Interpretation and Equivalence; *or*, Equivalence and Interpretation

Neil Dewar

March 10, 2017

> Here as always the theory itself sets the framework for its interpretation.
>
> Everett (1957)

Introduction

Philosophers of science spend a lot of time "interpreting" scientific theories. In this paper, I try to get a handle on what it is they might be up to. My main contention is that a certain picture of interpretation is widespread (though implicit) in contemporary philosophy of science: a picture according to which interpretation of theories is relevantly analogous to the interpretation of foreign literature. On this picture, which we might call the external account of theory-interpretation, meaning is to be imported into the equations by putting them in correspondence with some discourse whose signs and symbols are already endowed with significance. Of course, the prevalence of this picture wouldn't be much of a problem if that picture were the only way to think about interpretation, or was clearly the best way to do so (though even then, there would be a value to bringing it out into the light). I contend, however, that it is neither. There is an alternative way of thinking about interpretation—what we can call the "internal" account of interpretation—which instead takes interpretation to be a matter of delineating a theory's internal semantic architecture. At a minimum, I hope to convince you the internal picture highlights an aspect of interpretation that

we are otherwise at risk of neglecting. But I also aim to show that the internal picture offers a richer and more satisfying account of interpretation than the external picture does.

The paper proceeds as follows. I start (section 1) by assembling various platitudes about what interpretation is for, so that we can get a bead on the notion we are after. Section 2 outlines the external account of interpretation in more detail, looking in particular at three examples of the form: the reductionist approach to interpretation found in Carnap's *Aufbau*, the quest for primitive ontology in quantum mechanics, and the interpretation of physics by metaphysics. I take a slight step back in section 3, to explore the question of what it is we are interpreting—that is, what I am taking a "theory" (prior to interpretation) to be. This lays the groundwork for the internal account of interpretation, which I give in section 4. Section 5 extends the internal account to discuss inter-theoretic relations; and section 6 confronts the issue of how interpretation relates to representation.

1 The role of interpretation

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.

First, interpreting a theory is a necessary component of determining 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. So an uninterpreted theory is not 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.[1] Maybe, after analysis, we will succeed in showing that there is no such multiplicity of interpretative options—but doing so will only be

[1] cf. Jones (1991).

possible after the application of some philosophically rich account of interpretation, so we are still required to develop such an account.

Second, for this reason, the notion of interpretation is crucial in explicating (certain forms of) the realist-antirealist debate.[2] 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] The first factor 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).

Similarly, anti-realists may characterised by their attitudes towards how best to interpret theories (or whether to interpret theories at all). 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. Only quietists are marked out as those who think that scientific theories ought not to be interpreted at all; and even then, they will presumably think that the *observational* parts of a theory require interpretation, at least if the theory is to be tested or used. Thus, attitudes towards the practice of interpretation (compulsory vs. 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.

Third, 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

[2] cf. Stanford (nd).

[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.

that these are virtues of *interpreted* theories.[5] So not only is interpretation important for *understanding* realism, it is also 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.

Finally, there is a close relationship between equivalence and interpretation (a relationship which will be of much concern to us in this paper). The heart of the notion of theoretical equivalence is a certain sort of *ecumenicism* with regards to equivalent theories: if theories A and B 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 determinations of equivalence are interesting and important, since they will tell us when we do or don't need to make choices amongst theories. But it also makes clear that interpretation and equivalence are closely associated notions: for a pair of uninterpreted theories, there is no sense to be made of the question of whether or not they are equivalent, since (as discussed above) they do not have unambiguous rosters of commitments.

2 The external approach to interpretation

So what kind of practice could interpretation be, which would have the effects considered above? In this section, I want to articulate and critique one answer to this question, which is popular but—I contend—flawed. This is the "external" account sketched above, which takes the interpretation of a theory to be, in some relevant sense, analogous to the interpretation of a passage written in a foreign language. "Interpreting" such a passage is a matter of translating it into the home language, and thereby coming to understand the passage by dint of one's facility with the home language. The analogous move, in the case of scientific theories, is to exploit some antecedent semantic facility in order to come to understand (i.e., to interpret) the target theory. This is all very vague; with luck, the following examples will help spell out the phenomenon I have in mind.

First, consider attempts to reduce scientific (and ordinary) discourse to a phenomenological basis: for example, Russell's project in *Our Knowledge of the External World* (Russell, 1993). Russell is motivated by epistemic considerations, in particular a concern to ward off skeptical doubt:

> We are thus led to a distinction between what we may call "hard" data and

[5] I take this observation from Ruetsche (2011, Chapter 1).

> "soft" data. [...] I mean by "hard" data those which resist the solvent influence of critical reflection, and by "soft" data those which, under the operation of this process, become to our minds more or less doubtful. The hardest of hard data are of two sorts: the particular facts of sense, and the general truths of logic.[6]

If it is only the immediate objects of sensory experience and the truths of logic that enjoy primitive epistemic privilege, then (claims Russell) the only way for science to enjoy that same privilege is if the objects of science are, in fact, logical constructs from the objects of sense: "it may be laid down quite generally that *in so far* as physics or common sense is verifiable, it must be capable of interpretation in terms of actual sense-data alone."[7]

We need not be concerned with this (rather dubious) epistemic motivation for the project. Rather, we should be interested in the project itself: specifically, Russell's characterisation of it as providing an *interpretation* in terms of sense-data. So although the reductionist process has in mind an epistemic goal, the goal is to be accomplished by semantic means, by providing a certain sort of account of what the theory is about. In Russell's hands, it doesn't seem to be a requirement on the *coherence* or *intelligibility* of a theory that it be cashed out into the currency of experience—merely a requirement on its *knowability*. But it would not take long for the means and ends of such reductionism to be brought together. After reading *Our Knowledge of the External World* in 1921, Carnap was inspired to undertake his own version of the reductionist project, culminating in 1928's *Der Logische Aufbau der Welt* (Carnap, 1967).

As with Russell, the overall project is to show how all the objects of science may be constructed from the "autopsychological" basis of first-personal experience. This basis is comprised of "erlebs", primitive and elementary such experiences, standing in relations of recollected similarity; from these austere ingredients, we are to construct first the world of physical objects, then the "heteropsychological" world of third-personal mental configurations, and finally the world of sociocultural institutions. Unlike Russell, however, the core motivation for such a construction is not (or at least, not only) that of showing how our knowledge of the constructed world derives from our knowledge of the constructive basis. There is now a further notion that this will show how the *meaning* of discourse concerning the constructed world is cooked up out of the meaningfulness of terms regarding the basis. As Carnap put it in an

[6] Russell (1993, pp. 77–78)
[7] Russell (1993, pp. 88–89)

unpublished lecture,

> Quite generally, everything that we talk about must be reducible to what I have experienced. Everything that I can know refers either to my own feelings, representations, thoughts and so forth, or it is to be inferred from my perceptions. Each meaningful assertion, whether it concerns remote objects or complicated scientific concepts, must be *translatable* into a statement that speaks about contents of my own experience and, indeed, at most about my perceptions.[8]

