Learning to Represent: Mathematics-first accounts of representation and their relation to natural language

David Wallace*

March 23, 2024

Abstract

I develop an account of how mathematized theories in physics represent physical systems, in response to the frequent claim that any such account must presuppose a non-mathematized, and usually linguistic, description of the system represented. The account I develop contains a circularity, in that representation is a mathematical relation between the models of a theory and the system as represented by some other model — but I argue that this circularity is not vicious, in any case refers in linguistic accounts of meaning and representation, and is simply a consequence of the fact that we have no unmediated, representation-independent access to the world.

1 Introduction

What is a scientific theory? On the traditional ('received') view of theories, a scientific theory is a collection of statements in a (natural or formalized) language: it tells us what things there are in the world according to that theory, what features they have, and what they do. If this were so, the relation between a scientific theory and the world would be a special case of the familiar and much-studied relation between a language and the world: scientific terms refer, or fail to refer, to entities in the world; scientific predicates refer, or fail to refer, to properties and relations of those entities; scientific statements are true, or else they are not true.

But at least in physics, it is increasingly recognized that scientific theories are not like this. Theories in physics are heavily mathematized; what is described in a physics textbook or monograph is in general not some piece of the physical world but some mathematical model or collection of models. Physicists use language to describe the world, for sure, including in textbooks and monographs,

*Department of History and Philosophy of Science, and Department of Philosophy, University of Pittsburgh; david.wallace@pitt.edu

but faced with requests for increasing detail or precision, the words eventually give way to the mathematics. Ask a condensed matter theorist to explain Bose-Einstein condensation, or a quantum field theorist to explain asymptotic freedom, and often you will initially get an answer in words, but push sufficiently hard on that answer and sooner or later they will write an equation on the blackboard and explain that *really* what they mean is *this*. On this 'mathematics-first' view of theories (usually, if awkwardly, called the 'semantic' view), a physical theory itself is a (possibly structured) collection of mathematical models, not a collection of claims in language, and the relation between a physical theory and the world is not a special case of linguistic representation, but another form of representation altogether.

If this view of physics[1] is correct, it becomes urgent to ask: what *is* this mathematical representation, this relation that a model bears to a physical system iff it represents it correctly? (The question remains important, if less central, even if *theories* are not to be thought of in math-first terms: almost any observer of physics accepts that mathematized *models* are ubiquitous in physics, even if theories are more than just structured collections of those models.) A natural thought — all but universal in the literature on scientific representation — is that any such answer must have this general form: the mathematical model has features ABC; the physical system being represented has features XYZ; such-and-such relation holds between ABC and XYZ. Perhaps more is needed — perhaps it is necessary, for instance, that a scientist *intends* to represent XYZ by ABC — but at the least, there is a describable relationship between representing and represented system.

But crucially, this means that we can give no account of how the model represents the system unless we already possess a means of describing the system, independent of the model. For prior to understanding that ABC represents XYZ, we must already grasp that the system has XYZ. And this seems to suggest a circularity in any attempt to explain mathematical representation — unless that attempt is derivative on a *non*-mathematicized description of the system, normally taken to be a description in words, an account of what the objects are that comprise the system and what properties they have. Mathematical representation must ultimately be derivative on linguistic representation, and so-called 'math-first' accounts of scientific theorizing are not so math-first as all that.

If so, the consequences are substantial. For one thing, if understanding how a theory mathematically represents a system requires an antecedent linguistic description of that system, then it is opaque how we could ever acquire that description. If we are considering examples close to the manifest image it is at least plausible that we might already *have* a non-mathematical description of a system prior to representing it mathematically, things are otherwise in the

[1] Explications and defenses of the view have been given by, e.g., Saunders (2003, 2016), French (2014), French and Saatsi (2006), Ladyman (1998), French and Ladyman (2003), Ladyman and Ross (2007), and McKenzie (2024); for my own preferred version, see (Wallace 2022b). It is normally called 'ontic structural realism', though that term is also used to refer to quite different views (see, e. g. , (Esfeld, Deckert, and Oldofredi 2017), (McKenzie 2014)).

more abstruse areas of modern physics. We seem to have no way of describing, say, the quark-gluon plasma, or the inflaton field, without using the heavily mathematized formalism of modern theoretical physics. Certainly we in fact had no such description prior to developing that formalism.

One can perhaps imagine some kind of bootstrapping approach. We have a verbal description of the manifest image, and so insofar as some parts of a mathematized theory represent the manifest image, we can understand how *that* representation works. The rest of the mathematized theory is initially just an operational calculus, but we can hypothesize various linguistic descriptions of the unobservable features of a system under the constraint that the description underpins some representational account of the mathematics. Most likely the linguistic description will be highly underdetermined, but perhaps we can resolve that underdetermination through the traditional extra-empirical virtues: simplicity, elegance, familiarity, and the like.

This brings us to a second consequence. If this search for a linguistic description is what is is to understand a physical theory — to interpret a formalism, if you like — then physicists by and large are not in the business of understanding their own theories, since this search for a purely linguistic description of a system that can then be related to its mathematical description plays no apparent role in physics practice. (The linguistic descriptions physicists in fact give are too heuristic, and too intertwined with mathematics, to count.) Interpreting a theory becomes a task separate from physics practice, a task for someone other than a physicist: a philosopher, perhaps?

And indeed, this conception of interpretation of theories as something that philosophers of physics do but physicists do not is very common in the philosophy literature. Sometimes it is described as a beneficial division of labor (in an introduction to philosophy of physics, Dean Rickles (2016) writes:"If the primary task of the physicist is to construct models and theories of the world, the primary task of a philosopher of physics is to interpret these products of physics"), sometimes as a lamentable feature of physics that philosophers should endeavor to correct (Tim Maudlin (2018a, p.1) writes: "Metaphysicians rightly look to physics for insight into the nature of the physical world. And once upon a time, they would get clear and articulate answers... Nowadays nothing is clear and sharp in the area at all.") Closely related is the idea that physicsts seek to *calculate* while philosophers seek to *understand*, and that these are orthogonal goals best served by quite different ways of conceptualizing a physical theory (see, e. g. , Callender's (1999, p.349) extended critique of Gibbs statistical mechanics while conceding that it is to be preferred for the practice of science, or Kuhlmann's (2010) parallel take on Lagrangian quantum field theory). Physicists, needless to say, would strongly contest the idea that they do not understand their own theories and need to outsource that problem to philosophers.[2]

But both consequences — the mystery of how we can get a description of

[2]They would also contest — or scathingly dismiss — the idea that being able to calculate in a theory is orthogonal to understanding it, and rightly so, but that's another story (to which I return briefly in section 8).

physical systems that does not rely on the mathematized theories we use to describe those system, and the idea that physicists are not in the business of interpreting their own theories — follow only if mathematical representation is ultimately dependent on linguistic description. My goal in this paper is to show that this is not so, or at least need not be so, by formulating an account of mathematical representation freed of that dependence.

