the snowglobe) they continue to drift gently as usual. In the original scenario, we were imagining that the accelerative forces were applied (at least in the first instance) only to the exterior globe itself. The "sensible shock" that we associate with a sudden acceleration is, in fact, just the relative motions caused by applying an accelerative effect to some but not all parts of the system under consideration.

The second is to observe that Proposition 22 is illustrating something quite striking. We begin with a theory containing a relatively minimal amount of geometrical structure; this structure is nevertheless sufficient to permit the characterisation of a non-trivial dynamics; that dynamics then permits us to introduce further geometrical structure (to wit, the connection) as a codification of the behaviour of the dynamical objects.[37] In other words, Proposition 22 provides an (extremely partial) illustration of the dynamical approach to spacetime geometry,[38] in which one seeks to characterise spacetime geometry as a codification of the behaviour of dynamical structures.[39] A natural question is whether this can be extended: can we do the same trick, but starting from a more minimal geometrical basis yet? A starting-point for an answer would be to try and replicate the analysis here, but using the Künzle-Ehlers Recovery Theorem[40] rather than the Trautman Recovery Theorem. It would also be of interest to know to what extent, if any, such techniques could be extended to relativistic rather than Newtonian spacetimes. I postpone these questions to another time.

6.A Proofs of propositions

Proposition 12. Let $\langle \mathfrak{L}, W \rangle$ be a Maxwell spacetime, and consider any $\overline{\nabla} \in W$. For any other connection $\overline{\nabla}'$, $\overline{\nabla}' \in W$ iff $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$, for some spacelike field η^a such that $\overline{\nabla}^a \eta^b = 0$.

[37] cf. [Knox, 2014]
[38] [Brown, 2005], [Stevens, 2015]
[39] [Wallace, 2015b] discusses these issues in more depth.
[40] [Malament, 2012, Proposition 4.5.2]

Proof. First, suppose that $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$ for some spacelike rigid η^a. To show that ∇ and ∇' are rotationally equivalent, let θ^a be any unit timelike field, and observe that

$$\begin{aligned} \nabla'^{[a}\theta^{b]} &= \nabla^{[a}\theta^{b]} - h^{c[a}\eta^{b]}t_c t_n \theta^n \\ &= \nabla^{[a}\theta^{b]} \end{aligned} \tag{6.30}$$

It remains to show that ∇' is flat. Using the standard expression relating two Riemann tensors, we can obtain

$$R'^a{}_{bcd} = R^a{}_{bcd} + 2t_b t_{[d}\nabla_{c]}\eta^a \tag{6.31}$$

But $R^a{}_{bcd} = 0$, and as discussed in the proof of proposition 11, if $\nabla^a \eta^b = 0$ then $t_{[d}\nabla_{c]}\eta^a = 0$. So $R'^a{}_{bcd} = 0$.

□

Proposition 14. Let $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ be a model of Newton-Cartan gravitation. Consider all pairs $\langle \overline{\nabla}, G^a_{\overline{\nabla}} \rangle$ where $\overline{\nabla}$ is a flat affine connection, $G^a_{\overline{\nabla}}$ is a spacelike twist-free vector field, and $\widetilde{\nabla} = (\overline{\nabla}, G^a_{\overline{\nabla}} t_b t_c)$. Define W as consisting of all the flat connections $\overline{\nabla}$ that feature in some such pair; and define G^a_* as the map $W \to \mathfrak{X}(\mathfrak{L})$ which assigns $\overline{\nabla}$ to $G^a_{\overline{\nabla}}$. Then $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ is a model of Maxwell-Weatherall gravitation.

Proof. Consider any pair $\overline{\nabla}, \overline{\nabla}' \in W$. Equation (6.30) indicates that they are both rotationally equivalent to $\widetilde{\nabla}$, and hence to one another; so $\langle \mathfrak{L}, W \rangle$ is a Maxwell spacetime. Moreover, if $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$, then since

$$(\overline{\nabla}, G^a_{\overline{\nabla}} t_b t_c) = \widetilde{\nabla} = (\overline{\nabla}', G^a_{\overline{\nabla}'} t_b t_c) \tag{6.32}$$

we obtain that $\eta^a = G^a_{\overline{\nabla}} - G^a_{\overline{\nabla}'}$; so equation (6.20) is satisfied, and G^a_* is a relative gravitational field.

We now show that $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi \rangle$ satisfies the equations (6.21). That equation (6.21b) is satisfied is given by the definition of G^a_*. As for the others, let $\overline{\nabla} = (\widetilde{\nabla}, -G^a_{\overline{\nabla}} t_b t_c)$ be an arbitrary member of W. Then:

$$\begin{aligned} \xi^n \overline{\nabla}_n \xi^a &= \xi^n \widetilde{\nabla}_n \xi^a + G^a_{\overline{\nabla}} \\ &= G^a_{\overline{\nabla}} \end{aligned}$$

So equation (6.21c) is satisfied. Next, expressing the Riemann tensors of $\widetilde{\nabla}$ and $\overline{\nabla}$ in

terms of one another, and using the fact that the latter's Riemann tensor vanishes:

$$\widetilde{R}^{a}{}_{bcd} = 2t_b t_{[d}\overline{\nabla}_{c]}G^{a}_{\overline{\nabla}} \tag{6.33}$$

Then, using equation (6.16a),

$$\begin{aligned} 4\pi\mu t_b t_d &= \widetilde{R}_{bd} \\ &= -\widetilde{R}^{a}{}_{bad} \\ &= -t_b t_d \overline{\nabla}_a G^{a}_{\overline{\nabla}} \end{aligned}$$

So equation (6.21a) is satisfied.

□

Proposition 15. Let $\langle \mathfrak{L}, W, G^a_*, \mu, \Xi\rangle$ be a model of Maxwell-Weatherall gravitation. Let $\overline{\nabla}$ be an arbitrary element of W. Define $\widetilde{\nabla} = (\overline{\nabla}, G^a_{\overline{\nabla}} t_b t_c)$. So defined, $\widetilde{\nabla}$ is independent of the choice of $\overline{\nabla}$; and $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi\rangle$ is a model of Newton-Cartan gravitation.

