What is it to interpret a theory?

May 31, 2016

Abstract

This paper seeks to give an account of what could be involved in interpreting a theory. The aim is to try and provide a robust conception of theory-interpretation which operates in terms internal to the representational architecture of the theory, rather than importing meaning by stipulative correspondence to external terms.

1 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 2) by assembling various platitudes about what interpretation is for, so that we can get a bead on the notion we are after. Section 3 outlines the external account of interpretation in more detail, looking in particular at two examples of the form: the reductionist story about interpretation found in the *Aufbau*, and the motivating ideas behind the quest for primitive ontology in quantum mechanics. I take a slight step back in section 4, 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 5.

2 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 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 the

[1] cf. [Jones, 1991].

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).[5]

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 that these are virtues of *interpreted* theories.[6] So not only is interpretation

[2] cf. [Stanford, MS].

[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.

[5] cf. what Steven French (in [French, 2014, chap. 3]) calls "Chakravartty's Challenge": the claim that "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." [Chakravartty, 2007, p. 26].

[6] I take this observation from [Ruetsche, 2011, Chapter 1].

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.

3 Against external interpretation

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 the "external" approach mentioned above, which takes the interpretation of a theory 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.

One famous example of this kind of project are attempts to reduce scientific (and ordinary) discourse to some privileged "phenomenological" language. As a starting-point, consider 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 "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.[7]

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."[8]

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 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

[7] [Russell, 1993, pp. 77–78]
[8] [Russell, 1993, pp. 88–89]

> 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.[9]

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 the broader positivist program (of which Carnap's work was a part) exemplifies the connections we canvassed in section 2 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).[10] 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.[11]

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. 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

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

[10] 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).

[11] See e.g. [Reichenbach, 1938], or [Putnam, 1983] for a critique.

experience, are (for whatever reason) nowhere near as widespread as they once were.

But this is not to say that interpretation by translation died with the positivists. Interpreting a theory by reducing it to a *phenomenological* basis may no longer be popular, but (I claim) there is still a widespread idea that interpretation is, in the first instance, a matter of cashing the theory out in *some* privileged basis.[12] For instance, 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.[13] Nevertheless, there is an important continuity. Maudlin's account of the relationship between the two approaches is exemplary, and worth quoting at length:

> 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 one 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.

[12] It may be worth noting 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]

[13] 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.

> 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.[14]

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 §5 below.

Second, however, there is the claim that this correspondence is most easily done 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". But what this means is that

[14][Maudlin, 2007b, p. 3158–3159]

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 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?[15]

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";[16] "ignoring [the primitive ontology of particle positions in Bohmian mechanics], the theory becomes a theory about nothing";[17] "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."[18] Thus, interpreting a theory is a matter of identifying the primitive ontology of the theory (or providing it

[15][Dürr et al., 1992, pp.848–849]
[16][Ghirardi, 2016]
[17][Dürr, 2008, p. 117]; the context makes it reasonably clear that the claim generalises to other forms of primitive ontology.
[18][Dürr and Teufel, 2009, p. 38]

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.

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 commitments.[20]

So we've now seen two examples of the external approach to interpretation. Other examples could be adduced: in particular, certain questions in the metaphysics of science (e.g. are physical properties dispositions?) could plausibly be characterised as bids to "translate" our theories of physics into metaphysicalese. But for now, we have enough examples to make clear the overall character of the external approach—and my concerns about it.

There are two problems in particular with an external approach to interpretation. 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.[21] We should be extremely skeptical that the reflections of philosophers will offer a better mechanism for theory choice in science than the practice of science does.

Secondly, there is a concern about the prima facie coherence of the project. This view treats meaning as a kind of special resource, imported into our theories by establishing appropriate pipelines to other, already meaningful theories. But the source has to be somewhere. So at some point, the question will need to be addressed of why the transparent theory comes with meaning pre-equipped, in contrast to other theories. What is it about our starting-point that lets it be spangled in content; and

[19] [Allori et al., 2008, p. 365]—although as discussed in n. 31 below, they also seem open to applying the converse direction.

[20] See e.g. [Miller, 2014], [Callender, 2014], [Esfeld, 2014], or [Bhogal and Perry, 2015].