The short version of the argument is: yes, there is a circularity in accounting for mathematical representation without some prior, non-mathematical way to represent; but the circularity is not vicious, and in any case essentially recurs in trying to explain how *linguistic* representation works. In either case it is simply a consequence of the ancient philosophical observation — understood by Kant, understood by the Socrates of the *Meno*[3] — that we have no access to the world *an sich*, no way to explain our representational capacities without using those very capacities.

The long version goes like this. After clarifying just how accounts of representation threaten circularity (section 2), I revisit the central philosophy of language question of what it is to learn a new *language* (section 3) and argue that (a) initially learning that language involves various forms of translation into an already-known language, but (b) eventually we transcend that translation and understand the language in its own right, and (c) that understanding does not consist of some introspected account of the language's meaning in an already-known language, but of correct usage — where 'correct' is to be assessed by other competent language-users and/or through Davidsonian radical interpretation. Applying these ideas to the way physicists learn new *theories* (section 4) leads to an account of what it is to interpret a mathematized theory closely analogous to David Lewis's famous distinction between *languages* (formal constructions equipped with semantics) and *language* (a human activity), which I call the 'theories and theorizing', or 'TT' account (section 5). With this account available, I consider the case for the alternative, 'language-first' approach to scientific represention (section 6) and argue that it is question-begging and that it falsely assumes a priority of linguistic over non-linguistic forms of representation in our pre-scientific activities. In section 7 I apply my account of representation to the question of when two physical theories are equivalent, where it provides a middle way between 'formal' and 'interpretational' accounts of equivalence. I conclude (section 8) with some broader thoughts as to how philosophers should engage with physical theories.

2 Representation and circularity

Probably the most common account of scientific representation[4] is that it is a form of *isomorphism*: mathematical model M represents physical system S be-

[3] I have in mind especially Socrates' discussion of the innate grasp of geometry.

[4] See, e. g. ,(French and Ladyman 1999; Redhead 2001; Bueno and French 2011; French 2003; Bartels 2006; Bueno and French 2011); for systematic discussion, see (Frigg and Nguyen 2020, ch.4) and references therein.

cause M and S are (perhaps partially) isomorphic. The intuition is clear: what we want to represent is the structure of a system; isomorphisms are structure-preserving maps; so if M is isomorphic to S, the mathematical structure of M is in an appropriate sense the same as the physical structure of S.

A standard objection to the isomorphism account (see, e. g. , (Frigg and Nguyen 2020, section 4.3) and references therein) is that it is insufficient, since radically different theories are formally isomorphic. A common example (Frigg 2002) is the simple harmonic oscillator, which can be used to represent systems as disparate as a pendulum and an electric circuit; another (Weatherall 2019) is the Langevin equation for Brownian motion, used to describe both the random-walk behavior of a pollen grain and the fluctuations in the share prices on a stock exchange. The suggestion is often that what is missing is something like the intention of the scientist: the harmonic oscillator does not simply passively represent a pendulum, it is used to represent the pendulum by a physicist, and the representation cannot be understood independently of that intention.

We will return to this issue of representation and intention shortly, but for now I want to raise a more basic concern about representation as isomorphism (a concern also recognized in the literature: see the discussion in (Nguyen and Frigg 2021) and references therein). Formally, an isomorphism is a relation *between mathematical entities.* In set-theoretic terms, for instance, we might characterize a mathematical model as a tuple $\langle \mathcal{O}, \mathcal{F} \rangle$, where $\mathcal{O}$ is a set of objects and $\mathcal{F}$ is a set of relations between those objects; given another such model $\langle \mathcal{O}', \mathcal{F}' \rangle$, an isomorpism between the two is a pair of maps $\rho : \mathcal{O} \to \mathcal{O}'$ and $\sigma : \mathcal{F} \to \mathcal{F}'$, so that $R(x_1, \ldots x_n)$ is true in the first model iff $\sigma(R)(\rho(x_1), \ldots \rho(x_n))$ is true in the second. Other notions of isomorphism — categorical equivalence, for instance — exist, but in each case they are again relations between mathematical systems. (In general in this paper I will set aside the question of whether the relevant mathematical relation for representation (and, later, theoretical equivalence) is set-theoretic isomorphism or something more sophisticated — these questions are important, but largely orthogonal to my concerns here.)

But then what does 'isomorphism' mean when we speak of isomorphism between a mathematical entity (or structure) and a *physical* system (or structure)? I can think of two strategies for making sense of it. Firstly, and most straightforwardly: we might say that mathematical structure A is isomorphic to physical system B iff B is represented faithfully by mathematical structure C and A is isomorphic to C. That notion is perfectly clear — yet it is a notion that relates different representations of the same physical system. It does not relate a mathematical model directly to the system it supposedly represents.

Secondly (and this seems to be the main strategy adopted in the literature[5]), we might possess a *linguistic* description of B, that says that it consists of (physical) objects $\mathcal{O}$, standing in (physical) relations $\mathcal{F}$. Then we can naturally associate a set-theoretic mathematical structure to B, and then that structure might be isomorphic to A in the familiar sense.

The first strategy tells us how a mathematical model represents a system only

[5]See, e. g. , (French and Saatsi 2006) or (Frigg and Nguyen 2020, section 4.6).

if we know already how to represent it by a mathematical model. The second tells us how to represent a system mathematically only if we know already how to describe it linguistically. In neither case have we an account of representation that does not already presuppose representation — although it might seem that the second strategy, but not the first, offers a way to understand mathematical representation without outright circularity, since it makes it derivative on the (supposedly) clearer notion of *linguistic* representation.

The general issue — that in order to explain how a model represents a system, we must already be able to represent that system — goes well beyond isomorphism accounts, as we can see by considering a very different proposal: the Gricean model proposed by Callender and Cohen (2006). To Callender and Cohen, scientific representation is downstream of mental representation: M represents S because a scientist intends M to represent S, and there is nothing more to it. But of course this intention needs to be articulated: the minimal statement that a scientist intends such-and-such system of differential equations to represent the flight of a space probe, say, is scientifically useless without some account of which features of the system represent which features of the probe's flight. We need to know, say, that the t coordinate in the model represents time in such-and-such units, that the x, y, z coordinates represent space in such-and-such coordinate system, and that the solutions of the differential equation represent possible trajectories of the space probe. And to know that we need to know already what 'such-and-such units', 'such-and-such coordinate system', and 'possible trajectories' actually mean. Any actually useful representation requires a detailed association between features of the model and features of the physical system; if that association holds simply by virtue of a scientist intending it to hold, that scientist must already understand what those latter features are.

I know of no extant account of representation that can be articulated without some (normally linguistic) prior representation of the system being represented.[6] But then the dilemma sketched in section 1 appears unavoidable: unless we rely on a prior notion of linguistic meaning, *mathematical* representation cannot be understood without already understanding it, and we seem trapped in a circle.