So what we have here is a particular story about where meaning comes from, informing and underpinning a particular way of imbuing theories with content. According to this story, meaning flows in the first instance from experience; and so, the ultimate topic of all meaningful (interpreted) discourse must be sensory experience itself. So we see an intimate relationship between the positivist or empiricist account of meaning, and the associated conception of what is involved in interpreting a theory. Note that this work of Carnap's, and the broader positivist program of which it is a part, exemplifies the connections we canvassed in section 1 above between interpretation, commitment, and equivalence. A theory's true commitments are, it is suggested, exhausted by the claims it makes about what is observable (identified, in the positivist program, with the claims statable in the observation-language).[9] And what it is for two theories to be equivalent is just for them to have the same observational consequences: empirical equivalence is a sufficient condition for theoretical equivalence.[10]

(A brief digression: there are subtleties about the extent to which a project such as Russell's or Carnap's is really best understood as translation into another language, except in a very broadly analogous sense.[11] At least on some accounts, sense-data or erlebs are objects with intrinsic semantic content, insofar as grasping the inherent properties of such an object is simply what it is to manifest a thought of the relevant associated kind; languages, by contrast, are symbolic systems whose semantic content arises from their relation to other objects, not from intrinsic or inherent features of the elements of the system themselves. That said, I think the analogy is robust enough

[8](Carnap, 1929, p. 12); quoted and translated in (Coffa, 1991, p. 227).

[9]Hence the significance of Craig's theorem, insofar as it was taken to show that one could find a recursively axiomatisable theory capturing just the "observational content" of any other theory (see Craig (1953), or Putnam (1965) for critical discussion).

[10]See e.g. Reichenbach (1938), or Putnam (1983) for a critique.

[11]I thank Erik Curiel for pressing me on this.

that my criticisms of the external approach do carry over—indeed, it seems to me that the claimed difference between the autopsychological basis and (other) languages (on the grounds that the former, but not the latter, are endowed with intrinsic semantic content) is exactly what is at issue. If I'm wrong about this, however, then one can just take my arguments to apply to a version of these phenomenological reduction-projects, where the process of reduction is thought of as a form of translation in some more robust sense than that envisioned by Russell or Carnap.)

Actually carrying out a project such as Carnap's, however, turns out to be fraught with difficulties. The main problem is that scientific discourse does not, in general, associate to each concept it employs a distinctive or canonical class of observable "indicators", or "criteria", or "verification-conditions"; and even in the (rather artificial) cases where such indicators are to be had, there may be further barriers to uniquely associating indicators with purely phenomenological data. For example, radioactive decay *may*, under appropriate circumstances, be associated with the clicking of a Geiger counter: but it is not always so associated (not even in all experimental contexts where radiation is successfully detected), and it is hard to spell out "the clicking of a Geiger counter" in terms of pure autopsychology. It should be emphasised that the post-*Aufbau* Carnap was well aware of these problems; they were a significant motivation for the replacement of explicit definition by reduction sentences or correspondence rules. Even these liberalised versions of the doctrine, however, still take for granted that interpretation of a theory is a matter of relating it to an external empirical basis.

At the same time, the popularity of the epistemic or semantic theses motivating these projects has severely waned. Claims that we only "really" have knowledge of that with which we are immediately acquainted, or that we only "really" understand claims about the immediate contents of experience, are (for whatever reason) nowhere near as widespread as they once were. However, this doesn't mean that the external approach to interpretation has gone away—just that it takes a different form. As a second example, consider the primitive-ontology approach to quantum mechanics. Advocates of this approach often stress the problems with explicating a theory's (empirical) content in terms of its phenomenological implications.[12] Nevertheless, there is an important continuity. Maudlin's account of the relationship between the two approaches is exemplary, and worth quoting at length:

[12]Dialectically, this is because such explications are often associated with Copenhagen-style interpretations of quantum theory, of the kind which primitive ontology seeks to oppose.

There was a reasonable concern behind all this foolery [i.e., the project of reducing physics to phenomenalist terms]. In order to be of interest, physical theories have to make contact with some sort of evidence, some grounds for taking them seriously or dismissing them. And the acquisition of evidence by humans clearly does involve experience at some point. So it is not surprising that one might focus on how physical claims relate to experience in an attempt to get a handle on the problem of evidence. But for all that, it turns out to be the wrong handle to grasp since the connection between physical descriptions and experience has never been made precise enough to admit analysis.

Rather, in classical physics the evidential connection is made between the physical description and a certain class of *local beables*, such as the positions of macroscopic objects. [...] Our ability to reliably observe such facts [i.e., facts about the local beables] is not itself derived from the physics: it is rather a presupposition used in testing the physics. So the contact between theory and evidence is made exactly at the point of some local beables: beables that are predictable according to the theory and intuitively observable as well.

The pre-theoretical intuition that certain physical states of affairs are unproblematically observable is not couched in the terminology of a physical theory: it is couched in everyday language. If Galileo drops rocks off the Leaning Tower, what is important is that we accept that it is observable *when the rocks hit the ground*. If the physical theory itself asserts that rocks are made up of atoms, then it will follow *according to the theory together with intuition* that we can observe when certain collections of atoms hit the ground, but this latter is clearly not the content of the observation. If the theory says instead that rocks are composed of fields, then it will follow that we can observe when certain fields hit the ground, or when the field values near the ground become high. It is easy enough to see how to translate the claim that we can see the rocks into the proprietary language of atomic physics or continuum mechanics or string theory. But the critical point is that *the principles of translation are extremely easy and straightforward when the connection is made via the local beables of the theory*. Collections of atoms or regions of strong field or regions of high mass density, because they are local beables, can unproblematically be rock-shaped and

> move in reasonably precise trajectories. If the theory says that this is what rocks really *are*, then we know how to translate the observable phenomena into the language of the theory, and so make contact with the theoretical predictions.[13]

Let's count the steps here. First, there is the claim that the empirical content of a theory is better identified with its implications for the behaviour of macroscopic objects, rather than its implications for sense-experience. Then follows the observation that we already have a language for talking about such objects: namely, English (or French, or Chinese, or whatever). So to pick out the implications of the theory for such objects is—perhaps *inter alia*—to put certain terms of the theory into correspondence with certain terms in English (or whatever). This idea is well taken, and we will return to it in §4 below. Second, there is the observation that this correspondence is reasonably straightforward when the theory contains designated local beables. For, given the local beables, we may give a straightforwardly mereological account of how to accomplish this correspondence: if rocks, tables, etc., are composed of the local beables, then "rock" is just translated as "rock-shaped collections of local beables".