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

what is it about all other places that they can only be rendered contentual by being hooked up to this primary source, rather than being made significant directly and on their own terms? In the end, after all, we hope for a theory which holds all there is within its compass. Interpreting such a theory cannot draw upon resources external to it, by definition; it must be the case that "the theory itself sets the framework for its interpretation", as Everett famously claimed for his interpretation of quantum mechanics.[22] So if nothing else, there is a value to be had in looking for an alternative to external interpretation, against the day when we are called upon to interpret such a theory.

4 What are we interpreting?

Towards developing an alternative picture of theory-interpretation, we should first of all pause to spell out in more detail just what it is we take ourselves to be interpreting. That is: what, for our purposes, is 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 comprised by 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 isolation from the sentences, will fail to capture everything of interest.[23]

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:

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

[22] [Everett, 1957]

[23] In this, I follow [Halvorson, 2012], [Halvorson, 2013], and [Lutz, 2015].

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 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 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":[24] 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

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

thoughtful version of either view.[25] I merely wish to signal that it does not, so far as I can see, coincide with thoughtless versions of either.

The theory above was formulated in terms of differential equations, the workhorses of modern physics. 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 in the form of differential equations, governing how systems evolve over time (or how fields may be distributed over spacetime).[26] 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 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*.[27] 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,

[25] 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.

[26] cf. [Maudlin, 2007a].

[27] 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).

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).[28] 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.

5 Internal interpretation

As we have now seen, 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. To interpret a theory is to indicate what parts, or components, or aspects of the putative semantics are to be taken seriously. How is that to be done, other than putting into correspondence with some already-serious theory?

Towards an answer, let's return to one of the issues that came up when discussing external interpretation. We saw there that one of the characteristic features of these interpretations was that they tended to induce a criterion of equivalence: for two theories (say) to be equivalent was just for them to correspond to the same theory in the external language. [Coffey, 2014] has argued that this demonstrates, more 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

[28] 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.

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 that can easily accommodate the preceding puzzles:

> 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.[29]

As discussed in §2 above, I am inclined to think that "taking the interpretive project seriously" is not just the province of realists, but that is not my main concern here. Rather, it is that this conclusion only follows if one understands the interpretive project along the lines of the external account: i.e., as Coffey says, as the project of articulating the "physical pictures" associated with a pair of formalisms. Now that we've seen some of the problems in how such picturing could take place,[30] Coffey's account suggests a natural place to get off the boat: reverse the order of dependence between commitment and equivalence. When interpreting by translation, one articulates the theory's commitments in the privileged language, and uses that as a means of determining equivalence. As Coffey notes, and as we saw above, two theories or models are equivalent just in case they say the same things in the privileged language. On the alternative conception—which we shall call the *internal* approach to interpretation—we begin by making determinations of equivalence (between models within one theory, or between the models of one theory and those of another), and use those determinations to get a fix on the theory's commitments. We do this by employing the following

[29] [Coffey, 2014, pp. 834–835]

[30] To be fair, Coffey acknowledges that there are difficulties in unpacking the notion of external interpretation: his claim is merely that "if we are already committed to the coherence of the interpretive project in foundational physics, [...] then we are already presupposing the necessary semantic knowledge. On this approach to theoretical equivalence, we incur no new semantic debts not already incurred in virtue of our commitment to the interpretive project itself." [Coffey, 2014, p. 835]. But this does not hold if, as I suggest here, judgments of equivalence are what undergird the interpretive project.

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.[31]

Thus, on the internal view, interpreting a theory is a matter of postulating certain equivalences between elements of the putative semantics. In effect, we abstract away from the differences between the (declared-to-be) equivalent models. The results of this process of abstraction—that is, the things we obtain by abstracting from the models in this way—are naturally understood as possible worlds: "possible", that is, in the sense of being nomologically possible relative to taking the claims of the theory 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 spontaneously accelerate? 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.[32]

For instance, at least if we are using standard mathematical tools,[33] models can be distinct whilst still being isomorphic: perhaps one model has a domain comprising the

[31] It should be noted that [Allori et al., 2008] are sympathetic to such an idea. The quotation given in §3 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]

[32] 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].)