And indeed we are trapped. But — I will argue — this is unproblematic and, on reflection, only to be expected. My demonstration of this will be a bit indirect, via considerations of the very similar problems that arise in philosophy of language, when we try to provide a theory of linguistic meaning. Understanding how those problems are resolved in that context will provide insight into how to solve them for mathematical representation.

3 Learning new theories: the language-first story

What does it mean to give the referent of some natural-language word: say, 'London'? Well, if we happen to *be* in London, we can state it by ostension: *this* (waves arms) is London. But that is inconclusive even if we are so located

[6] See (Jacobs 2024) for a more extended argument to the same effect.

(do you mean London, or the South Bank, or England, or just this specific pub?) And mostly ostension is not available to us, either contingently (I'm not in London as I type this, and anyway you're not reading it over my shoulder as I do so) or necessarily (try ostending 'quark' or 'injustice'.) So, mostly, we explain reference in language: 'London is the capital city of Britain'; 'London is the big city upriver from the Thames estuary'; 'London is *Londres*'. In each case we assume our interlocutor already understands a language sufficiently rich to pick out London: then we explain how to translate between that language extended to include 'London' and that language without extension.

The same phenomenon occurs in formalized theories of meaning. What is the meaning of 'Snow is white'? A standard starting point is Tarski's 'T sentence':

> 'Snow is white' is true (in English) iff snow is white.

Whether understood extensionally (*a la* Davidson 2001) or intensionally (*a la* Lewis 1983), the statement is uninformative inasmuch as it will not tell you what 'snow is white' means unless you know already. One sees the value of such statements by supposing that the language and metalanguage differ:

> 'Schnee ist weiss' is true (in German) iff snow is white.

That's genuinely informative: if you speak English, but not German, sufficiently many such sentences will teach you the theory of meaning for German, the referents of German nouns, and the like. But it is informative *because* you already know a language that succeeds in describing those systems that others describe in German. What does 'Schnee' refer to? Snow! But this relies on that fact that you already know the referent of 'snow'. If you know a language adequate to describe phenomena in a given domain, you can learn another language by learning how to translate between the two, or by stating a theory of meaning of the new language in the old, but all of this will avail you nothing if you do not in fact know even that old language.

And now we confront the same problem as for mathematical representation: if our accounts of linguistic meaning for a given language only make sense when we presuppose that we already have a linguistic framework adequate to the system being described in that language, how are we supposed to give an account of meaning that is not circular?

We have already seen one answer: ostension. This was the positivists' answer to the puzzle, and a central motivation for their whole program: we can define x in terms of y, and y in terms of z, but ultimately we explain what z means by pointing to the world. To the positivists[7], the Verification Criterion for meaningfulness is not simply a demarcation criterion but a consequence of their theory of linguistic meaning: the meaning of a linguistic expression is given by a chain of definitional equivalences ending in something immediately given in experience.

[7] See, e.g., (Waismann 1979, pp.243-9), (Schlick 1979, pp.309-12,361-9), (Ayer 1936), all reprinted in (Hanfling 1981); see also Hanfling's introduction in the same volume.

But few accept their answer any more: the notions of sense data, or the 'given', or an unanalyzed 'object language' on which it relies have flaws too numerous and too well known to rehearse here. The general position in post-positivist philosophy is that there is no way out of this circle, but that the circle is not vicious: we have to take our current linguistic practice as a starting point, but we can still analyze parts of it using the resources from other parts. As Quine (1969, pp.126-127) famously put it:

> I see philosophy and science as in the same boat — a boat which, to revert to Neurath's figure as I so often do, we can rebuild only at sea while staying afloat in it. There is no external vantage point, no first philosophy.

To spell out how this has been developed in post-positivist analytic philosophy: suppose we have some theory, formulated in natural language, fairly adequate for a certain class of phenomena but not the last word in science. (For now, let us indulge the fiction that scientific theories after all are linguistically given: I wish to make contact here not with actual science but with the philosophical literature on scientific theories.) A scientist suggests a new theory, supposedly subsuming and improving upon the old one; never mind whether it is *true*, how are we to even *understand* it?

Part of the answer will be formal. The scientist can point out to us that some of the vocabulary of the two theories is in common: perhaps both speak of space, time, acceleration. Furthermore, she can explain how the statements of our old theory are derivable — perhaps with some corrections or tolerable approximations — from the new theory, even where there is vocabulary not in common between the two theories, if the new theory is supplemented by 'bridge principles' that translate between the two. From this starting point we can consider three answers to the question. (I don't intend these to be exhaustive of the literature, but I think they show its main shape from the positivists to the present.)

The positivists' answer: (As presented in, e. g. , (Schlick 1932); see (Dewar 2023) for a helpful review): the new theory has only a pragmatic or calculational value. Its theory of meaning is forever distinct from the old theory: whereas we understand the old theory as a genuine description of what is going on, the new theory is but a fiction that helps us derive claims in the old theory. Its assertions cannot be understood literally; its terms do not genuinely refer.

Lewis's answer: (As presented in (Lewis 1970)): Terms in the new theory are referential just as in the old. But if you want to understand how that reference works, or understand the meaning of the new predicates in the new theory, you need to look back to the old theory. The properties of the new theory are, definitionally, understood in terms of what they entail about the old theory.

Sider's answer: (As presented in (Sider 2011, section 7.4) and (Sider 2023)): Lewis's answer is right, *for a while*.[8] But eventually, you understand the new theory well enough that you can grasp it in its own terms. At that point, the logical order reverses: where once the new theory was to be understood by reference to the old, now the old theory is to be interpreted by reference to the new. (So, for instance, a physicist learning quantum theory might initially understand quantum concepts via their relation to the classical, but there comes a time when they have 'got the hang' of quantum theory, and have a grasp of what these terms mean without appealing to the classical connection. At this point, the understanding of the new theory has become direct.

Sider's answer, I think, is the nearest to a consensus answer in contemporary analytic metaphysics. (It might equally well be called *Quine's answer*, since it is close in spirit to Quine's idea that we rebuild our science and our philosophy as we go along: in the first place, we extend our theorizing by using our old theory to understand our new one; in due course, we hold fixed our new theory and, from that new starting point, reconceive our old one.) It suggests that we can eventually get a direct grasp of what a new theory (or new language) means, without that grasp being parasitic on our understanding of an old theory or language. What does that consist in? — what makes it true that someone understands that new language?

There are two ways to answer. One is by reference to other speakers of the new language. I (the expert, the fluent speaker of the new language) understand that language just fine. You (the student, seeking understanding through fumbling steps) seek to gain that same understanding. I can judge which utterances are correct and reasonable, and which are false or confused; insofar as your use of the language gains my imprimatur, then you understand it, and in your turn you can judge the students that come after you.