Where the external approach comes in, however, is in the conclusion drawn from these claims: that the local-beables portion of the theory's language acquires meaning by being translated into ordinary English, with the rest of the theory then acquiring meaning from its implications for the behaviour of those beables—and hence, possessing meaning only insofar as it has implications for those beables. Thus Dürr, Goldstein and Zanghí write:

> According to (pre-quantum-mechanical) scientific precedent, when new mathematically abstract theoretical entities are introduced into a theory, the physical significance of these entities, their very meaning insofar as physics is concerned, arises from their dynamical role, from the role they play in (governing) the evolution of the more primitive—more familiar and less abstract—entities or dynamical variables. For example, in classical electrodynamics the *meaning* of the electromagnetic field derives solely from the Lorentz force equation, i.e., from the field's role in governing the evolution of the positions of charged particles, through the specification of the forces, acting upon these particles, to which the field gives rise; while in general relativity a similar statement can be made for the

[13](Maudlin, 2007b, p. 3158–3159)

> gravitational metric tensor. That this should be so is rather obvious: Why would these abstractions be introduced in the first place, if not for their relevance to the behavior of *something else*, which somehow already has physical significance?[14]

The result of all this is that for theories without local beables, there is no interpretative project available. If a theory does not posit a "primitive ontology" of local beables, then it is uninterpretable, since there is nothing to be translated into English. So the primitive ontology plays a privileged role in investing the theory with content: "the fundamental requisite of the [primitive ontology] is that it should make absolutely precise what the theory is fundamentally about";[15] "ignoring [the primitive ontology of particle positions in Bohmian mechanics], the theory becomes a theory about nothing";[16] "in a particle theory, [. . .] particle positions are what the theory is about. The role of all other variables is to say how the positions change."[17] Thus, interpreting a theory is a matter of identifying the primitive ontology of the theory (or providing it with one, if none is forthcoming); the "extremely easy and straightforward" mereological translation into ordinary language then gives meaning to the claims the theory makes about the primitive ontology, and thence to the theory as a whole. In other words, we have here a project that is structurally very similar to the Russellian or Carnapian project discussed above—only with the phenomenological basis replaced by a basis of material objects in space and time.[18]

Again, we find a close-knit web of connections between interpretation, commitment and equivalence. For example, Allori et al. (2008) "suggest that two theories be regarded as physically equivalent when they lead to the same history of the PO [primitive ontology]".[19] And an interesting recent trend in the primitive-ontology literature is towards treating other aspects of a theory besides the primitive ontology—such as the wavefunction or the electromagnetic fields—as not fully part of the theory's

[14](Dürr et al., 1992, pp.848–849)

[15]Ghirardi (2016)

[16](Dürr, 2008, p. 117); the context makes it reasonably clear that the claim generalises to other forms of primitive ontology.

[17](Dürr and Teufel, 2009, p. 38)

[18]In this connection, note that the *Aufbau* is more pluralist about the choice of basis than one might expect. In particular, Carnap explicitly allows that one could use a physical basis (such as (§62) that consisting of elementary material particles or spacetime points), rather than a psychological one, and notes that such a system"would have the advantage that it uses as its basic domain the only domain (namely, the physical) which is characterized by a clear regularity of its processes." (Carnap, 1967, §59)

[19](Allori et al., 2008, p. 365)—although as discussed in n. 37 below, they also seem open to applying the converse direction.

commitments.[20]

Finally, I want to adduce one more example of the external approach—or rather, not so much a specific example, as a suggestion that the external approach is implicit in much of the practice of contemporary philosophy of science. I have in mind the pervasive metaphor of interpretation as a matter of "picturing" our scientific commitments—where providing such a picture is often understood as a matter of giving the metaphysics to accompany the science. For example, Chakravartty holds that

> The neglect of metaphysics in the context of realism [...] is a mistake. For there is a sense in which the metaphysics of science is a precursor to its epistemology. One cannot fully appreciate what it might mean to be a realist until one has a clear picture of what one is being invited to be a realist about.[21]

As discussed in the introduction, I am sympathetic to the claim that interpretation is a precondition of measuring what commitments a realist is getting themselves into; but here, that is coupled to an image of interpretation as the provision of a "picture". Note that this image of interpretation transcends divisions over what kind of metaphysics is appropriate—it is a higher-level, more methodological, commitment than that. Thus French, despite disagreeing strongly with Chakravartty over what kind of metaphysics scientific realism demands, agrees that providing such a metaphysics is a precondition for realism (and by extension, for interpretation):

> In order to obtain Chakravartty's clear picture and hence obtain an appropriate realist understanding of the world we need to clothe the physics in an appropriate metaphysics. Those who reject any such need are either closet empiricists or ersatz realists.[22]

This methodological stance then has implications for what kinds of projects are worth pursuing in the philosophy of science, and how we should go about pursuing them. For instance, I've stressed above that one characteristic feature of external interpretations is that they induce a criterion of equivalence: for two theories, or models of a theory, or sentences of a theory (etc.) to be equivalent is for the external interpretation to assign them the same content. Coffey (2014) has argued that this demonstrates,

[20] See e.g. Miller (2014), Callender (2014), Esfeld (2014), or Bhogal and Perry (2015).
[21] (Chakravartty, 2007, p.26)
[22] (French, 2013, p. 85)

quite generally, that there is no interesting independent question of when two theories are equivalent:

> For those of us who think sense can be made of a theory's physical content beyond what the theory says about the empirically confirmable or disconfirmable—in short, for those of us who take the interpretive project seriously in the philosophy of physics—there's a natural and seemingly simple account of theoretical equivalence [. . .]:
>
> > Two theoretical formulations are theoretically equivalent exactly if they say the same thing about what the physical world is like, where that content goes well beyond their observable or empirical claims. Theoretical equivalence is a function of interpretation. It's a relation between completely interpreted formulations.
>
> Insofar as we can understand the physical pictures provided by different interpreted formalisms, and insofar as we're capable of comparing those pictures, we can straightforwardly determine whether two interpreted formulations are theoretically equivalent, whether they say the same thing about what the physical world is like.[23]

I think Coffey is essentially correct here: on the external account, there is little that can be independently said about equivalence. However, as we'll see, I take this to be an artefact of that account, rather than an indication that inquiries into conditions of equivalence are misdirected.

We now have enough examples to make clear the overall character of the external approach—and my concerns about it. To my mind, there are two problems that are especially pressing. First, since this approach involves pretheoretically privileging some particular model of description, it gives rise to naturalist 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.[24] We should be extremely skeptical that the reflections

[23](Coffey, 2014, pp. 834–835)
[24]e.g. Egg and Esfeld (2014), Esfeld et al. (2014)

of philosophers will offer a better mechanism for theory choice in science than the practice of science does.

Second, there is a concern about circularity. We are confronted by a pair of theories (say T_1 and T_2), and we inquire into what kind of relationship obtains between them: whether they are equivalent, say, or whether one theory is some kind of limiting case of the other. On the external approach, these questions should be deferred until after the theories have been interpreted, i.e., until we have settled on an account of what the theories say. But when we look at any specific such account, what form does it have? Well, just some more sentences—that is to say, just more theory![25] So the status of this third theory—call it T_I—as an interpretation of T_1 or T_2 depends on the propriety of claiming that T_I stands in some appropriate kind of relationship to T_1 or T_2. But of course, articulating the conditions for such relationships to hold was exactly the task with which we began. So whatever the virtues of introducing T_I might be, it cannot be the case that doing so obviates the need for articulating those conditions.