[33] Rather than, say, homotopy type theory (see [The Univalent Foundations Program, 2013]).

natural numbers as its domain, whereas its isomorphic cousin has the integers. But perhaps we should be sceptical that this distinction corresponds to any difference, if we want to deny that the particular identities of the objects used to populate the models are playing any substantive representational role.[34] If so, we should interpret isomorphic models as equivalent, and deny that the differences between them are to be taken seriously.[35]

Or, more generally, one might maintain that if a theory contains symmetries (or at least, symmetries of a certain kind), then the theory should be interpreted as committed only to those components of its ontology which are invariant under those symmetries.[36] Again, the internal approach cashes out this lesson as a matter of asserting equivalence between symmetry-related models of the theory. Two such models represent the same possibility; but this is as much a clarification of what we mean by "possibility" as anything else.

Turning to relationships between theories (rather than within theories): as we have noted above, models (at least as I am understanding them) are language-soaked entities: they wear their syntactic labels on their sleeves, as it were. (If we want to be more formal about this, we can observe that the relevant notion of isomorphism for models of a theory is one which preserves those labels: an isomorphism from a model $\mathcal{M}$ to another model $\mathcal{N}$ is a bijection $f : |\mathcal{M}| \to |\mathcal{N}|$ such that (say) the extension $F^{\mathcal{M}}$ is mapped to $F^{\mathcal{N}}$, $R^{\mathcal{M}}$ is mapped to $R^{\mathcal{N}}$, etc.) But we don't think that these labels should be taken seriously. So we interpret the theory in such a way that they are not—*which is just to say* that we regard two theories as equivalent if they agree up to a choice of labels. Thus, for example, the theory

$$\forall x(Fx \to Gx) \tag{3}$$

is equivalent to

$$\forall x(Px \to Qx) \tag{4}$$

And this criterion of equivalence gets cashed out in the observation that a model $\mathcal{A}$ of the former theory should be regarded as equivalent to a model $\mathcal{B}$ of the latter theory if they are related by such a swapping of labels: if, that is, the two models are related

[34] cf. [Kaplan, 1975].

[35] As e.g. [Pooley, 2006] and [Weatherall, MS] advocate as a solution to the Hole Argument.

[36] There is a large literature on the merits or demerits of this as an interpretative stance: see e.g. [Saunders, 2003], [Roberts, 2008], [Baker, 2010], [Dasgupta, 2014], [Caulton, 2015], or [Dewar, 2015].

by

$$\begin{aligned} |\mathcal{A}| &= |\mathcal{B}| \\ F^{\mathcal{A}} &= P^{\mathcal{A}} \\ G^{\mathcal{A}} &= Q^{\mathcal{B}} \end{aligned} \tag{5}$$

More generally, we might take the view that if two theories are related by a systematic translation, then they are equivalent. This could be made precise in terms of definitional or translational equivalence.[37] More generally still, we could allow that even if the two theories disagree on how many things there are, they are still equivalent if the extra objects in one are all appropriately "constructible" from the objects of the other, and vice versa. This could be made precise in terms of so-called "Morita equivalence".[38] In either case, a notable feature of these accounts of equivalence is that if two theories can be shown to be equivalent (in either sense), there is a corresponding equivalence between the classes of models of the two theories.

Now, this is not to say that all of the above interpretative moves can be defended. For example, fans of grounding or fundamentality may want to resist the idea that translationally equivalent theories should be regarded as equivalent: which terms are primitive and which are defined, they could insist, encodes differing commitments about which properties are fundamental and which are derivative.[39] A larger audience will want to resist the idea that Morita-equivalent theories are equivalent: that opens the way, for instance, for mereological nihilism and universalism to collapse into one another. My point here is not to defend this or that specific interpretative move, but just to observe that each such move can be characterised as the presentation and advocacy of some criterion of equivalence or other.

Thus far, I have concentrated on how an internal interpretation affects our understanding of the relationships between the semantic structures of the theories (i.e., their classes of models). However, doing so also has profound implications for our understanding of the relationships amongst and between the syntactic structures (i.e., the sentences). Suppose that we are dealing with a single theory, and we decide to interpret the theory in such a way that some non-isomorphic models of the theory are equivalent. In general, this will mean that some of the sentences (in the theory's lan-

[37] See [Glymour, 1970], [Barrett and Halvorson, 2015a].