The second, and deeper, answer is externalist, via Davidson's notion of *radical interpretation* (Davidson 1973; Lewis 1974). I (the dispassionate linguist, the external observer) already speak and understand a language adequate to the phenomena we are jointly studying. You (the subject of study) make noises and draw sigils that look linguistic in nature. Here is a theory of meaning, stated in *my* language: *your* utterances, *your* symbols, can be interpreted as statements in a language L whose theory of meaning is M. Given that theory of meaning, your utterances and drawings make sense: they are reasonable utterances, reasonable drawings, on the understanding that you are speaking L, and hence that I should understand its content according to M. These facts are constitutive of the claim that you are speaking L.

In either case, there is no escape from the circle: I could not have judged that you were speaking L without my own representational tools, antecedently assumed to be understood: either my knowledge of L itself, or of another language whose representational capacities are at least as strong as L. (Note in

[8] To be fair, Lewis himself might endorse this: he is concerned with how we *initially* come to learn a new theory.

particular that my being a fluent speaker of L cannot consist of my knowing the explicit theory of meaning of L. That leads to a genuinely vicious regress, or else straight back to positivism — and in any case, it is plainly true observationally that few if any speakers of a natural language could actually state the explicit theory of meaning of that language.)

The basic ideas here recur in the more familiar context of learning a new word in an existing language. Take the somewhat obscure English word 'lachrymose': it is possible to someone be quite fluent in English without having ever come across it. Suppose that you have never come across it before now, and looked it up in a dictionary; you would find a definition like this (from Merriam-Webster[9]): "given to tears or weeping". And that is indeed *roughly* what it means: indeed, since dictionary writers are good at their jobs, it is probably as good a definition as one could reasonably expect. But it would be wrong, or at least an odd affectation, to say that a tantrum-prone toddler or colicky newborn, or someone unusually sensitive to onions, was lachrymose.

How do I know this? Because, as it happens, I know what the word 'lachrymose' means. Can I define it for you? Not any better than the dictionary-writers could. I can help you to understand what it means by giving you contexts in which it would or would not be appropriate; indeed, dictionary writers do this too (Merriam-Webster gives the example phrase "tended to become lachrymose when he was drunk"). But I cannot give you an exhaustive definition, because words do not work that way: they are not exhaustively defined by their definitions in terms of other words except in those rare circumstances where they are introduced explicitly that way (Quine 1951). My knowing what 'lachrymose' means is no more or less than my ability to use it correctly; that is, in accordance with the rules of English. And of course I don't *know* what those rules are, except tacitly: it is highly contested among linguists and philosophers of language what a theory of meaning for a language even is, and certainly I do not know how to state them explicitly — unless you allow me to use English, complete with 'lachrymose', as my metalanguage, in which case I can tell you that 'T is lachrymose' is true iff there is some x such that T refers to x and x is lachrymose. (You're welcome.)

To sum up: learning a new language, or a theory stated in a new language, or an extension of an old language, is in each case a two-step process. Initially we understand it through definitional links to the old language; but those links do not fully capture the new language; in due course we come to understand the new language directly, and then our understanding of the old language may be recast in terms of the new language; what consistutes understanding of the new language is correct use; no explicit theory of 'correct use' can be given except by one who already understands a language rich enough to describe the phenomena to which the new language applies.

Let us now see what happens when we apply this framework to mathematical, rather than purely linguistic, representation.

[9] "Lachrymose", *Merriam-Webster.com Dictionary*, Merriam-Webster, https://www.merriam-webster.com/dictionary/lachrymose. Accessed 8 Mar. 2024.

4 Learning new theories: the math-first story

Consider a physics student who attempts to learn general relativity. That student almost certainly knows plenty of physics already: they are pretty likely to have studied classical mechanics, nonrelativistic gravitation, electromagnetism, and special relativity already, and so they are familiar, for instance, with the idea of force-free motion, of fields, of the Newtonian gravitational potential, and of Minkowski spacetime. More specifically, they understand how to represent particular physical systems — colliding particles, say, or light rays, or the motion of the planets, or travellers who go to distant stars and return younger than expected — through mathematical models. (Perhaps the student cannot theorize interestingly about what, philosophically, such representations consist in, but if they can't actually construct and use them, they won't have got this far through their studies: it is constitutive of understanding classical mechanics, nonrelativistic gravitation, electromagnetism, or special relativity that one can do these things.[10])

When the student encounters general relativity, the new concepts are explained not *ab initio* but in terms of their existing understanding.[11] The idea of 'spacetime curvature' is introduced via the motion of force-free bodies; 'geodesic' is introduced as a generalization of 'straight line'; it will be demonstrated how Minkowski spacetime can be represented as a pseudo-Riemannian manifold with a certain flat metric, and how one can generalize that spacetime by dropping the requirement of flatness, and analogies will be drawn with flat and curved surfaces in three-dimensional Euclidean space despite the well-known weaknesses of those analogies; it may well be pointed out how the equivalence principle works in Newtonian gravity, and how that motivates the construction of general relativity; an analogy may be drawn to curvature of *space* (though that analogy can often mislead students: it is the curvature of space*time* that does the work in geometrical theories of gravity).

Fairly quickly, the student will encounter concrete models of general-relativistic spacetimes and will learn how to use them to represent physical systems. They will study geodesic motion in the Schwarzschild solution, and establish how it can be identified to subleading order with Newtonian orbital motion plus a small correction; they already know how a Newtonian solar-system model represents, and they learn via this how the new model does so. They will also study small perturbations around a flat metric, and be led to the discovery of a wave equation, structurally similar to those they know already, that can be understood as describing the propagation of disturbances across spacetime.

The Newtonian-gravity example brings out an aspect of scientific representation worth pausing on. There is both a kinematic and a dynamical aspect to the way Newtonian models represent, say, the motion of a planet around the sun.

[10] Kuhn (1962, 1970) famously recognizes this: his conception of a 'disciplinary matrix' explicitly recognizes central examples and methods of application as part of the content, and not just the pragmatics, of scientific theories.

[11] Here I draw *inter alia* on (Misner, Thorne, and Wheeler 1973), (Carroll 2003) and (Zee 2013).

Even without considering dynamics, Newtonian gravity gives us the resources to describe the motion of the Earth; it further makes dynamical claims about that motion, claims that can be true or false. So, for instance, Newtonian gravity *mis*represents the motion of Mercury, making claims about the precession of its perihelion that turn out to be not quite correct, even while it provides the representational machinery to *describe* that motion. (Lest that motion be considered directly observational and unmediated by theoretical representation, recall (Barbour 2001) that motions in Newtonian physics are relative to inertial frames and measurements of space and time that take significant, highly-theorized work to extract from the observational data.) It is precisely because of this possibility of misrepresentation that we can say that general relativity *corrects* Newtonian gravity in its account of Mercury's orbit.