Proof. First, suppose (having defined $\widetilde{\nabla}$ with respect to $\overline{\nabla}$) that $\overline{\nabla}'$ is any other member of W. By Proposition 12, $\overline{\nabla}' = (\overline{\nabla}, \eta^a t_b t_c)$, for some η^a such that $\overline{\nabla}^a \eta^b = 0$. Then, using equation (6.20),

$$\begin{aligned} (\overline{\nabla}', G^a_{\overline{\nabla}'}) &= ((\overline{\nabla}, \eta^a), G^a_{\overline{\nabla}'}) \\ &= (\overline{\nabla}, \eta^a + G^a_{\overline{\nabla}'}) \\ &= (\overline{\nabla}, G^a_{\overline{\nabla}}) \\ &= \widetilde{\nabla} \end{aligned}$$

So the definition is independent of the choice of connection in W. Next, using equation (6.33), we immediately get that

$$\widetilde{R}^{ab}{}_{cd} = 0 \tag{6.34}$$

and using equation 6.21b, that

$$\begin{aligned} \widetilde{R}^{a}{}_{b}{}^{c}{}_{d} &= t_b t_d \overline{\nabla}^c G^a_{\overline{\nabla}} \\ &= t_d t_b \overline{\nabla}^a G^c_{\overline{\nabla}} \\ &= \widetilde{R}^{c}{}_{d}{}^{a}{}_{b} \end{aligned}$$

So $\langle \mathfrak{L}, \widetilde{\nabla} \rangle$ is a Newton-Cartan spacetime. Using equation (6.33) in conjunction with equation (6.23a) yields

$$\begin{aligned} \widetilde{R}_{bd} &= -t_b t_d \overline{\nabla}_a G^a_{\overline{\nabla}} \\ &= 4\pi\mu t_b t_d \end{aligned}$$

Thus, equation (6.16a) is satisfied. Finally, for any $\xi^a \in \Xi$, equation (6.21c) ensures that

$$\begin{aligned} \xi^n \widetilde{\nabla}_n \xi^a &= \xi^n \overline{\nabla}_n \xi^a - G^a_{\overline{\nabla}} \\ &= 0 \end{aligned}$$

□

So equation (6.16b) is satisfied.

Proposition 16. Let $\langle \mathfrak{L}, J \rangle$ be a Maxwell-Cartan spacetime, and consider any $\nabla \in J$. For any other connection ∇', $\nabla' \in J$ iff $\nabla' = (\nabla, \eta^a t_b t_c)$, for some spacelike field η^a such that $\nabla^{[a}\eta^{b]} = 0$.

Proof. First, suppose that $\nabla' = (\nabla, \eta^a t_b t_c)$ for some spacelike twist-free field η^a. Equation (6.30) shows that ∇ and ∇' are rotationally equivalent. So all we need to show is that ∇' satisfies the homogeneous Trautman conditions (6.15). Using (6.31),

$$\begin{aligned} R'^{ab}{}_{cd} &= R^{ab}{}_{cd} + 2h^{bn} t_n t_{[d} \nabla_{c]} \eta^a \\ &= R^{ab}{}_{cd} \end{aligned}$$

So clearly, $R'^{ab}{}_{cd} = 0$ iff $R^{ab}{}_{cd} = 0$.

Next, suppose that $R^a{}_b{}^c{}_d = R^c{}_d{}^a{}_b$. Then

$$\begin{aligned} R'^a{}_b{}^c{}_d &= R^a{}_b{}^c{}_d + 2h^{cn} t_b t_{[d} \nabla_{n]} \eta^a \\ &= R^a{}_b{}^c{}_d + t_b t_d \nabla^c \eta^a \\ &= R^c{}_d{}^a{}_b + t_d t_b \nabla^a \eta^c \\ &= R'^c{}_d{}^a{}_b \end{aligned}$$

where the third equality uses our supposition, and the twist-freedom of η^a. Showing that if $R'^a{}_b{}^c{}_d = R'^c{}_d{}^a{}_b$ then $R^a{}_b{}^c{}_d = R^c{}_d{}^a{}_b$ proceeds similarly.

Conversely, suppose that $\nabla' \in J$.[41] Since ∇ and ∇' are both compatible with $\mathfrak{L}$, there

[41] This half of the proof is adapted from a proof in [Weatherall, 2015c].

is some antisymmetric tensor field κ_{ab} such that $\nabla' = (\nabla, 2h^{an}t_{(b}\kappa_{c)n})$.[42] Now let θ^a be some unit timelike field such that $\nabla^{[a}\theta^{b]} = 0$ (some such field is guaranteed to exist, since ∇' obeys the homogeneous Trautman conditions). It follows that

$$\begin{aligned}
0 &= \nabla'^{[a}\theta^{b]} \\
&= \nabla^{[a}\theta^{b]} + 2h^{d[b}h^{a]n}t_{(n}\kappa_{m)d}\theta^m \\
&= h^{d[b}h^{a]n}t_n\kappa_{md}\theta^m + h^{d[b}h^{a]n}t_m\kappa_{nd}\theta^m \\
&= \tfrac{1}{2}(\kappa^{ab} - \kappa^{ba}) \\
&= \kappa^{ab} (= h^{ac}h^{bd}\kappa_{cd})
\end{aligned}$$

So $\kappa_{ab} = t_{[a}\sigma_{b]}$, for some 1-form σ_b; and so $2h^{an}t_{(b}\kappa_{c)n} = \eta^a t_b t_c$ for the spacelike field $\eta^a = 2h^{an}\sigma_n$.

It remains to show that η^a is twist-free. By using equation (6.31), we obtain

$$R'^a{}_b{}^c{}_d = R^a{}_b{}^c{}_d + 2t_b t_d \nabla^c \eta^a \tag{6.35}$$

So by exchange of indices, and applying the second homogeneous Trautman condition,

$$t_b t_d \nabla^c \eta^a = t_b t_d \nabla^a \eta^c \tag{6.36}$$

Since $t_a \neq 0$, $\nabla^{[c}\eta^{a]} = 0$. □

Proposition 17. Let $\langle \mathfrak{L}, W \rangle$ be a Maxwell spacetime. Define the class of connections J by taking the closure of W under the corotationality relation (so $\nabla \in J$ iff $\nabla = (\overline{\nabla}, \eta^a t_b t_c)$, for some $\overline{\nabla} \in W$ and some spacelike twist-free η^a). Then $\langle \mathfrak{L}, J \rangle$ is a Maxwell-Cartan spacetime.

Proof. Since any flat connection satisfies the homogeneous Trautman equations, this is a straightforward consequence of Proposition 16. □

Proposition 18. Let $\langle \mathfrak{L}, J \rangle$ be a Maxwell-Cartan spacetime. Define the class of connections W to consist of just those members of J which are flat. Then $\langle \mathfrak{L}, W \rangle$ is a Maxwell spacetime.