These observations suggest that the external approach puts the cart before the horse: that the "rendering" of a theory into a particular language or framework is the *end* of an interpretative analysis, rather than merely constituting it. Before I spell this claim out in more detail, though, I want to take a bit of a step back, and consider just what it is we're interpreting. Since the internal account thinks of interpretation as a matter of altering a theory's internal structure, rather than in postulating connections to external fixed points, it will behove us to say a little more about the structure of theories.

3 What is the structure of a theory?

The standard take on this question holds that we have two available choices. We can take a *syntactic* view of theories, according to which theories are comprised by sets of sentences, formed and manipulated according to some appropriate formal calculus. Or we can take a *semantic* view of theories, according to which theories are composed of sets of models. In this section, I suggest that we need make no such choice: rather, we should take a theory to comprise both syntactic and semantic elements. Considering the sentences in isolation from the models, or the models in

[25] It's in this sense that the internal approach to interpretation makes contact with better-known doctrines under the "internal" label, such as Putnam's internal realism.

isolation from the sentences, will fail to capture everything of interest.[26]

Let's consider an example theory. And let's take about the simplest example possible: the theory of a single Newtonian particle. 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:

$$m\frac{d^2x}{dt^2} = -\frac{dV}{dx} \tag{1}$$

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

$$f(t) = at + b \tag{2}$$

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). 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 (2) of (1).

Thus, our toy theory could be described as a set of syntactic conditions, *together with* an account of the structures to which those conditions apply, and of what it would be for them to be satisfied. It is for this reason that I take both the syntactic and

[26] In this, I follow Halvorson (2012), Halvorson (2013), and Lutz (2015).

semantic views to be a poor fit for the actual character of theories, at least if those views are taken at face value. I think it matters what semantic constructions 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 satisfying some specific set of conditions. Moreover, I am quite happy with the idea that these models are "yoked to a particular syntax":[27] the spaces T and X are explicitly labelled (by the variables t and x respectively), in order to make manifest how to assess whether the condition (1) holds of a given function. (Although we will return to the issue of language-independence in §5 below.) All this said, I don't wish to rule out the notion that some more subtle conception of the syntactic or semantic view 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.[28] I merely wish to signal that it does not, so far as I can see, coincide with thoughtless versions of either.

One traditional difficulty with relating the philosophical literature to the practice of science is the former's focus on theories formulated in terms of the first-order predicate calculus, despite the paucity of such theories in scientific practice. At least within physics, one is far more likely to come across laws that—as in the example above—take the form of differential equations, governing how systems evolve over time, how fields may be distributed over spacetime, etc.[29] However, the differences between first-order theories, and theories stated in terms of differential equations, should not be overstated. In fact, there are a series of useful and illuminating analogies between the two formalisms, which can guide us in how concepts from the one can be usefully applied to the other—and which indicate that an account of interpretation should be applicable to theories in *either* form.

To see this comparison, recall that a "theory" in first-order model theory is typically taken to be a set of sentences of a specified first-order language. Such a language may be identified with the set of well-formed formulae generated from a particular signature (set of relation- and function-symbols) Σ, according to the recursive syntax rules of the predicate calculus. That sounds a lot like the syntactic conception of

[27] (van Fraassen, 1989, p. 366)

[28] 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.

[29] cf. Maudlin (2007a).

theories. But model theory, of course, is not interested in such sets of sentences in isolation. Say that a Σ-*theory* is a set of sentences of signature Σ. Then a Σ-*structure* $\mathcal{S}$ is a set S, equipped with "interpretations" of the elements of Σ (maps from relation-symbols to relations over S, and from function-symbols to functions over S). $\mathcal{S}$ may make Σ-sentences true or false via the standard Tarskian clauses. If a Σ-structure $\mathcal{M}$ makes all the sentences of a Σ-theory $\mathbb{T}$ true, then $\mathcal{M}$ is said to be a *model* of $\mathbb{T}$; the class of all models of $\mathbb{T}$ is denoted $\mathrm{Mod}(\mathbb{T})$. So model theory, as the name suggests, is interested in analysing the various relationships between sets of sentences *and their models*.[30] Hence, a theory in the sense of model theory exhibits the same tripartite structure that we saw a moment ago. There is a specification of the kinds of mathematical structures that will be used for representation (i.e. Σ-structures). 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: we can think of a model of a first-order theory as describing the distribution of certain properties and relations over a set of individuals, and we can think of a model of the Newtonian theory as describing the distribution of a monadic determinable property (position) over some set of individuals (particle-stages).[31] I take this to be prima facie evidence that the form I describe for theories in general (a set of syntactic conditions governing some mathematical structures of an appropriate type) is indeed an appropriately generic form for theories to take. Hence, I will suppose that this kind of form is an appropriate target for our account of interpretation. I now turn to giving that account.

4 Models and modality

So, suppose that we are presented with a theory in the form above (i.e., a theory comprising both syntactic conditions and an appropriate model theory). The models bestow truth-values on the sentences of the theory in some kind of appropriately

[30] I intend this to include relationships that hold between sets of sentences in virtue of their models: for instance, the relation of logical equivalence (i.e., of having the same models).

[31] 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.

systematic way. However, in interpreting the theory, we need not take all of the aspects of the theory to faithfully encode commitments required when believing the theory. That is, a crucial feature of interpretation is that there is scope to treat some aspects of the theory as (mere) *artefacts*. Thus, for example, the "facts" about which specific coordinates an object occupies in a coordinate-based model of some physical system are typically regarded as merely artefactual: accepting the theory from which this model is taken does not mean accepting that there are genuine physical correlates to such facts. The project of *internal interpretation* is exactly this separation of artefactual and representational features in the theory's models. To illustrate, let's look at some examples of doing so, in order to demonstrate how the separation can be done in a suitably internal fashion.

First, consider the case of *isomorphic* models. At least if we are using standard mathematical tools,[32] models can be distinct whilst still being isomorphic: perhaps one model has a domain comprising the natural numbers as its domain, whereas its isomorphic cousin has the integers. But it has seemed plausible to many that we should be sceptical that this distinction corresponds to any difference. For one thing, this view gets motivation from "anti-haecceitist" doctrines, i.e., views which deny that there are any metaphysically substantive facts about the "intrinsic identities" of objects or individuals (above and beyond their qualitative profiles).[33] Since isomorphic models agree on the distribution of qualitative properties, argues the anti-haecceitist, they are representing the same possible world. Alternatively, or more generally, one can also argue for the equivalence of isomorphic models from considerations about the very nature of representation by mathematical structures. On this view, even if one is a haecceitist, one should still interpret isomorphic models as representing the same possible world—it's just that one should include representatives of non-qualitative properties in one's models, so that models which are qualitatively isomorphic need not be isomorphic *tout court*.[34] Thus, debates about what we should take the content of our theories to be, and over what kind of representational role certain aspects of a theory might permissibly have, get cashed out in the question of whether or not certain models of the theory ought to be regarded as equivalent.