In any case: what is happening here, I want to suggest, is a representational version of the first stages of learning a new *linguistic* theory. The student learns general relativity — learns how the models of general relativity represent — through relating that theory to the theories, and the representational tools, they know already.

But those relations are temporary heuristics: they cannot be taken as permanently fixing how general relativity represents, because the relations are partial and imperfect. We can understand the weak-field regime of the Schwarzschild solution through its connection to Newtonian methods of modelling the solar system — but those methods break down as the geometry becomes more strongly curved. The weak-field approximation establishes a tight and informative analogy between gravitational and electromagnetic waves — but not all fields are weak, and the analogy eventually comes apart. Even the idea of a geodesic as the trajectory of a force-free test particle can be sustained only so long — true massive point particles are an impossibility in general relativity, and the motion of realistic, finitely-large, particles is a complicated matter only approximately mapping to geodesic motion.

The former student of general relativity who has achieved mastery in the subject understands all of this. They know that Newtonian physics can be reinterpreted as a useful approximation to general relativity in certain regimes, and that special relativity can be reinterpreted likewise. They understand, directly, how to use models of general relativity to represent physical systems, and they cannot give a *perfect* explanation of how in terms that presuppose any less than an understanding of general relativity itself. The best they can do without that presupposition, the best they can do in explaining these representations to someone who does not know general relativity, is to show, partially, heuristically, how general relativity represents systems already represented — at least kinematically — within already-known theories. And that partial, heuristic understanding is the beginning of the path by which their interlocutor, too, learns general relativity.

We see here the second half of the two-step process. Initially the student understands the new theory through the relations it holds to already-understood concepts. But in time they lose the need for these relations, and reconstrue them in reverse order, as re-explicating elementary concepts in terms of the concepts of the more advanced and general theory.[12]

5 Theories and theorizing

What makes it *true* that the student understands general relativity? The question is closely analogous to one we have already asked about learning a new language, and the same answers are available. One appeals to expertise: the student understands general relativity insofar as they use it correctly, as judged by *other* competent users of general relativity. The other, and deeper, asks how external observers — aliens visiting Earth, say — would assess the student's (or, indeed, the experts') understanding. Let's suppose that these aliens already understand a substantially more advanced theory of gravity, T — perhaps full quantum gravity, perhaps 'low-energy quantum gravity' (Wallace 2022a), the quantum-field-theory version of general relativity. What, from their perspective, makes it *true* that the student understands general relativity?

I suggest that the answer needs to be a mathematized version of radical interpretation. The aliens first deduce that the student's theory is formally represented by mathematical structure X (this itself will be a non-trivial task for them). They then observe that there is some kind of partial isomorphism (or similar mathematical mapping; I will keep the details intentionally vague) between X and their theory T (which *ex hypothesi* is also a mathematical structure). On the assumption that features of X represent that which is represented by the corresponding features of T, the aliens can make sense of the student's actions; indeed, they can do so better that way than via any other assumption. (Again, I leave 'make sense of' intentionally vague.)

There is (again) a very close analogy with how interpretation works in language. David Lewis's influential 'Languages and Language' (Lewis 1983) distinguished 'languages' (formal objects, understandable without direct reference to human speakers of that language) from 'language' (a human activity). To Lewis, a language is a lexicon and syntax along with a theory of meaning for that language, which consists of some kind of set of mapping rules between terms, phrases and sentences in the language and entities, properties and propositions. What then makes it true that a given linguistic community is speaking that language is that interpreting them as doing so makes the best sense of their behavior. (Lewis advanced specific accounts both of what the mapping rules looked like and what 'makes the best sense of' consists in, but the details are less important than the general contours of the proposal.)

Here similarly we can distinguish between (interpreted) theories, and the communities of theorists who interpret them. Given a class of systems represented by T, an interpreted theory for that class of systems is a pair (X, ρ) where X is a mathematical structure and ρ — the *theory of representation* for the theory — is a morphism of appropriate kind from X to T. What makes it

[12]For a discussion of parallel ideas in quantum mechanics, see (Bacciagaluppi and Ismael 2015).

true, assessed from the external perspective, that a given scientific community is using (X, ρ) is that this assumption best makes sense of their behavior.

The account I offer here — call it the *TT account*, for 'theories and theorizing', in parallel to Lewis's terminology — bears strong resemblance to isomorphism accounts of representation, and indeed inherits many of its subdivisions: the various strategies in the literature to delineate what sort of morphism is appropriate for representation, for instance (surveyed in (Frigg and Nguyen 2020, ch.4)) can be taken over fairly directly to the present context. But it differs in presupposing, even to state what a theory of representation is, that we are already given the system to be represented *as represented*, so that these morphisms unproblematically relate mathematical structures, and do not relate mathematics to the system directly.

The TT account resolves (or, if you like, bypasses) what Frigg and Nguyen (2020, s4.5) identify as the central problem for isomorphism-based accounts of representation (see also their discussion in Nguyen and Frigg 2021): how we are to assign structure to a physical system in order to construe that structure as isomorphic to a mathematical representation of it. As we have seen, the usual starting presumption in these approaches is that the system consists of some collection of objects and that a structure for that system is a set-theoretic construction from those objects; on pain of Newman's objection (Newman 1928; Demopoulos and Friedman 1985)), we cannot permit that structure to be built from *arbitrary* set-theoretic constructions, and so we are led to restrict to some subset, dilineated by reference to the system's physical relations and properties.

From the TT point of view, what is going on here is that neither Frigg and Nguyen, nor anyone else, can access the genuinely-unrepresented system. Their discussion already presupposes a representation: specifically, a description in language, which permits us to speak of the objects and their physical properties and relations. The morphism in question is then a morphism from the new theory to the set-theoretic models of this linguistically-stated theory, or to some similar construction.

There is nothing contradictory about this setup, but as we have seen, it is question-begging against any math-first conception of physics, in that it grants primacy to *linguistic* representations of a system. Frigg and Nguyen recognize this, and indeed suggest that it undermines the structural realist claim that an isomorphism-based account of representation removes the need for linguistic descriptions of physical systems. But this follows only if one assumes exactly what the structural realist will wish to deny: that the representation of the system that we must have already in order to state how another representation works must be a representation in language. The TT account requires no such assumption: once we have recognized that an account of representation already requires a representation, there is no particular advantage to that representation being linguistic.

It is also instructive to compare the TT account to various proposals in the representation literature that tie representation to a scientist's intention. To Callender and Cohen as we have seen, that A intends to represent T by X is both necessary and sufficient for T to represent X. But even more structuralist accounts of representation have tended to respond to obvious counterexamples to the idea the isomorphism alone is representation by including a user's intention to represent as part of the definition of representation: see, e.g., (van Fraassen 2008, p.23), (Bueno 2010, p.94); see also (Frigg and Nguyen 2020, 4.3) and references therein.