Proof. First, we show that J contains at least one flat connection (so the set W is nonempty).[43] Let ∇ be some member of J, with curvature tensor $R^a{}_{bcd}$. Since $R^{ab}{}_{cd} = 0$,

[42] [Malament, 2012, Proposition 4.1.3]

[43] This part of the proof is essentially just the first part of the Trautman recovery theorem; I follow Malament's treatment.

there is some unit timelike field χ^a which is rigid and twist-free ($\nabla^a\chi^b = 0$).[44] Let $\overline{\nabla} = (\nabla, t_b t_c \chi^n \nabla_n \chi^a)$; that is, let the acceleration field of $\overline{\nabla}$ relative to ∇ be the four-acceleration of χ^a. By Proposition 16, $\overline{\nabla} \in J$; one can also show that it is flat.

Now consider the closure of $\overline{\nabla}$ under the relation of rigid corotationality, $[\overline{\nabla}]$. By Proposition 12, $\langle \mathfrak{L}, [\overline{\nabla}] \rangle$ is a Maxwell spacetime. So all we need to show is that $W = [\overline{\nabla}]$. So consider any other flat operator $\overline{\nabla}' \in J$. Using equation (6.31), we obtain that $t_b t_{[d} \nabla_{c]} \eta^a = 0$. Acting on both sides with h^{cn} yields $t_b t_d \nabla^c \eta^a = 0$; since $t_a \neq 0$, $\nabla^c \eta^a = 0$. So all the flat operators in J are related to one another by rigid accelerations, from which it follows that $W = [\overline{\nabla}]$, and hence that $\langle \mathfrak{L}, W \rangle$ is a Maxwell spacetime. □

Proposition 19. Let $\langle \mathfrak{L}, J, \mu, \Xi \rangle$ be a possible model of Maxwell-Cartan gravitation, and consider any $\nabla, \nabla' \in J$. Then (for any $\xi^a \in \Xi$) the equations (6.23) hold with respect to ∇ iff they hold with respect to ∇'.

Proof. By Proposition 16, $\nabla' = (\nabla, \eta^a t_b t_c)$, for some spacelike field η^a such that $\nabla^{[a}\eta^{b]} = 0$. Now, using equation (6.31),

$$\begin{aligned} R'_{bd} &= R_{bd} + 2 t_b t_{[d} \nabla_{a]} \eta^a \\ &= R_{bd} + t_b t_d \nabla_a \eta^a \end{aligned}$$

and so for any timelike ξ^a,

$$R'_{bd} \xi^b \xi^d = R_{bd} \xi^b \xi^d + \nabla_a \eta^a \tag{6.37}$$

On the other hand,

$$\begin{aligned} \nabla'_a (\xi^n \nabla'_n \xi^a) &= \nabla'_a (\xi^n \nabla_n \xi^a + \eta^a) \\ &= \nabla_a (\xi^n \nabla_n \xi^a + \eta^a) + \eta^a t_a t_r (\xi^n \nabla_n \xi^r + \eta^r) \\ &= \nabla_a (\xi^n \nabla_n \xi^a) + \nabla_a \eta^a \end{aligned}$$

Hence,

$$R_{bd} \xi^b \xi^d - \nabla_a (\xi^n \nabla_n \xi^a) = R'_{bd} \xi^b \xi^d - \nabla'_a (\xi^n \nabla'_n \xi^a) \tag{6.38}$$

[44] [Malament, 2012, Proposition 4.2.4 (1)]

So equation (6.23a) holds with respect to ∇ iff it holds with respect to ∇'. Second,

$$\begin{aligned}\nabla'^c(\xi^n\nabla'_n\xi^a) &= \nabla'^c(\xi^n\nabla_n\xi^a + \eta^a)\\ &= \nabla^c(\xi^n\nabla_n\xi^a + \eta^a) + h^{dc}\eta^a t_d t_e(\xi^n\nabla_n\xi^e + \eta^e)\\ &= \nabla^c(\xi^n\nabla_n\xi^a) + \nabla^c\eta^a\end{aligned}$$

Similarly,

$$\nabla'^a(\xi^n\nabla'_n\xi^c) = \nabla^a(\xi^n\nabla_n\xi^c) + \nabla^a\eta^c$$

Since $\nabla^c\eta^a = \nabla^a\eta^c$ (i.e., since η^a is twist-free), equation (6.23b) also holds with respect to ∇ iff it holds with respect to ∇'. □

Proposition 20. Let ξ^a be a unit timelike field, and suppose that $\{\overset{i}{\lambda}{}^a\}_i$ are three orthonormal spacelike fields such that $\mathcal{L}_\xi\overset{i}{\lambda}{}^a = 0$. Then

$$\frac{1}{3}\sum_{i=1}^{3}\overset{i}{\lambda}_a\xi^n\nabla_n(\xi^m\nabla_m\overset{i}{\lambda}{}^a) = \frac{1}{3}\big(\nabla_a(\xi^n\nabla_n\xi^a) - R_{bd}\xi^b\xi^d\big) \tag{6.27}$$

Proof. First, note that for any connecting field λ^a, $\xi^n\nabla_n\lambda^a = \lambda^n\nabla_n\xi^a$ (since $\mathcal{L}_\xi\lambda^a = 0$). Hence,[45]

$$\begin{aligned}\xi^n\nabla_n(\xi^m\nabla_m\lambda^a) &= \xi^n\nabla_n(\lambda^m\nabla_m\xi^a)\\ &= (\xi^n\nabla_n\lambda^m)\nabla_m\xi^a + \xi^n\lambda^m\nabla_n\nabla_m\xi^a\\ &= (\lambda^n\nabla_n\xi^m)\nabla_m\xi^a + \xi^n\lambda^m\nabla_m\nabla_n\xi^a + \xi^n\lambda^m R^a{}_{rmn}\xi^r\\ &= (\lambda^m\nabla_m\xi^n)\nabla_n\xi^a + \lambda^m\nabla_m(\xi^n\nabla_n\xi^a) - \lambda^m(\nabla_m\xi^n)(\nabla_n\xi^a) + \xi^n\lambda^m R^a{}_{rmn}\xi^r\\ &= \lambda^m(\nabla_m(\xi^n\nabla_n\xi^a) + R^a{}_{rmn}\xi^r\xi^n)\end{aligned}$$

Since the connecting fields are orthonormal,

$$\hat{h}_{ab} = \sum_i \overset{i}{\lambda}_a\overset{i}{\lambda}_b \tag{6.39}$$

and so[46]

$$\sum_i \overset{i}{\lambda}_a\overset{i}{\lambda}{}^c = \delta_a{}^c - t_a\xi^c \tag{6.40}$$

[45] This calculation is just an extension of the proof of [Malament, 2012, Proposition 1.8.5], for the case where ξ^a is not a geodesic.