[38] See [Barrett and Halvorson, 2015b].

[39] 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.

guage) which are true of one model are not true of the other.[40] Suppose, then, that σ is such a sentence. Then by interpreting the theory in this fashion, we commit ourselves to thinking that σ is somehow defective, since its truth-value varies between equivalent models. Or, more precisely, it shows that σ contains some element of conventionality. In practical terms, this means that if two parties disagree over σ, that need not mean that there is anything genuinely at issue. For instance, suppose that we interpret electromagnetism in such a way that models related by a gauge symmetry are equivalent (as is standard scientific practice). We are then committed to thinking that gauge-dependent sentences are defective in this fashion: if you and I disagree over a sentence such as "the electrical potential is everywhere vanishing", that does not show that we disagree in any substantive fashion. In order to crystallise a disagreement, we would have had to agree on a gauge convention. Then, and only then, would our disagreement over gauge-variant sentences be worth arguing over.

Thus, interpreting a theory can mean identifying certain portions of the theoretical language as privileged: namely, those elements of it which are appropriately invariant under the equivalence-relations that have been identified. This leads to a certain commonality with the external approach. For example, one *could* be a primitive ontologist on the internal approach: the criterion for equivalence would be that the models agree on what the chosen local beables of the theory are up to. It would follow from this that statements about the behaviour of those theoretical entities *other* than the local beables are problematic in the same way that statements about the electromagnetic potential are. So it risks sounding like we are back at the same overall picture that I described (and criticised) in §3: a privileged subset of the theoretical vocabulary describes the true commitments of the theory, so that models are equivalent just in case their reducts to that vocabulary agree.

This impression can become even stronger if we think about how theories make contact with evidence. Earlier, I made approving noises about the idea that the empirical content of a theory should be sought in the picture—if any—which the theory provides of the motions of macroscopic bodies in space. Moreover, we might worry that it just seems *true* that—as a matter of scientific practice—these stipulations are what it is to interpret a theory. Sure, a critic might contend, there is *some* role for the sort of semantic clarification discussed above; but it's surely obvious that such

[40] It is not guaranteed, of course. In first-order theories, models may be non-isomorphic yet elementarily equivalent (i.e., such as to make all the same sentences true); and in theories formulated in terms of differential geometry, two models might be locally identical (thereby satisfying precisely the same local differential equations), whilst differing in their global topological character.

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 connection of the theoretical architecture to our overall picture of the world, plays an enormously significant role in endowing the theory with semantic content. Can these observations be incorporated into the internal account, without having it collapse into the external account?

Let us consider the second issue first. What role does specifying the "meaning" of certain variables play, according to internal interpretation? Simply this: they are prescriptions for what identities (if any) should be postulated in the event that two theories are conjoined. For example, if we have specified that the term ρ in one theory represents "charge density" (say), and that the term μ in another theory does the same, then that serves to commit ourselves to adding a condition $\mu = \rho$ in the event we conjoin the two theories. An important special case of this is when we seek to conjoin some theory with what might be called our *empirical* theory, which summarises our empirical or experimental data.[41] To use an example we have seen before, we might specify that the pointlike variables in one theory represent "atoms", and furthermore that rocks are made up of atoms. This commits us to thinking that the trajectories of rocks should coincide with the trajectories of collectives of the pointlike variables: in other words, that the motions of rocks will be predicted by the theory.[42]

Hence, the claims which the external approach describes as "specifications of meaning" are understood as a way of getting a larger and richer theory, rather than as a form of interpretation *per se*. This is the major difference between the internal and external approaches to interpretation. As a result, there is a big difference in how claims such as "$\mathbf{x}$ represents the position of the particle" bear upon the interpretation of the theory, and (correlatively) the kind of authority they are taken to enjoy. For the advocate of external interpretation, these claims possess a particular kind of semantic authority: they are *definitions*, and so bind the term to express a certain kind of thing (i.e., to express the same thing as its translation in the antecedently meaningful transparent language). On the internal approach, however, these claims differ only in degree from assertions such as (1), not in kind.