There are two problems with centering intention in this way. The first (yet again familiar from philosophy of language) is that — at least as an account of generally-accepted *theories*, if perhaps not of specifically proposed *models* — is that it neglects the communal nature of physics. If our student of general relativity, asked what the second rank tensor g represents, answers 'the stress-energy of the matter fields', he is not to be understood as expounding his own personal interpretation of general relativity: he is simply *wrong* — as wrong as I would be if I told you that cats typically have three legs, no matter my protestations about my nonstandard usage of 'three'. (Of course, the student can explicitly state that he is not describing general relativity, but a new and idiosyncratic theory, just as I can state that I am speaking not English but a new language in which 'three' denotes what 'four' denotes in English, but neither changes the fact that absent that explicit statement, we are just wrong.)

Grice recognized this, and his notion of linguistic meaning as a sort of collective development of speaker meaning[13] (Grice 1989) is an alternative, and in some ways a precursor, to Lewis's framework: roughly speaking, to Grice, a linguistic expression means P if there is a shared convention among speakers to mean P by it. But whatever the virtues of this approach in the semantics of ordinary language, in the physics context it fails to address my second problem: if T represents X in such-and-such a way because someone intends T to represent X in that way, we cannot state their intention without presupposing a way (linguistic or otherwise) to represent X.

If one is content with a resolutely externalist approach to intention, perhaps this is acceptable: I intend T to represent X in such-and-such a way just if attributing that intention to me is the best fit to my actions. But (not least because my actions in this context will in large part consist in me using scientific theories in certain ways) this account is only thinly distinguished from the TT account. I think what is normally intended (so to speak) by advocates of these approaches is that intentions are transparent to the intender, so that my intending T to represent X, if cashed out in terms of existing representation T' of X, requires me to understand that existing representation, and to know how to use it to represent X. And this circularity, unlike the circularity in standard accounts of learning language or in the TT account, is genuinely vicious: if my representing X by T requires me to already represent X by T', then my representing X by T' requires me to represent X by T'', and so on *ad infinitum*.

What I think Callender and Cohen, and other advocates of intention-based approaches to representation, have in mind is that *linguistic* representation is

[13] Callender and Cohen note this aspect of Grice's program, but note that it 'won't concern us in what follows since it seems to be more directly relevant in the praticular case of linguistic representation than in other cases of representation, such as scientific representation'. They don't defend this claim, though.

different. Their account is not supposed to apply to the way language represents: that is something we understand differently, and which is presupposed by accounts of non-linguistic representation, and so the vicious circularity is avoided: X is represented by T provided that I intend T to represent X, and I state that intention in language, using a linguistic representation of X that is to be understood in a quite different way. We have already seen that something similar is going on in much of the literature on isomorphism-based representation. Indeed, Jacobs (2024) makes a strong case that this priority of linguistic description over mathematical representation is all but universal in the representation literature.[14]

I don't think there is any very good reason to adopt this explicit double standard, which uses linguistic description to avoid circularity in our account of representation without recognizing the parallel circularity in accounts of linguistic meaning but I agree with Jacobs that it plays a central, underacknowledged role in much of the literature on scientific representation. In the next section I explore it in more detail.

6 Against language-first representation

Here is an explicit statement of the double-standards view, which we might more charitably call *language-first representation*[15]:

> In the first instance, we describe the world in language, and thus understand it in terms of objects, properties, and relations. How we learn languages, and learn to apply them in new situations and extend their vocabulary, is a subtle and difficult problem in linguistics, the philosophy of language, and metaphysics, but it is appropriate to consider it solved before considering the separate question of how mathematical models represent. We then have the resources to directly describe a physical system, and so consider the substantive question of how that system is indirectly described via representation by a mathematical model.

I don't know anywhere in the literature where the position is stated quite that explicitly[16] (though it is close to the surface in (Nguyen and Frigg 2021)) but

[14]An important recent exception is Dewar (2023), whose account of representation overlaps with mine in some respects.

[15]Jacobs (2024) calls it *descriptivism*.

[16]The nearest I know is Wilhelm (2023). In his discussion of representation and theoretical equivalence in quantum field theory, he advocates that '[p]hilosophical accounts of content should generally be formulated using the theoretical terms, and basic guiding principles, of our best science of content: namely, the naturalistic science of linguistics'. As stated I find this confusing: linguistics is our best science of *linguistic* content, but I have never seen its practicioners assert dominion over *all* forms of content, and indeed cognitive science is replete with theorizing about content and representation in non-linguistic contexts (see, e. g. ,(Pitt 2022) and references therein). (I also don't know how to make non-tautologous sense of 'naturalistic science'.) I suspect Wilhelm assumes that scientists are representing the world *in language* (cf .his footnote 8), but this is question-begging against math-first representation

I also don't know how to make sense of much of the extant literature without tacitly assuming it. But irrespective of this, it is true that *if* language-first representation is correct, it dissolves most of the problems I have raised in this paper so far. Non-linguistic representation becomes a useful additional tool in science, but is to be understood entirely through its relation to linguistic representation, or more precisely to the world as represented linguistically. (And apparently substantive problems, like *why does mathematics actually help here?*, open up.) So it is worth seeing how defensible language-first representation is, whether or not it is tacitly or explicitly assumed elsewhere.

I see two immediate problems for language-first representationalism. Firstly, at least for deep physics it is manifestly the case that we completely lack these supposed non-mathematized descriptions of systems. Field (1980) endeavored to give a nominalist account of Newtonian gravity; Arntzenius and Dorr (2012) struggled heroically to extend that account to gauge field theory; the degree to which they succeeded is contestable but in any case these are hard-won reconstructions, not something scientists *in fact* understand and know. So at any rate our account of how mathematized physics represents a given system cannot rely on our *knowing* a linguistic description of that same system. And secondly, we have seen both that there is an unavoidable (but not vicious) circularity in how languages are learned, and that we can give a closely parallel account of how mathematical representation is learned. So it is hard to see what, beyond intuition and familiarity, privileges language over other forms of representation.

The advocate of the language-first approach has answers to both problems (both prefigured in section 1). As for the first, there are two stories that might be told. One is externalist: it is not necessary for working physicists to *know* the linguistic representation of the systems that they represent mathematically; it is necessary only that *there exists* such a linguistic representation. The other is revisionist: yes, we lack a proper language-first statement of our deep theories of physics: a proper accounting of what exists according to those theories and what properties it has. So much the worse for those theories: until and unless such an accounting is provided, we have but a calculational algorithm and not a true description of reality. (Maudlin (2018b), for instance, argues that a 'canonical presentation' of a theory, underdetermined by its normal mathematized form, is necessary to make sense of its contents; in the quantum regime, he and other philosophers of quantum mechanics have argued (Maudlin 2007, 2013; Allori *et al* 2008; Esfeld *et al* 2017) that only theories formulated with a so-called 'primitive ontology' can be understood as descriptive of the world.)