[46] [Malament, 2012, Equation 4.1.12]

Therefore,

$$\begin{aligned}\frac{1}{3}\sum_{i=1}^{3}\overset{i}{\lambda}_a\xi^n\nabla_n(\xi^m\nabla_m\overset{i}{\lambda}{}^a) &= \frac{1}{3}(\sum_i\overset{i}{\lambda}_a\overset{i}{\lambda}{}^c)(\nabla_c(\xi^n\nabla_n\xi^a)+R^a{}_{bcd}\xi^b\xi^d)\\ &= \frac{1}{3}(\delta_a{}^c - t_a\xi^c)(\nabla_c(\xi^n\nabla_n\xi^a)+R^a{}_{bcd}\xi^b\xi^d)\\ &= \frac{1}{3}(\nabla_a(\xi^n\nabla_n\xi^a)-R_{bd}\xi^b\xi^d)\end{aligned}$$

□

Proposition 21. Let $\langle\mathfrak{L},\widetilde{\nabla},\mu,\Xi\rangle$ be a model of Newton-Cartan gravitation. Define

$$J=\{\nabla:\nabla=(\widetilde{\nabla},\eta^a t_b t_c)\} \tag{6.28}$$

for any spacelike η^a such that $\widetilde{\nabla}^{[a}\eta^{b]}=0$. Then $\langle\mathfrak{L},J,\mu,\Xi\rangle$ is a model of Maxwell-Cartan gravitation.

Proof. By Proposition 16, $\langle\mathfrak{L},J\rangle$ is a Maxwell-Cartan spacetime. The equations (6.16) guarantee that the equations (6.23) hold with respect to $\widetilde{\nabla}$; by Proposition 19, they therefore hold with respect to any $\nabla\in J$. Thus, $\langle\mathfrak{L},J,\mu,\Xi\rangle$ is a model of Maxwell-Cartan gravitation. □

Proposition 22. Let $\langle\mathfrak{L},J,\mu,\Xi\rangle$ be a model of Maxwell-Cartan gravitation. Let ∇ be an arbitrary element of J, and let ξ^a be an arbitrary element of Ξ. Define a derivative operator $\widetilde{\nabla}$ by

$$\widetilde{\nabla}=(\nabla,t_b t_c\xi^n\nabla_n\xi^a) \tag{6.29}$$

So defined, $\widetilde{\nabla}$ is independent of the choice of ξ^a and ∇; and $\langle\mathfrak{L},\widetilde{\nabla},\mu,\Xi\rangle$ is a model of Newton-Cartan gravitation.

Proof. First, suppose (having defined $\widetilde{\nabla}$ with respect to ξ^a and ∇) that ξ'^a is any other member of Ξ. Since all members of Ξ are acceleratively equivalent to one another,

$$(\nabla,t_b t_c\xi'^n\nabla_n\xi'^a)=(\nabla,t_b t_c\xi^n\nabla_n\xi^a) \tag{6.41}$$

so the definition of $\widetilde{\nabla}$ is independent of the choice of ξ^a. Now suppose that ∇' is any other member of J. By Proposition 16, $\nabla'=(\nabla,\eta^a t_b t_c)$ for some η^a such that $\nabla^{[a}\eta^{b]}=0$.

Then

$$\begin{aligned}(\nabla', t_b t_c \xi^n \nabla'_n \xi^a) &= ((\nabla, \eta^a t_b t_c), t_b t_c (\xi^n \nabla_n \xi^a - \eta^a)) \\ &= (\nabla, t_b t_c \xi^n \nabla_n \xi^a) \\ &= \widetilde{\nabla}\end{aligned}$$

So the definition is independent of the choice of ∇.

Since $\xi^n \nabla_n \xi^a$ is spacelike, and given (6.23b) (for ∇), we get (by Proposition 16) that $\widetilde{\nabla} \in J$, and so obeys the homogeneous Trautman conditions. So $\langle \mathfrak{L}, \widetilde{\nabla} \rangle$ is a Newton-Cartan spacetime. Moreover, for any $\xi^a \in \Xi$,

$$\xi^n \widetilde{\nabla}_n \xi^a = \xi^n \nabla_n \xi^a - \xi^n \nabla_n \xi^a = 0 \tag{6.42}$$

So we immediately have equation (6.16b): that is, the dynamically allowed trajectories are precisely the geodesics of $\widetilde{\nabla}$. Since $\widetilde{\nabla} \in J$, we also have that it satisfies (6.23a) for all such geodesic vector fields, i.e., for every $\xi^a \in \Xi$,

$$\widetilde{R}_{bd} \xi^b \xi^d = 4\pi\mu \tag{6.43}$$

But if this holds for *every* geodesic field only if (6.16a) holds. So $\langle \mathfrak{L}, \widetilde{\nabla}, \mu, \Xi \rangle$ is a model of Newton-Cartan gravitation. □

Conclusion

> When I married Humphrey I made up my mind to like sermons, and I set out by liking the end very much. That soon spread to the middle and the beginning, because I couldn't have the end without them.
>
> George Eliot, *Middlemarch*

To conclude, I will make a few brief remarks about some of the recurring themes of the above. First, I hope to have shown how attention to the internal character of a theory, or to the formal relationships between theories, can be an interesting interpretative project in its own right. I also hope to have shown how the appropriation of resources from the philosophy of logic and language can be brought to bear on that project, illuminating how rich and complex the notions of translation and synonymy can be. I suspect that an overly simplified picture of translation (and cognate notions) may contribute to the assumption that such "merely semantic" explication cannot yield philosophical insight. Providing a translation between theories is not easy; nor is providing a translation from a theory to itself. But by the same token, that means that the presence and character of such translations can shed a surprising amount of light on a theory's internal architecture.

Second, I have generally abjured appeals to notions like grounding or fundamentality, except insofar as they can be made out within the structure of natural laws. This is against the trend of much recent metaphysics, which trumpets the power of notions of such hyperintensional and extra-nomological structure. Indeed: in many ways, my problem with these notions is just that they are *so* powerful, they can make it *harder* to make philosophical progress. Grounding can provide such a rich superabundance of structure that one can use it to explicate almost any kind of ontological picture you like. I think there is a good methodological case for working with more limited metaphysical resources, and seeing how far it is possible to get. To be clear, I don't think that doing

so means that we are somehow "avoiding" metaphysics: metaphysical deflationists, after all, are just those who disagree with *both* sides in a given metaphysical debate, by maintaining that there is no disagreement there at all. Given that, it is important that deflationists do not cede the ground of metaphysical argument to their opponents. If it really is the case that two apparently different positions are really equivalent, then that is something that will need to be argued for.