As a second example, consider *symmetries* in physics: say, the gauge symmetry of electromagnetism. Recall that this symmetry arises when the electromagnetic

[32] Rather than, say, homotopy type theory (see The Univalent Foundations Program (2013)).
[33] Kaplan (1975), Pooley (2006)
[34] This is how I understand Weatherall (2016)'s critique of the usual dialectic surrounding the "Hole Argument".

potential is characterised as a 1-form A_a on (say) Minkowski space M. The equations of the theory are invariant under the transformation

$$A_a \mapsto A_a + \nabla_a \Lambda \tag{3}$$

where Λ is any smooth scalar function and ∇_a is the derivative associated with the Minkowski metric. As a result, given any model (solution) of the theory, we can obtain another solution by transforming A_a as in (3).

Generally, gauge-related models are understood as physically equivalent to one another.[35] However, it is controversial whether this means that such models can be *interpreted* as equivalent (so that someone could continue to use the original theory whilst affirming that gauge-related models are equivalent), or whether this is merely a way of saying that we ought to seek some alternative theory in which the models are isomorphic (or even identical).[36] Again, therefore, a dispute about interpretation gets parlayed into a dispute about whether certain kinds of models should be regarded as equivalent to one another.

So the internal approach holds that we should, in general, understand interpretive disputes as disputes over what kinds of equivalences hold amongst the models of a theory. That is, in interpreting a theory, we *begin* by making determinations of equivalence, and use those determinations to get a fix on the theory's commitments. We do this by employing the following principle, the converse of Coffey's: the theory is committed to whatever is invariant across equivalences, i.e., to all and only that which is shared by equivalent models.[37] Thus, on the internal view, interpreting a theory is a matter of postulating certain equivalences between elements of the model theory, abstracting away from the differences between the (declared-to-be) equivalent models.

I now want to defend a further claim about the results of this process of abstraction: namely that, at least for theories regarded as describing the world (on which more below), these results are naturally understood as possible worlds ("possible", that is, in the sense of being nomologically possible relative to taking the claims of the theory

[35] Exactly why we should do so is a matter of some dispute: see e.g. Saunders (2003), Roberts (2008), Baker (2010), Dasgupta (2014), Caulton (2015), and references therein.

[36] See Dewar (2015), Møller-Nielsen (nd), Dewar (nd) for discussion.

[37] It should be noted that Allori et al. (2008) are sympathetic to such an idea. The quotation given in §2 above continues, "Conversely, one could define the notion of PO [primitive ontology] in terms of physical equivalence: The PO is described by those variables that remain invariant under all physical equivalences." (Allori et al., 2008, p. 365)

as laws). This expresses the fact that we generally explicate theory-relative possibility by looking 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 simultaneously occupy an eigenstate of the position and momentum operators? No, because there is no model of the theory in which that is the case. But we do not straightforwardly associate models with possibilities, in a one-to-one fashion. Diffeomorphic models of General Relativity are standardly taken to represent the same possibility, as are a corresponding pair of wave-mechanical and matrix-mechanical models of quantum mechanics. So we should not identify the possible worlds with the models themselves, but rather with the results of abstracting from the models by the equivalence relation postulated in interpreting the theory. This suggestion provides the standard link between interpretation and modality: in an interpreted theory, equivalent models are those which represent the same possible world. In contrast to the standard account, however, our grasp of the possible worlds *follows* (or rather, is provided by) our postulation of the equivalence-relations between models.[38]

A brief digression is in order here, regarding the history of modal semantics. The view just described may remind the reader of Carnap's proposal to explicate necessity in terms of logical truth;[39] and this might sound like something of a problem, given that Carnap's modal semantics are generally agreed to be problematic. 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.[40]

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

[38]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 the view Lewis describes as "pictorial ersatzism" (Lewis, 1986, §3.3), although even that is only a partial match. (Which may be for the best, given that pictorial ersatzism seems to generally be reckoned implausible: e.g. "[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).)

[39]Carnap (1956)

[40](Carnap, 1956, p. 174)

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,[41] suffice to determine the sentence as true (respectively, false), then the sentence in question is L-true (respectively, L-false).

Here, I definitely part company from Carnap. The account of possible worlds given here is intended to furnish us with a *specific* Kripke model, i.e., that in which the worlds are the models of the associated theory. But attention to one Kripke model 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]."[42] This results from the fact that he is considering only one Kripke model $\mathcal{K}$ (that in which the worlds are all and only the Tarski-models of first-order logic), and identifies logical validity with truth in $\mathcal{K}$, rather than truth in all Kripke models. So, because P is true in some Tarski-models, $\Diamond P$ 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:[43] $(Q \wedge \neg Q)$ 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.

[41] Which are the same as those used here, except that Carnap employs a substitutional account of quantification.

[42] (Carnap, 1956, p. 175)

[43] I take this observation from (Williamson, 2013, §2.8).

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 analysis proposed here is that it is about the only means I can imagine by which we could have any access to those facts at all.[44]

I will conclude this section by considering a final issue, which may have been perturbing the reader. If it really is the case that the internal approach to interpretation puts the postulation of equivalences prior to the possible worlds, then what kinds of considerations are to be deployed in advocating one interpretation over another? That is, what makes something a *good* interpretation or not? If the possible worlds are somehow "there" prior to and independently of the process of interpretation, and if the models of the theory are just in the business of representing those worlds, then we could give a straightforward criterion for whether an interpretation is good or not: it's good just in case it judges two models to be equivalent exactly when they represent the same possible world. But if the possible worlds are (in some sense) constructions from an interpretation, then it looks as though all interpretations will be on a par. If I have an interpretation you dislike, then you cannot charge me with being mistaken about what the possible worlds are like. By definition, *my* possible worlds (i.e., those appropriate to the modality associated to my interpretation of the theory) are in line with my interpretation; just as your possible worlds are in line with your interpretation. So what can you say to persuade me out of my interpretation?

The answer is that you can say exactly the sorts of things you would normally say in criticising someone's interpretation—just without the detour via metaphysically robust possible worlds. For example, suppose that you think my interpretation is too fine-grained: it takes some models as inequivalent (i.e., to represent distinct

[44] 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?

possibilities), which you think should be taken as equivalent. Suppose further that you think this for essentially epistemic reasons: on my interpretation there are certain facts (those concerning which of the allegedly distinct possibilities is actual) that would be in principle inaccessible to knowers in those possibilities. That's still a good argument against my interpretation! For, what interpretation one plumps for affects what sentences will have determinate truth-values (in worlds governed by the theory), and hence what kinds of arguments one thinks are worth having about the theory. If you're right in your epistemic argument, then I'm committed to there being certain kinds of arguments that are worth having, but which cannot (even in principle) be settled by appeal to empirical evidence. That's a problem, though not an insurmountable one. Perhaps the kinds of explanation that can be given in my interpretation are better, or perhaps the ontology associated with it is somehow better (e.g. it abides by a principle of local action).