Tied to this revisionist response is the idea that it is conceptually impossible for us to understand what a theory says about the world except insofar as we can understand it as making linguistic claims about the world. Perhaps this is is a cognitive limitation (as might be suggested by Fodor's (1975) 'language of thought hypothesis'); perhaps it is an unanalyzable truth; but in any case, understanding a theory in language is constitutive of understanding it at all.

and indeed against the semantic view of theories: it does not follow from the fact that scientists communicate in natural language *about* their representational tools that those tools are themselves linguistic.

(Maudlin 2013, p.152) is fairly explicit here: “studying the mathematics in which a theory is couched is not the royal road to grasping its ontology” — from which it follows that there is more to a theory’s ontology than its mathematical presentation.

Without further development, though, this first answer is question-begging. *Why* should linguistic representation play such a special role in our general account of representation, and *why* should it be seen as a requirement on understanding that it be couched in language? Physicists claim to understand their theories, despite having no purely linguistic presentation of them; are they lying? Mistaken? What are our criteria for making those claims? Come to that, we don’t need to treat ‘physicists’ as some alien tribe here. *I* claim to understand plenty of bits of modern physics, even though I don’t know how to describe them in object-property terms; am *I* lying or mistaken? Do you, reader, understand general relativity? If so, do you know how to present it in object-property-relation form? We don’t generally accept a principle that some concept C cannot be understood unless it can be exactly and exhaustively described using some fixed set of resources. (I note that Maudlin, for instance, is extremely dismissive (Maudlin 2011) of anyone who claims not to understand what ‘probability’ is, or to ask for an explication of it in terms that do not already presuppose an understanding of it.)

This brings us to the second answer, which can provide the necessary underpinnings for the first. Linguistic representation (this second answer presumes) is something we understand pre-theoretically; indeed, our grasp of the manifest image is through linguistic representation. Prior to studying modern physics, we represent the world non-mathematically, through our linguistic descriptions of it. *Mathematical* representation is something new, something alien, something learned only since the scientific revolution, and so to mathematical representation must be understood in terms of linguistic representation.

I am not sure how decisive this would be, even if true. From a math-first perspective, linguistic representation (or, if you like, representation through set-theoretic models) is a perfectly valid form of representation. There is no particular problem telling the story we have told about how new mathematical representations are learned even if our first-known representations happen to be set-theoretic. The idea that we cannot go from knowing a set-theoretic representation to knowing a non-set-theoretic representation is again question-begging, except perhaps understood as a hypothesis about our cognitive capacities.

But in any case, it is not true. For one thing, we understand *space*: we have a clear and immediate understanding of the way in which matter can be spatially located. Probably it is innate (blank-slate theories where we theorize space on the grounds of sense data look extremely implausible neurologically) but in any case we have it, and not later than we have language. And spatial representation is mathematized: yes, there are presentations of Euclidean geometry in predicate logic, but they took thousands of years to develop and cannot plausibly be said to constitute our pre-theoretic conception of space. Much the same could be said of time; indeed, famously Kant regarded our basic intuitions of objects in space and of the unfolding of time as the conceptual core of, respectively, geometry and algebra. He didn't have a linguistic presentation of those intuitions in mind; indeed, his expectation that no purely logical presentation of geometry or algebra could exhaust its content, and if he was wrong about that as a matter of mathematics, still he was right that no such presentation is pre-theoretically known to us.

Comparatives more generally offer similar lessons. Even qualitative comparatives like 'nicer than' or 'smarter than', where there is no clear numerical basis, are naturally formalized via topology: there is a space of niceness or smartness properties, and that space has an ordering relation on it. First-year undergraduates learn how to treat relations like these via explicit assumptions of transitivity and asymmetry, but in general they seem to be working out how to cast an already-understood notion into the austere framework of first-order logic, not explicating the conceptual core of that notion. The point is only sharpened when we consider quantitative comparatives like 'twice as heavy as' or 'a third taller'. There is a recent literature on comparatives of length and mass, in which one finds (Martens 2022, pp.327-8) the idea that they should be formalized by positing a continuous infinity of length or mass properties, along with a long list of axioms very clearly constructed in order to imbue those properties with the structure of the positive reals up to scale transformations; whatever the virtues of this approach as metaphysics, it is implausible as an account of our pre-theoretic conceptual scheme.

Or take fluids. We don't have any particular problem understanding the idea of a continuum, or that at any point in a fluid, that fluid has a certain velocity or temperature, long before we get an explicit understanding of the fairly subtle mathematics required to treat these notions. (My three-year-old daughter can't yet count to twenty[17], let alone master the idea of smooth functions on a manifold; she has a perfectly clear understanding of the idea that the water is hotter in some parts of the bath than in other parts.) Nor do we struggle with the idea that each part of the fluid has a mass, and that we can define a density which is the fluid's mass per unit volume. These ideas were understood in antiquity, indeed probably prehistorically (the first boats long precede writing) but they are substantially mathematized and not straightforwardly cast into object/property form.

The point can be pushed both interpretationally and neuroscientifically. As for interpretation: if you were to try to ascribe representational content to humans on the basis of their behavior, your ascription would be rich with mathematics: the most natural way to represent our discourse and our non-linguistic activity with respect to space, time, mass, continuous matter, . . . , is mathematically, with tools that substantially exceed nominalistic first- or second-order logic. As for neuroscience: there is to my knowledge exactly no evidence that all of our representational capacities rely on some preferred language-first notion of representation, and a very extensive literature[18] presuming and/or arguing for the contrary. Of course it might be true (Fodor famously argued for

[17]Note added in proof: she has worked it out now.
[18]See, e. g. , (Pitt 2022) and references therein.

something similar) but it seems pretty implausible, not least because language, in the sense of compositional semantics, seems restricted to modern humans, whereas the ability to deal with space, time, and continuous matter is universal among vertebrates and probably beyond. (At the very least, it seems pretty implausible that humans developed a qualitatively different approach to (e.g.) spatial representation when they developed compositional semantics, and that spatial representation remains qualitatively different in humans compared to chimpanzees.)

I conclude that there are no good reasons to take language-first representation seriously. Our innate, pre-theoretic grasp of the world is heavily mathematized; we've known it at least since the *Meno*. We start off already understanding plenty of ways to represent that aren't, or aren't fully, linguistic; there is no barrier to us learning more.

This concludes my defense of the TT approach to representation. In the next section I apply it to the vexed question of theoretical equivalence.