Finally, and also on a methodological note, I hope that the above illustrates the value of an eclectic approach to formalisms. Rather than alighting on some framework—first-order logic, differential geometry, category theory, or whatever—as the be-all and end-all, we should be pluralistic about what tools are best applied to the formal study of scientific theories. For example, if we want a tight grip on how the derivable consequences of some axioms relate to the models of those axioms, then we should make use of model theory; but, we should bear in mind that virtually no realistic theory will be expressible in those terms. If we want to abstract away and apply a uniform condition for equivalence, then we should characterise our theories as categories; but, we should bear in mind that not all of the essential information about a theory is likely to reside in that category we have rendered it as. By shifting between methods and means as circumstances demand, we can discern similarities and analogues between different formalisms, and use these to cross-fertilise our investigations into one area with insights from another.

Bibliography

[Adams, 1979] Adams, R. M. (1979). Primitive thisness and primitive identity. *The Journal of Philosophy*, 76(1):5–26.

[Aharonov and Bohm, 1959] Aharonov, Y. and Bohm, D. (1959). Significance of Electromagnetic Potentials in the Quantum Theory. *Physical Review*, 115(3):485–491.

[Ainsworth, 2009] Ainsworth, P. M. (2009). Newman's Objection. *The British Journal for the Philosophy of Science*, 60(1):135–171.

[Alexander, 1956] Alexander, H. G., editor (1956). *The Leibniz-Clarke correspondence*. Manchester University Press, Manchester.

[Allori et al., 2008] Allori, V., Goldstein, S., Tumulka, R., and Zanghì, N. (2008). On the Common Structure of Bohmian Mechanics and the Ghirardi–Rimini–Weber Theory. *The British Journal for the Philosophy of Science*, 59(3):353–389.

[Anderson, 1967] Anderson, J. L. (1967). *Principles of relativity physics*. Academic Press, New York.

[Arntzenius, 2012] Arntzenius, F. (2012). *Space, Time, and Stuff*. Oxford University Press, Oxford.

[Awodey, 2010] Awodey, S. (2010). *Category theory*. Oxford University Press, Oxford; New York.

[Bacon, 2014] Bacon, A. (2014). Representing Counterparts. *The Australasian Journal of Logic*, 11(2):90–113.

[Baez and Muniain, 1994] Baez, J. C. and Muniain, J. P. (1994). *Gauge fields, knots, and gravity*. World Scientific, Singapore; River Edge, N.J.

[Bain, 2003] Bain, J. (2003). Einstein algebras and the hole argument. *Philosophy of Science*, 70(5):1073–1085. WOS:000220274500019.

[Baker, 2010] Baker, D. J. (2010). Symmetry and the Metaphysics of Physics. *Philosophy Compass*, 5(12):1157–1166.

[Barrett, 2015a] Barrett, T. W. (2015a). On the Structure of Classical Mechanics. *The British Journal for the Philosophy of Science*, 66(4):801–828.

[Barrett, 2015b] Barrett, T. W. (2015b). Spacetime structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 51:37–43.

[Barrett and Halvorson, 2015] Barrett, T. W. and Halvorson, H. (2015). Glymour and Quine on Theoretical Equivalence. Unpublished draft.

[Belot, 1995] Belot, G. (1995). New Work for Counterpart Theorists: Determinism. *The British Journal for the Philosophy of Science*, 46(2):185–195.

[Belot, 2007] Belot, G. (2007). The representation of time and change in mechanics. In Butterfield, J. and Earman, J., editors, *Philosophy of Physics*, volume A of *Handbook of the philosophy of science*, pages 133–227. North-Holland, Amsterdam.

[Belot, 2011] Belot, G. (2011). Background-independence. *General Relativity and Gravitation*, 43(10):2865–2884.

[Belot, 2013] Belot, G. (2013). Symmetry and equivalence. In Batterman, R. W., editor, *The Oxford Handbook of Philosophy of Physics*. Oxford University Press, New York.

[Benacerraf, 1965] Benacerraf, P. (1965). What numbers could not be. *Philosophical Review*, 74:47–73. References are to the reprint in [Benacerraf and Putnam, 1983].

[Black, 1952] Black, M. (1952). The Identity of Indiscernibles. *Mind*, 61(242):153–164.

[Blizard, 1988] Blizard, W. D. (1988). Multiset theory. *Notre Dame Journal of Formal Logic*, 30(1):36–66.

[Brading and Brown, 2004] Brading, K. and Brown, H. R. (2004). Are Gauge Symmetry Transformations Observable? *The British Journal for the Philosophy of Science*, 55(4):645–665.

[Breckenridge and Magidor, 2010] Breckenridge, W. and Magidor, O. (2010). Arbitrary reference. *Philosophical Studies*, 158(3):377–400.

[Bricker, 2006] Bricker, P. (2006). Absolute actuality and the plurality of worlds. *Philosophical perspectives*, 20(1):41–76.

[Brighouse, 1997] Brighouse, C. (1997). Determinism and Modality. *The British Journal for the Philosophy of Science*, 48(4):465–481.

[Brighouse, 2008] Brighouse, C. (2008). Understanding Indeterminism. In Dieks, D., editor, *The Ontology of Spacetime II*, volume 4 of *Philosophy and Foundations of Physics*, pages 153–173. Elsevier.

[Brown, 2005] Brown, H. R. (2005). *Physical relativity: space-time structure from a dynamical perspective*. Oxford University Press, Oxford.

[Carnap, 1956] Carnap, R. (1956). *Meaning and necessity: a study in semantics and modal logic*. University of Chicago Press, Chicago.

[Cartwright, 1999] Cartwright, N. (1999). *The dappled world: a study of the boundaries of science*. Cambridge University Press, Cambridge, UK; New York, NY.

[Chalmers, 2006] Chalmers, D. (2006). The Foundations of Two-Dimensional Semantics. In Garcia-Carpintero, M. and Macià, J., editors, *Two-Dimensional Semantics: Foundations and Applications*, pages 55–140. Oxford University Press, Oxford.

[Chalmers, 1997] Chalmers, D. J. (1997). *The Conscious Mind: In Search of a Fundamental Theory*. OUP USA.