Whatever the details, the point for our purposes is just that this kind of familiar back-and-forth is not, so far as I can tell, improved by holding that we are arguing about the nature of antecedently existing possible worlds. Indeed, doing so would seem to merely add to the mystery. Why think that these worlds are never epistemically distinguishable? Or that their ontologies are especially intelligible? It's reasonably easy to think of pragmatic virtues for interpretations which are epistemically or explanatorily well-behaved, or which involve more readily intelligible ontologies. But that suggests that some more deflationary account of possible worlds fits *better* with making sense of disagreements over the best interpretation. It opens up the space for pragmatic virtues to be decisive in anointing one interpretation as "best", without being crowded out by the simple virtue of being right or wrong.

5 Internal inter-theoretic relations

So, this is how the internal approach characterises the project of interpreting a given theory: as one of elaborating its internal networks of synonymy and equivalence. However, it should be clear that this process (at least, taken naively) cannot be all there is to interpretation. After all, there are plenty of cases where we have theories whose internal structures are *identical*, and yet which—as we say—ought to receive different interpretations. To take a well-worn example, the mathematics of a simple harmonic oscillator may be used to represent small pendulums, or masses on springs, or vibrating strings, or individual modes of electromagnetic radiation, or inductive

electric circuits, or many others besides. So in comparing (say) the theory of a mass on a spring and that of an inductive electric circuit, paying attention only to their internal structure would lead one to the conclusion that they are equivalent theories; only by attending to the relationships those theories bear to the world can we recognise the representational difference between them.

Well, so you might think. However, it seems to me that there is a way in which the internalist can make sense of the distinctions between these theories, without making primitive use of notions like representation or reference. (In the next section, I'll discuss in more detail where such notions could come in.) The solution is to think of inter-theoretic relations in a particular way: namely, as intra-theoretic relations. To see this, it is helpful to focus upon the role that such judgments play in our scientific practice.

Let's take a (super-simple) example. Suppose that you and I both write down Maxwell's equations—but whereas I use ρ to indicate charge density, you use μ. It seems clear that we should judge the two theories to be equivalent. What is involved in doing so? Simply that in speaking the *combined* vocabulary (that involving both ρ and μ), certain kinds of inferences are licenced: for example, from

> In this region, ρ vanishes.

we may infer

> In this region, μ vanishes.

And of course, this generalises: any statement about ρ may equally well be phrased as a statement about μ, since these are just two different notations for the same thing. In other words, if M_μ is your version of Maxwell's equations and M_ρ is mine, then the proper combined theory is *not* just $M_\mu \cup M_\rho$: rather, it is $M_\mu \cup M_\rho \cup \{\mu = \rho\}$. Thus, it is in the act of integrating two theories into a single theory that we can make use of a judgment of equivalence.

Here, the formal relationship between the two theories was very clearly apt for underwriting a judgment of equivalence. More generally, we might take the view that if two theories are related by a systematic translation, then they may be taken as equivalent: perhaps a translation that permits us to construct new predicates and functions,[45] or which permits us to construct new sorts and quantifiers,[46] or

[45] i.e. definitional equivalence: see Glymour (1970), Barrett and Halvorson (2015a)
[46] i.e. so-called *Morita equivalence*: see Barrett and Halvorson (2015b).

which is a translation in some more general and abstract sense yet.[47] This isn't to claim that any of these views on permissible varieties of translation are, or should be, uncontroversial; rather, it is to claim that such controversies are important and vital, precisely because they are a precondition to interpretation. Indeed, I think many debates about the "richness" of the ontology we attribute to the world may be perspicuously recast as debates about what criteria of translation are appropriate. For example, fans of grounding or fundamentality may want to resist the idea that definitional equivalence gives a good notion of translation: which terms are primitive and which are defined, they could insist, encodes differing commitments about which properties are fundamental and which are derivative.[48] A larger audience will want to resist the claim that Morita equivalence is a species of translation: that opens the way, for instance, for mereological nihilism and universalism to collapse into one another. So the acceptability of such criteria of equivalence is (*pace* Coffey) not a mere corollary of the interpretive project, but rather an integral part of it. In particular, it seems to me that the clearest way to be an ontological deflationist is to provide such criteria, and defend the claim that they can support judgments of equivalence when theories are combined.

Of course, to say that we *can* employ a judgment of equivalence in combining theories does not mean that we *must* do so. And this, I claim, is precisely what happens with the case of the simple harmonic oscillators. We have several theories which are perfectly apt for equivalence—the equations in each case are the same in form, differing only (let us suppose) in their choice of variables.[49] The difference, then, lies in how those theories are combined: if x is the position of the mass on the spring, and I the current through the circuit, then in treating of both at once we certainly *cannot* infer that $I = 2$ from $x = 2$—notwithstanding the fact that x and I play exactly analogous roles in the two sets of equations.

Equivalence between theories is not the only inter-theoretic relation that is relevant

[47] I think of the research on categorical equivalence of theories (Weatherall (2015), Rosenstock et al. (2015), Rosenstock and Weatherall (2016)) as exploring what the most general and abstract constraints on a notion of translation might be.

[48] See e.g. Maudlin (2007b)'s claim that one can have two versions of electromagnetism: one in which charge density is primitive, and correlated by the laws with the divergence of the electric field; and one in which charge density is *defined* as the divergence of $\mathbf{E}$. Hicks and Schaffer (2015) also discuss the relationship between definability and non-fundamentality.

[49] Of course, in practice we often don't have distinct variables: one typically uses ω, for instance, to denote the angular frequency of whatever SHO system is under investigation, whether it be a mass on a spring, an inductive circuit, or whatever. (This also illustrates that the same goes for theoretical but natural-language terms such as "angular frequency".) But this is no more a cause for puzzlement than the existence of many bearers of the name "John" (or even "John Smith").

to the question of how theories ought to be combined. A thesis of reduction, for instance, seeks to show how the terms of one theory may be identified with (perhaps higher-level constructions from) those of another; whilst showing that one theory is a limiting case of another is a matter of showing how the terms of one theory may, upon application of the appropriate limit, be identified with those of another. Again, the virtue of doing so is that it permits cross-theoretical inferences of a certain kind. From a series of statistical-mechanical claims, I can (given a reduction of thermodynamics to statistical mechanics) infer certain thermodynamical claims; once I see the sense in which Newtonian gravitation is a limiting case of General Relativity, I can import or export data between (say) a general-relativistic model of the solar system, appropriate for determining motions near the sun, and a more computationally tractable Newtonian model of the solar system, adequate for the motions further out. There are more subtle relations as well: for example, one might use theoretical quantum chemistry to predict reaction rates, which can then be fed into one's chemistry theory as parameters.

In the above examples, it is (we now think) obvious whether the terms of the one theory should be identified with those of the other. But whether the terms of one theory may be identified with those of another, and the concomitant issue of how the two theories ought to be integrated with one another, is often a matter for substantive scientific investigation. Think of Maxwell's proposal that the notion of light (as that term occurs in the theory of optics) ought to be identified with propagating electromagnetic waves (understood via Maxwell's own theory of the electromagnetic field). Or, for a more contemporary example, consider black-hole thermodynamics. Supposing one accepts that theory, then one will agree that there are quantities in the relevant equations that play analogous roles to certain quantities in the equations of "traditional" thermodynamics—but it is a further question whether the horizon area (the analogue of entropy) ought to be identified with thermodynamical entropy when the two theories are combined.