It follows that according to the internal approach, such claims are not compulsory

[41] cf. Nagel's notion of an "experimental law" [Nagel, 1979, chap. 5]

[42] 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].

for the business of interpretation. Sure, *one* way of fixing the relationships amongst the models of the theory (or between a pair of theories) is to translate that theory (or both theories) into a common tongue, whose interpretation is asserted to be literal—that is, to be such that different models always represent different possibilities. But there is no compulsion to do so—nor will it always be the case that all equivalence-relations can be represented in terms of sharing a common vocabulary. For instance, it is not clear how to make sense of this idea for translating between two theories in different languages. Even within a single theory, it will not always be the case that the invariant fragment of the theory's language itself constitutes a well-formed language: in general, it need not have the recursive structure of sentences generated by a compositional syntax. (Specifically: there are, in general, complex invariant statements which cannot be generated by assembly from simpler invariant statements; they can only be generated by assembly from simple *variant* statements.)[43] And even when the equivalence-relations do correspond to a common tongue in this way, there is certainly no requirement that the common tongue must take a particular form (e.g. that it must speak in purely phenomenological terms, or that it must talk only of local beables). This is important if we desire an account of interpretation which respects the naturalistic constraints canvassed at the end of §3.

On the internal picture, then, we always wind up with a theory, of one sort or another; there is no way of stepping outside all networks of representation altogether, and standing cheek by jowl to the world. Or, turning it around, we *do* stand cheek by jowl with the world when we represent it, in one form or another, and we should not defer the task of representation to an impossible standard whereby scientific representation is somehow not good enough. In other words, what the external approach thought of as the life and soul of interpretation—the fusion of one theory with another—turns out to be merely a prelude to interpretation proper. Hence, a full understanding of how such external explication operates must, perforce, depend on a full understanding of the mechanics of internal explication.

6 Conclusion

I will conclude 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

[43] It is for this reason that it is highly non-trivial to "reduce" a theory with symmetries to a theory that traffics only in quantities invariant under the symmetry.

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 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, as discussed in the previous section, 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 principal 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.

References

[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.

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

[Barrett and Halvorson, 2015a] Barrett, T. W. and Halvorson, H. (2015a). Glymour and Quine on Theoretical Equivalence. Unpublished draft.

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

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

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

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

[Carnap, 1929] 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, 1967] Carnap, R. (1967). *The logical structure of the world; Pseudoproblems in philosophy.* University of California Press, Berkeley.

[Caulton, 2015] 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, Part B:153–162.

[Chakravartty, 2007] Chakravartty, A. (2007). *A metaphysics for scientific realism: knowing the unobservable*. Cambridge University Press, Cambridge; New York.

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

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

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

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

[Dewar, 2015] 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.

[Dürr, 2008] 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 et al., 1992] 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 and Teufel, 2009] Dürr, D. and Teufel, S. (2009). *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer Science & Business Media.

[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.

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

[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.

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

[French, 2014] French, S. (2014). *The structure of the world: metaphysics and representation.* Oxford University Press, Oxford.

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

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

[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.

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

[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.

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

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

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

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

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

[Nagel, 1979] Nagel, E. (1979). *The structure of science: problems in the logic of scientific explanation*. Hackett, Indianapolis.

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

[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.

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

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

[Reichenbach, 1938] Reichenbach, H. (1938). *Experience and prediction: an analysis of the foundations and the structure of knowledge*. University of Chicago Press, Chicago.

[Reichenbach, 1965] 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, 2008] Roberts, J. T. (2008). A Puzzle about Laws, Symmetries and Measurability. *The British Journal for the Philosophy of Science*, 59(2):143–168.

[Ruetsche, 2011] Ruetsche, L. (2011). *Interpreting quantum theories: the art of the possible*. Oxford University Press, Oxford; New York.

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

[Saunders, 2003] 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, 1996] Sider, T. (1996). All the World's a Stage. *Australasian Journal of Philosophy*, 74(3):433–453.

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

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

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

[Weatherall, MS] Weatherall, J. O. (MS). Regarding the "Hole Argument". Forthcoming in the British Journal for the Philosophy of Science; available at arXiv:1412.0303 [physics.hist-ph].