[Coffey, 2014] Coffey, K. (2014). Theoretical Equivalence as Interpretative Equivalence. *The British Journal for the Philosophy of Science*, 65(4):821–844.

[Correia and Schnieder, 2012] Correia, F. and Schnieder, B., editors (2012). *Metaphysical Grounding*. Cambridge University Press.

[Dasgupta, 2009] Dasgupta, S. (2009). Individuals: An essay in revisionary metaphysics. *Philosophical Studies*, 145(1):35–67.

[Dasgupta, 2011] Dasgupta, S. (2011). The Bare Necessities. *Philosophical Perspectives*, 25(1):115–160.

[Dasgupta, 2014a] Dasgupta, S. (2014a). On the plurality of grounds. *Philosophers' Imprint*, 14(20):1–28.

[Dasgupta, 2014b] Dasgupta, S. (2014b). Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science*, Forthcoming.

[Dasgupta, 2015] Dasgupta, S. (2015). Inexpressible Ignorance. *Philosophical Review*, 124(4):441–480.

[Dasgupta, 2016] Dasgupta, S. (2016). Quality and Structure. In Barnes, E., editor, *Current Controversies in Metaphysics*. Routledge.

[de Bouvère, 1965] de Bouvère, K. (1965). Synonymous Theories. In Addison, J. W., Henkin, L., and Tarski, A., editors, *The Theory of Models: Proceedings of the 1963 International Symposium at Berkeley*, Studies in Logic and the Foundations fo Mathematics. North-Holland, Amsterdam.

[Dennett, 2000] Dennett, D. C. (2000). With a Little Help from My Friends. In Ross, D., Brook, A., and Thompson, D., editors, *Dennett's Philosophy: A Comprehensive Assessment*, pages 327–388. MIT Press, Cambridge, MA.

[Dorato, 2000] Dorato, M. (2000). Substantivalism, relationism, and structural spacetime realism. *Foundations of Physics*, 30(10):1605–1628.

[Earman, 1977] Earman, J. (1977). Leibnizian Space-Times and Leibnizian Algebras. In Butts, R. E. and Hintikka, J., editors, *Historical and Philosophical Dimensions of Logic, Methodology and Philosophy of Science*, number 12 in The University of Western Ontario Series in Philosophy of Science, pages 93–112. Springer Netherlands.

[Earman, 1989] Earman, J. (1989). *World enough and space-time: absolute versus relational theories of space and time*. MIT Press, Cambridge, Mass.

[Earman, 2003] Earman, J. (2003). Tracking down gauge: an ode to the constrained Hamiltonian formalism. In *Symmetries in Physics: Philosophical Reflections*. Cambridge University Press, Cambridge.

[Earman and Norton, 1987] Earman, J. and Norton, J. (1987). What Price Spacetime Substantivalism? The Hole Story. *The British Journal for the Philosophy of Science*, 38(4):515–525.

[Egg and Esfeld, 2014] Egg, M. and Esfeld, M. (2014). Primitive ontology and quantum state in the GRW matter density theory. *Synthese*, 192(10):3229–3245.

[Einstein, 1905] Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik*, 322(10):891–921. Translated as {"On the Electrodynamics of Moving Bodies"}.

[Esfeld et al., 2014] Esfeld, M., Hubert, M., Lazarovici, D., and Dürr, D. (2014). The Ontology of Bohmian Mechanics. *The British Journal for the Philosophy of Science*, 65(4):773–796.

[Friedman, 1983] Friedman, M. (1983). *Foundations of space-time theories: relativistic physics and philosophy of science*. Princeton University Press, Princeton.

[Geroch, 1972] Geroch, R. (1972). Einstein algebras. *Communications in Mathematical Physics*, 26(4):271–275.

[Gibbard, 1975] Gibbard, A. (1975). Contingent Identity. *Journal of Philosophical Logic*, 4:187–221.

[Greaves, 2011] Greaves, H. (2011). In Search of (Spacetime) Structuralism. *Philosophical Perspectives*, 25(1):189–204.

[Greaves and Wallace, 2014] Greaves, H. and Wallace, D. (2014). Empirical Consequences of Symmetries. *The British Journal for the Philosophy of Science*, 65(1):59–89.

[Hall, 2012] Hall, N. (2012). Physical Magnitudes (supplement to David Lewis's Metaphysics). In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2012 edition.

[Halvorson, 2012] Halvorson, H. (2012). What Scientific Theories Could Not Be. *Philosophy of Science*, 79(2):183–206.

[Halvorson, 2013] Halvorson, H. (2013). The Semantic View, If Plausible, Is Syntactic. *Philosophy of Science*, 80(3):475–478.

[Halvorson and Tsementzis, 2015] Halvorson, H. and Tsementzis, D. (2015). Categories of scientific theories. `http://philsci-archive.pitt.edu/11596/`.

[Hawthorne, 2001] Hawthorne, J. (2001). Causal Structuralism. *Noûs*, 35:361–378.

[Hazen, 1979] Hazen, A. (1979). Counterpart-Theoretic Semantics for Modal Logic. *The Journal of Philosophy*, 76(6):319–338.

[Healey, 1995] Healey, R. (1995). Substance, modality and spacetime. *Erkenntnis*, 42(3):287–316.

[Healey, 2006] Healey, R. (2006). Symmetry and the Scope of Scientific Realism. In Demopoulos, W. and Pitowsky, I., editors, *Physical Theory and its Interpretation*, number 72 in The Western Ontario Series in Philosophy of Science, pages 143–160. Springer Netherlands.

[Healey, 2007] Healey, R. (2007). *Gauging What's Real*. Oxford University Press.

[Healey, 2009] Healey, R. (2009). Perfect Symmetries. *The British Journal for the Philosophy of Science*, 60(4):697–720.

[Hildebrand, 2015] Hildebrand, T. (2015). Two Types of Quidditism. *Australasian Journal of Philosophy*, 0(0):1–17.

[Hodges, 1997] Hodges, W. (1997). *A shorter model theory*. Cambridge University Press, Cambridge; New York.

[Holland and Brown, 2003] Holland, P. and Brown, H. R. (2003). The non-relativistic limits of the Maxwell and Dirac equations: the role of Galilean and gauge invariance. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 34(2):161–187.

[Jones, 1991] Jones, R. (1991). Realism about What? *Philosophy of Science*, 58(2):185–202.

[Kaplan, 1975] Kaplan, D. (1975). How to Russell a Frege-Church. *The Journal of Philosophy*, 72(19):716–729.