6 Theory meets world

I turn now to one final concern: just how, on this account, do theories come to have empirical, physical content? For, one might feel, no matter how much careful explication is done of a theory's internal semantic architecture (and/or its relations to other theories) it will remain marooned—cut off from contact with the world—unless

we can provide it with appropriate referential links to that world.

The answer is to recognise a fiction in which we have been indulging for most of this essay. Namely, we have supposed that the aspiring philosopher of science finds themselves in the same boat as Quine's intrepid field linguist:[50] confronted by a wholly unknown representational practice, and faced by the daunting task of how that practice is to be made intelligible, how it is (as the objection puts it) to be endowed with physical or referential content. This picture is the final vestige of the externalist viewpoint, and underpins their supposition that the task of the philosopher of science is (like that of the linguist) to come up with an appropriate dictionary. So far, I've discussed what could be done by the field linguist without translating the theory into their home language—how (as it were) they could try to construct a grammar or lexicon, rather than a dictionary, for the target language. Now, though, we should drop the analogy altogether. We don't begin our analysis of scientific theories by taking some mysterious equations carved on stone tablets and puzzling out what they might mean: theories are born as bearing all kinds of semantic or interpretational relationships to our broader representational practice. So the interpretative task which confronts us is not that of the field linguist, but that of the lexicologist: the problem is not how to comprehend an alien practice, but how to fully understand a practice which we already—at least to some extent—inhabit.

In particular, we do not begin our analysis with a multiplicity of isolated theories and languages, but rather with a *single* theory. It is part of the task of science to work out how parts of that theory can be parcelled up and separated off, in order to better systematise the nature of our scientific knowledge. So austere calculi are not the starting-point for scientific inquiry, but rather a result of it. It was a substantial scientific achievement to get to the point where our understanding of the electromagnetic field was so well-encapsulated by a single set of equations that Hertz could identify Maxwell's theory with those equations. It is similarly part of the fruits of scientific knowledge to equip us with a rich language for describing the results of experiments, and to synthesise our empirical knowledge into what might be called an "empirical theory" (relative to a given domain of theoretical inquiry).[51] Having done both of these tasks, we are then in a position to show how the austere theory and the empirical theory can be successfully integrated, via the kinds of term-identifications that I discussed in the previous section—or, as we say, to show how the theory is sup-

[50] See (Quine, 1960, chap. 2).

[51] cf. Suppes' discussion of "the theory of the experiment" Suppes (1969), or Nagel's notion of an "experimental law" (Nagel, 1979, chap. 5).

ported by evidence. Of course, it may be that no such integration can be successfully performed. In that case, our only option is to reject or modify parts of our theoretical framework until such consilience can be achieved.[52]

Note that this amounts to a kind of recovery of the external picture within the internal approach: certain connections are postulated between the particular theory at hand and an appropriate empirical theory for it, at least so far as possible. However, there are (at least) two important differences. First, there is no *a priori* commitment to a particular empirical theory as that to which all other theories must be related, nor even to a particular form of language that the empirical theory must take. Second, the relationship between a scientific theory and its empirical theory is just one instance of a broader class of intertheoretic relations: there is not a principled difference (at least at the level of semantics, rather than epistemology) between the relation of thermodynamics to statistical mechanics, and the relation of theoretical chemistry to summaries of lab experiments. Relatedly, the relationships between scientific theories are not (necessarily) mediated by their respective relationships to some empirical theory. By contrast, we saw that on the external view, questions about the relations between a pair of scientific theories lose their autonomy; they just supervene on questions about the relations those theories bear to the privileged language.

Now, all this leaves an important question unanswered: just what relationship, if any, obtains between our unified (theoretical plus empirical) theory and "the world". However, it seems to me that addressing that question is not part of the purview of philosophy of science. For it is a question that arises even prior to our engaging in anything recognisably like science, with its distinctive problems and questions of interpretation. Any kind of representational practice will, as a matter of course, raise the question of how the representational apparatus relates to the entities represented. Moreover, it is a question whose answer seems, for the most part, to float free of anything tangibly related to the distinctive purposes of science. Once we have an account of how a theory relates to our empirical theory, we are in a position to use that theory to augment the empirical theory, and adjust our expectations of the future (especially those related to the effects of specific kinds of intervention in the world) appropriately. What more is added to our scientific practice by the assumption, or

[52]So, a theory being falsified is better described as our larger theory (the conjunction of the particular theory with the empirical theory, together with appropriate bridging claims) turning out to be inconsistent. This conception of truth in terms of consistency was defended by the early Reichenbach: see (Reichenbach, 1965, chap. IV).

the requirement, that the theory's terms "genuinely refer"?[53] To be clear, this is not to claim that the question of realism (for it is he!) is not of philosophical importance. It is just to deny that it is of *specific* importance to the philosophy of science, or that it is most appropriately handled by the methods of philosophy of science—rather than, say, metaphysics (in something more like the Kantian than the analytic sense, i.e. as the inquiry into the relationship between the noumenal and the phenomenal) or philosophy of language.

So what, then, are the problems of interpretation to which philosophy of science *is* best addressed? Well, at the general level, there are the kinds of projects that I have already canvassed in this essay. Are isomorphic models, or symmetry-related models, representationally equivalent? What kind of relationship must hold between the respective architectures of two theories, if they are to be meshed together via equivalence or reduction? And in the philosophies of the special sciences, there are also important questions of interpretation. Does general relativity admit of a coherent notion of gravitational energy: that is, is there anything in general relativity which may be identified with "energy" as it appears in other theories? What notion of "species" is best used by biologists—or are there (as seems plausible) different ways of unpacking such a notion, which are apt for different contexts? Such projects remain, and keep their importance, on the internal approach. It may even be of value to use the standard representationalist tools (reference, truth) to analyse such questions: e.g., to what in the theory of statistical mechanics might the term "temperature" refer? Under what conditions, as described by general relativity, is Newtonian gravitation approximately true? Only the detour through a single, privileged theory, or (what comes to the same thing) via the world, is omitted—and I cannot see that philosophy of science should regard that as a very great loss.

Acknowledgements

Many thanks to Thomas Barrett, Jeff Barrett, Erik Curiel, Josh Eisenthal, Alex Meehan, Oliver Pooley, Emma Ruttkamp-Bloem, David Schroeren, Kyle Stanford, Jim Weatherall and David Wallace for comments on and/or discussions of earlier drafts of this material; I'm also very grateful to audiences at The Semantics of Theories conference and the SoCal Philosophy of Physics Reading Group for their insightful questions.

[53] cf. Stein's discussion of "the fallacy of something more" Stein (1989).

References

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.

Baker, D. J. (2010). Symmetry and the Metaphysics of Physics. *Philosophy Compass*, 5(12):1157–1166.

Barrett, T. W. and Halvorson, H. (2015a). Glymour and Quine on Theoretical Equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

Barrett, T. W. and Halvorson, H. (2015b). Morita equivalence. *arXiv:1506.04675*.

Bhogal, H. and Perry, Z. (2015). What the Humean Should Say About Entanglement. *Noûs*.