[Kearns and Magidor, 2012] Kearns, S. and Magidor, O. (2012). Semantic Sovereignty. *Philosophy and Phenomenological Research*, 85(2):322–350.

[Ketland, 2004] Ketland, J. (2004). Empirical Adequacy and Ramsification. *The British Journal for the Philosophy of Science*, 55(2):287–300.

[Knox, 2011] Knox, E. (2011). Newton–Cartan theory and teleparallel gravity: The force of a formulation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42(4):264–275.

[Knox, 2013] Knox, E. (2013). Effective spacetime geometry. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 44(3):346–356.

[Knox, 2014] Knox, E. (2014). Newtonian Spacetime Structure in Light of the Equivalence Principle. *The British Journal for the Philosophy of Science*, 65(4):863–880.

[Kosso, 2000] Kosso, P. (2000). The empirical status of symmetries in physics. *The British Journal for the Philosophy of Science*, 51(1):81–98.

[Kripke, 1971] Kripke, S. (1971). Identity and necessity. In Munitz, M. K., editor, *Identity and Individuation*, pages 161–191. New York University Press, New York.

[Langton, 2004] Langton, R. (2004). Elusive Knowledge of Things in Themselves. *Australasian Journal of Philosophy*, 82(1):129–136.

[Le Bellac and Lévy-Leblond, 1973] Le Bellac, M. and Lévy-Leblond, J. M. (1973). Galilean electromagnetism. *Il Nuovo Cimento*, 14(2):217–234.

[Leeds, 1995] Leeds, S. (1995). Holes and Determinism: Another Look. *Philosophy of Science*, 62(3):425–437.

[Lewis, 1983] Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377.

[Lewis, 1986] Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell Publishers Ltd, Oxford.

[Lewis, 2009] Lewis, D. (2009). Ramseyan Humility. In Braddon-Mitchell, D. and Nola, R., editors, *Conceptual Analysis and Philosophical Naturalism*, pages 203–222. Mit Press.

[Lewis, 1968] Lewis, D. K. (1968). Counterpart Theory and Quantified Modal Logic. *The Journal of Philosophy*, 65(5):113–126.

[Loll, 1994] Loll, R. (1994). The Loop Formulation of Gauge Theory and Gravity. In *Knots and Quantum Gravity*, pages 1–20. Clarendon Press, Oxford.

[Lutz, 2014a] Lutz, S. (2014a). Empirical Adequacy in the Received View. *Philosophy of Science*, 81(5):1171–1183.

[Lutz, 2014b] Lutz, S. (2014b). What's Right with a Syntactic Approach to Theories and Models? *Erkenntnis*. Published online.

[Lutz, 2015] Lutz, S. (2015). What Was the Syntax-Semantics Debate in the Philosophy of Science About? *Philosophy and Phenomenological Research*, 91(3).

[Malament, 2012] Malament, D. B. (2012). *Topics in the foundations of general relativity and Newtonian gravitation theory*. University of Chicago Press, Chicago.

[Manfredi, 2013] Manfredi, G. (2013). Non-relativistic limits of Maxwell's equations. *European Journal of Physics*, 34(4):859–871.

[Maudlin, 1988] Maudlin, T. (1988). The Essence of Space-Time. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1988:82–91.

[Maudlin, 1993] Maudlin, T. (1993). Buckets of Water and Waves of Space: Why Space-time Is Probably a Substance. *Philosophy of Science*, 60(2):183–203.

[Maudlin, 2007a] Maudlin, T. (2007a). *The Metaphysics Within Physics*. Oxford University Press, Oxford.

[Maudlin, 2007b] Maudlin, T. (2007b). A Modest Proposal Concerning Laws, Counterfactuals, and Explanations. In *The Metaphysics Within Physics*. Oxford University Press, Oxford.

[Maudlin, 2007c] Maudlin, T. (2007c). Suggestions from Physics for Deep Metaphysics. In *The Metaphysics Within Physics*, pages 78–103. Oxford University Press, Oxford.

[Maudlin, 2007d] Maudlin, T. W. E. (2007d). Completeness, supervenience and ontology. *Journal of Physics A: Mathematical and Theoretical*, 40(12):3151.

[Melia, 1999] Melia, J. (1999). Holes, Haecceitism and two conceptions of determinism. *The British Journal for the Philosophy of Science*, 50(4):639–664.

[Milnor and Stasheff, 1974] Milnor, J. W. and Stasheff, J. D. (1974). *Characteristic Classes*. Princeton University Press.

[Mundy, 1992] Mundy, B. (1992). Space-time and isomorphism. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pages 515–527.

[Newman, 1928] Newman, M. H. A. (1928). Mr. Russell's "Causal Theory of Perception". *Mind*, 37(146):137–148.

[Newton, 2004] Newton, I. (2004). De Gravitatione. In Janiak, A., editor, *Newton: Philosophical Writings*, Cambridge Texts in the History of Philosophy, pages 12–39. Cambridge University Press, Cambridge.

[Nolan, 2015] Nolan, D. (2015). *David Lewis*. Routledge.

[Norton, 1993] Norton, J. D. (1993). General Covariance and the Foundations of General Relativity: Eight Decades of Dispute. *Reports of Progress in Physics*, 56:791–858.

[Pais, 1982] Pais, A. (1982). *Subtle is the Lord: The science and the life of Albert Einstein*. Oxford University Press.

[Paul, 2012] Paul, L. A. (2012). Building the world from its fundamental constituents. *Philosophical Studies*, 158(2):221–256.

[Peacocke, 2014] Peacocke, C. (2014). Ontology and Intelligibility. Draft of October 2014.

[Pitts, 2006] Pitts, J. B. (2006). Absolute objects and counterexamples: Jones-Geroch dust, Torretti constant curvature, tetrad-spinor, and scalar density. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 37(2):347–371.

[Pooley, 2002] Pooley, O. (2002). *The reality of spacetime*. Dphil thesis, University of Oxford.

[Pooley, 2006] Pooley, O. (2006). Points, particles, and structural realism. In Rickles, D., French, S., and Saatsi, J., editors, *The Structural Foundations of Quantum Gravity*, pages 83–120. Oxford University Press, Oxford.

[Pooley, 2013a] Pooley, O. (2013a). The Reality of Spacetime. Unpublished draft.

[Pooley, 2013b] Pooley, O. (2013b). Substantivalist and relationalist approaches to spacetime. In *The Oxford Handbook of Philosophy of Physics*, pages 522–586. Oxford Univerity Press, Oxford.

[Pooley, 2015] Pooley, O. (2015). Background independence, diffeomorphism invariance, and the meaning of coordinates. In Lehmkuhl, D., Schiemann, G., and Scholz, E., editors, *Towards a theory of spaceme theories*, number 13 in Einstein Studies. Birkhäuser, Basel.

[Putnam, 1983] Putnam, H. (1983). Equivalence. In *Realism and Reason*, volume 3 of *Philosophical Papers*, pages 26–45. Cambridge University Press, Cambridge.

[Quine, 1948] Quine, W. V. (1948). On what there is. *The Review of Metaphysics*, 2(5):21–36.

[Quine, 1953a] Quine, W. V. O. (1953a). Reference and Modality. In *From a Logical Point of View*. Harvard University Press, Cambridge, MA.

[Quine, 1953b] Quine, W. V. O. (1953b). Three Grades of Modal Involvement. In *Proceedings of the XIth International Congress of Philosophy*. North-Holland Publishing Co.

[Quine, 1969] Quine, W. V. O. (1969). Ontological Relativity. In *Ontological Relativity and Other Essays*, pages 26–68. Columbia University Press, New York.

[Rickles, 2008] Rickles, D. (2008). *Symmetry, structure, and spacetime*. Elsevier, Amsterdam.

[Roberts, 2008] Roberts, J. T. (2008). A Puzzle about Laws, Symmetries and Measurability. *The British Journal for the Philosophy of Science*, 59(2):143–168.

[Rosenstock et al., 2015] Rosenstock, S., Barrett, T. W., and Weatherall, J. O. (2015). On Einstein algebras and relativistic spacetimes. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52, Part B:309–316.

[Ruetsche, 2011] Ruetsche, L. (2011). *Interpreting quantum theories: the art of the possible*. Oxford University Press, Oxford; New York.

[Russell, 1910] Russell, B. (1910). Knowledge by acquaintance and knowledge by description. In *Proceedings of the Aristotelian Society*, volume 11, pages 108–128. JSTOR.

[Russell, 2013] Russell, J. S. (2013). Actuality for Counterpart Theorists. *Mind*, 122(485):85–134.

[Russell, 2015] Russell, J. S. (2015). Temporary Safety Hazards. *Noûs*, pages n/a–n/a.

[Russell, 2016] Russell, J. S. (2016). Quality and Quantifiers. Draft of January 2016.

[Rynasiewicz, 1992] Rynasiewicz, R. (1992). Rings, Holes and Substantivalism: On the Program of Leibniz Algebras. *Philosophy of Science*, 59(4):572–589.

[Rynasiewicz, 1994] Rynasiewicz, R. (1994). The Lessons of the Hole Argument. *The British Journal for the Philosophy of Science*, 45(2):407–436.

[Saunders, 2003] Saunders, S. (2003). Physics and Leibniz's principles. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 289–308. Cambridge University Press, Cambridge.

[Saunders, 2013] Saunders, S. (2013). Rethinking Newton's Principia. *Philosophy of Science*, 80(1):22–48.

[Saunders, 2016] Saunders, S. (2016). On the emergence of individuals in physics. In Guay, A. and Pradeau, T., editors, *Individuals Across the Sciences*, pages 165–192. Oxford University Press, Oxford.

[Schreiber, 2015a] Schreiber, U. (2015a). smooth algebra (Rev #46) in nLab. `https://ncatlab.org/nlab/revision/smooth+algebra/46`. For most recent version, see https://ncatlab.org/nlab/show/smooth+algebra.

[Schreiber, 2015b] Schreiber, U. (2015b). synthetic differential geometry (Rev #63) in nLab. `https://ncatlab.org/nlab/revision/synthetic+differential+geometry/63`. For the most recent version, see https://ncatlab.org/nlab/show/synthetic+differential+geometry.

[Sider, 1996] Sider, T. (1996). All the World's a Stage. *Australasian Journal of Philosophy*, 74(3):433–453.

[Skow, 2005] Skow, B. (2005). *Once upon a spacetime*. Ph.d. thesis, New York University.

[Stevens, 2015] Stevens, S. (2015). The Dynamical Approach as Practical Geometry. *Philosophy of Science*, 82(5):1152–1162.

[Swanson and Halvorson, 2012] Swanson, N. and Halvorson, H. (2012). On North's "The Structure of Physics". Unpublished note.

[Trautman, 1965] Trautman, A. (1965). Foundations and Current Problems of General Relativity. In *Lectures on General Relativity*, volume 1 of *Brandeis Summer Institute of Theoretical Physics 1964*. Prentice-Hall, Englewood Cliffs.

[van Fraassen, 1989] van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford University Press, Oxford; New York.

[Wallace, 2015a] Wallace, D. (2015a). Fields as Bodies: a unified presentation of spacetime and internal gauge symmetry. *arXiv:1502.06539 [gr-qc, physics:hep-th]*.

[Wallace, 2015b] Wallace, D. (2015b). Fundamental and emergent geometry in Newtonian physics. Unpublished draft (of 11.09.2015).

[Wallace, MS] Wallace, D. (MS). Who's Afraid of Coordinate Systems?

[Weatherall, 2012] Weatherall, J. O. (2012). Inertial motion, explanation, and the foundations of classical spacetime theories. In Lehmkuhl, D., Schiemann, G., and Scholz, E., editors, *Towards a theory of spaceme theories*, number 13 in Einstein Studies. Birkhäuser, Basel. Draft of July 2012; available at arXiv:1206.2980 [physics.hist-ph].

[Weatherall, 2015a] Weatherall, J. O. (2015a). Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent? *Erkenntnis*. Published online.

[Weatherall, 2015b] Weatherall, J. O. (2015b). Fiber bundles, Yang–Mills theory, and general relativity. *Synthese*. Published online.

[Weatherall, 2015c] Weatherall, J. O. (2015c). Maxwell-Huygens, Newton-Cartan, and Saunders-Knox Space-Times. *Philosophy of Science*, 83(1):82–92.

[Weatherall, 2016] Weatherall, J. O. (2016). Regarding the 'Hole Argument'. *The British Journal for the Philosophy of Science*, page axw012.

[Weatherall, MS] Weatherall, J. O. (MS). Understanding Gauge. Forthcoming in Philosophy of Science; available as arXiv:1505.02229 [physics.hist-ph].

[Williamson, 2013] Williamson, T. (2013). *Modal Logic as Metaphysics*. OUP Oxford.

[Wilson, 1993] Wilson, M. (1993). There's a Hole and a Bucket, Dear Leibniz. *Midwest Studies In Philosophy*, 18(1):202–241.