Bricker, P. (2006). Absolute actuality and the plurality of worlds. *Philosophical perspectives*, 20(1):41–76.

Callender, C. (2014). One world, one beable. *Synthese*, pages 1–25.

Carnap, R. (1929). Von Gott und Seele. Unpublished lecture. Item number RC 089-63-02, Archives of Scientific Philosophy in the Twentieth Century, Department of Special Collections, University of Pittsburgh.

Carnap, R. (1956). *Meaning and Necessity: A Study in Semantics and Modal Logic*. University of Chicago Press, Chicago.

Carnap, R. (1967). *The Logical Structure of the World; Pseudoproblems in Philosophy.* University of California Press, Berkeley.

Caulton, A. (2015). The role of symmetry in the interpretation of physical theories. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52:153–162.

Chakravartty, A. (2007). *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge University Press, Cambridge, UK.

Coffa, J. A. (1991). *The Semantic Tradition from Kant to Carnap: To the Vienna Station*. Cambridge University Press, Cambridge.

Coffey, K. (2014). Theoretical Equivalence as Interpretative Equivalence. *The British Journal for the Philosophy of Science*, 65(4):821–844.

Craig, W. (1953). On axiomatizability within a system. *The journal of Symbolic logic*, 18(01):30–32.

Dasgupta, S. (2014). Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science*, Forthcoming.

Dewar, N. (2015). Symmetries and the philosophy of language. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52, Part B:317–327.

Dewar, N. (n.d.). Sophistication about symmetries. *The British Journal for the Philosophy of Science*, forthcoming.

Dürr, D. (2008). Bohmian Mechanics. In Bricmont, J., Dürr, D., Galavotti, M. C., Ghirardi, G., Petruccione, F., and Zanghi, N., editors, *Chance in Physics: Foundations and Perspectives*, pages 115–132. Springer.

Dürr, D., Goldstein, S., and Zanghí, N. (1992). Quantum equilibrium and the origin of absolute uncertainty. *Journal of Statistical Physics*, 67(5-6):843–907.

Dürr, D. and Teufel, S. (2009). *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer Science & Business Media.

Egg, M. and Esfeld, M. (2014). Primitive ontology and quantum state in the GRW matter density theory. *Synthese*, 192(10):3229–3245.

Esfeld, M. (2014). Quantum Humeanism,or: Physicalism without properties. *Philosophical Quarterly*.

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.

Everett, H. (1957). "Relative State" Formulation of Quantum Mechanics. *Reviews of Modern Physics*, 29(3):454–462.

French, S. (2013). Handling Humility: Towards a Metaphysically Informed Naturalism. In Galparsoro, J. I. and Cordero, A., editors, *Reflections on Naturalism*, pages 85–104. Springer Science & Business Media.

Ghirardi, G. (2016). Collapse Theories. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition.

Glymour, C. (1970). Theoretical Realism and Theoretical Equivalence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pages 275–288.

Halvorson, H. (2012). What Scientific Theories Could Not Be. *Philosophy of Science*, 79(2):183–206.

Halvorson, H. (2013). The Semantic View, If Plausible, Is Syntactic. *Philosophy of Science*, 80(3):475–478.

Hicks, M. T. and Schaffer, J. (2015). Derivative Properties in Fundamental Laws. *The British Journal for the Philosophy of Science*.

Jones, R. (1991). Realism about What? *Philosophy of Science*, 58(2):185–202.

Kaplan, D. (1975). How to Russell a Frege-Church. *The Journal of Philosophy*, 72(19):716–729.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell Publishers Ltd, Oxford, UK.

Lutz, S. (2015). What Was the Syntax-Semantics Debate in the Philosophy of Science About? *Philosophy and Phenomenological Research*, 91(3).

Maudlin, T. (2007a). A Modest Proposal Concerning Laws, Counterfactuals, and Explanations. In *The Metaphysics Within Physics*. Oxford University Press, Oxford.

Maudlin, T. W. E. (2007b). Completeness, supervenience and ontology. *Journal of Physics A: Mathematical and Theoretical*, 40(12):3151.

Miller, E. (2014). Quantum Entanglement, Bohmian Mechanics, and Humean Supervenience. *Australasian Journal of Philosophy*, 92(3):567–583.

Møller-Nielsen, T. (n.d.). Invariance, Interpretation, and Motivation. Unpublished draft.

Nagel, E. (1979). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Hackett, Indianapolis.

Nolan, D. (2015). *David Lewis*. Routledge.

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, UK.

Putnam, H. (1965). Craig's Theorem. *The Journal of Philosophy*, 62(10):251.

Putnam, H. (1983). Equivalence. In *Realism and Reason*, volume 3 of *Philosophical Papers*, pages 26–45. Cambridge University Press, Cambridge, UK.

Quine, W. V. O. (1960). *Word and Object*. M.I.T. Press, Cambridge.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press, Chicago, IL.

Reichenbach, H. (1965). *The Theory of Relativity and A Priori Knowledge*. University of California Press, Berkeley. English translation by M. Reichenbach of "Relativitätstheorie und Erkenntnis Apriori" (1920).

Roberts, J. T. (2008). A Puzzle about Laws, Symmetries and Measurability. *The British Journal for the Philosophy of Science*, 59(2):143–168.

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.

Rosenstock, S. and Weatherall, J. O. (2016). A categorical equivalence between generalized holonomy maps on a connected manifold and principal connections on bundles over that manifold. *Journal of Mathematical Physics*, 57(10):102902.

Ruetsche, L. (2011). *Interpreting Quantum Theories: The Art of the Possible*. Oxford University Press, Oxford; New York.

Russell, B. (1993). *Our Knowledge of the External World*. Routledge, London.

Saunders, S. (2003). Indiscernibles, general covariance, and other symmetries: The case for non-eliminativist relationalism. In Ashtekar, A., Howard, D., Renn, J., Sarkar, S., and Shimony, A., editors, *Revisiting the Foundations of Relativistic Physics: Festschrift in Honour of John Stachel*. Kluwer, Dordrecht.

Sider, T. (1996). All the World's a Stage. *Australasian Journal of Philosophy*, 74(3):433–453.

Stanford, K. (n.d.). Reading Nature: The Interpretation of Scientific Theories. In Sklar, L., editor, *The Oxford Handbook of the Philosophy of Science*. Oxford University Press, Oxford.

Stein, H. (1989). Yes, but... Some Skeptical Remarks on Realism and Anti-Realism. *Dialectica*, 43(12):47–65.

Suppes, P. (1969). Models of data. In *Studies in the Methodology and Foundations of Science*, pages 24–35. Springer.

The Univalent Foundations Program (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. http://homotopytypetheory.org/book, Institute for Advanced Study.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford University Press, Oxford; New York.

Weatherall, J. O. (2015). Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent? *Erkenntnis*, 81(5):1073–1091.

Weatherall, J. O. (2016). Regarding the 'Hole Argument'. *The British Journal for the Philosophy of Science*, page axw012.

Williamson, T. (2013). *Modal Logic as Metaphysics*. OUP Oxford.