7 Theoretical equivalence

The question of when two theories are 'theoretically equivalent', and the prior question of what that even means, have been central in recent discussions of structuralism and more broadly in the metaphysics of physics. One version occurs where two theories share the same mathematical formalism, and are used to model phenomena in the same way, but are described differently in language. General relativity provides a classic example: it is generally described as a theory of curved spacetime, but can also be described as the theory of a spin-2 field propagating on (at least locally) Minkowski spacetime. Setting aside for the moment issues of global topology, these are two verbal descriptions of the same mathematical formalism.[19]

From a language-first perspective this looks like a classic example of underdetermination, but on the present approach to representation it is nothing of the kind. From an externalist perspective, general relativity as an interpreted theory (described as always with respect to an antecdently-given representation of the physical systems to which it is applied) is a morphism of some kind between the mathematical formalism of general relativity and that antecedently-given representation. Two theories sharing the same mathematical formalism are the same theory insofar as they also share interpretation maps — and, insofar as facts about usage fix a community's interpretation of a theory, if a given physics community uses general relativity to model physical systems in exactly the same way irrespective of whether they describe it as curvature or a spin-2 field, there are not two versions of the theory at all but two distinct attempts to describe in natural language the systems to which the theory is applied.

In Wallace (2022b) I called such descriptions 'predicate precisifications' of the theory. It is important to understand that they are not descriptions *of the*

[19] More precisely: there are many different choices of mathematical formalism for general relativity, but the curvature-versus-spin-2 issue can be replicated in any of them.

theory in language (pretty much all our descriptions of theories, and indeed of mathematics more generally, are given in language) but of the world according to the theory: precisely because the theory is not linguistic, such descriptions generally draw distinctions much finer than the mathematized theory requires (hence 'precisification'). I don't want to understate the value of predicate precisifications: they can build heuristic understanding, aid pedagogy, facilitate communication, and suggest new applications and new generalisations.[20] But they are not part of the theory itself, and it is not factive which description to use: one chooses between them on conceptual and pragmatic rather than epistemic grounds, and it is not inherently contradictory to use one in some contexts and a different one in another.

Or rather: it is not factive which precisification to use *provided* that the description genuinely is inert in choosing how to model physical systems. This need not be the case: a naive[21] advocate of the spin-2-field approach might insist that the topology of spacetime be Minkowskian, a substantial restriction of general relativity's modelling capacities. That restriction picks out a genuinely distinct theory — but its distinctness can be characterized exactly through that topological restriction, with the predicate precisification playing at most a heuristic role.

More interesting cases arise where we have two theories that are mathematically distinct but where some relation of 'mathematical equivalence' (such as isomorphism or categorical equivalence) holds between the two. In general relativity, a mild example is the equivalence between tensorial and frame-bundle presentations of the theory (see, e. g. , (Rovelli 2004, ch.1)); a more radical one is the equivalence between the theory presented on differentiable manifolds and the so-called 'Einstein algebra' formalism (Rosenstock, Barrett, and Weatherall 2015). Examples abound beyond general relativity: field vs. vector potential vs. holonomy vs. fiber-bundle versions of gauge theory; Schrödinger vs. Heisenberg vs. path-integral versions of quantum mechanics; Hamiltonian vs. Lagrangian vs. Hamilton-Jacobi formulations of classical mechanics; inverse-square vs. potential vs. Newton-Cartan formulations of nonrelativistic gravitation; the dualities of quantum field theory and quantum gravity.

Here it will be helpful to situate the TT approach in contrast to other approaches in the literature. Following (Weatherall 2019), we can distinguish two broad approaches to theoretical equivalence. On the first approach, theoretical equivalence is formal equivalence plus the preservation of some partial interpretation, usually taken to be empirical content: two theories are equivalent if they agree on empirical facts and if a mathematical equivalence of an appropriate kind exists between them. On the second approach, theoretical equivalence is interpretational equivalence: two theories are equivalent "if they say the same thing about what the physical world is like" (Coffey 2014, p.834), and formal

[20] For more on this point see (Knox and Wallace 2023).

[21] In practice physicists who talk about general relativity in particle-physics language have a more sophisticated attitude to global topology: look at the AdS/CFT correspondence, for instance (see, e. g. , (Kaplan 2016) for a good technical presentation, or (Wallace 2018) for a philosophical discussion.

equivalence is at most a guide to interpretation and cannot settle theoretical equivalence.

From the TT perspective, neither account is fully satisfactory, though both capture important aspects of the truth. The formal-equivalence approach has the great virtue of capturing (a large part of) physical practice: mathematical reformulation is routine in physics and one searches the physics literature in vain for any sign of a residual question of physical equivalence once formal equivalence is demonstrated. But it requires a clear once-and-for-all identification of the empirical content of a theory, against the lesson taught by theory-ladenness of observation, and absent interpretative guidelines, it is unclear just what sort of formal equivalence might be required: candidates tend either to be too narrow (missing standard physics examples of equivalent theories) or too broad (eliding much or all of the 'non-empirical' content of a theory, granting for the sake of argument that an empirical/non-empirical line can be drawn).

The TT perspective offers a diagnosis of the problem here. The formal-equivalence approach tacitly assumes a theory of interpretation, but it is the mathematized analog of the positivist approach or at best the Lewisian approach: some fixed theory (sufficient to represent the supposed 'empirical content') is chosen, and theories are interpreted relative to that fixed theory, having content only insofar as they have implications for the fixed theory.

Defenders of an interpretational view would agree, I think: for two theories to be equivalent they should agree in their descriptions of the whole system, not just some fragment of it. But from the TT perspective, the interpretational view is too mired in language: equivalent theories must *say* the same thing about the world, and mathematically-presented theories don't *say* anything at all. The interpretational view, thus understood,[22] takes us back to a language-first approach to representation, and away from a naturalistic reading of physics practice.

But this suggests a natural way forward: retain the view that two theories are equivalent if they represent systems in the same way, but drop the requirement of a linguistic presentation. To be more precise: suppose a class of phyiscal systems is represented by T, and let (X_1, ρ_1) and (X_2, ρ_2) be theories for that class of physical systems (recall that ρ_1, ρ_2 are morphisms of some (unspecified) appropriate kind from X_1, X_2 into T, and also that no theory of representation is expressible except with respect to some background representation like T). Then a *theoretical equivalence* between the theories is an formal equivalence s between X_1 and X_2 with the property that $\rho_2 = \rho_1 \cdot s$, and two theories are theoretically equivalent if there exists a theoretical equivalence between them.

There is a direct parallel, once again, with issues in the philosophy of language. There, we want to know what it is for one language to be a translation of another: this is not just a formal matter but must be asked of languages as interpreted. Thus asked, the answer in outline is straightforward: it is a formal mapping between the languages of some appropriate kind which commutes with

[22]The requirement for interpretation to be in language seems tacit in (Coffey 2014) though it is not spelled out explicitly; for a completely explicit statement see (Maudlin 2018b).

the mapping(s) that provide the theory of meaning of the language. There are significant questions of detail to answer but the basic shape of the answer can be understood without resolving those details (and indeed provides important constraints which can help us fill in those details).

Of course, the TT definition of theoretical equivalence is *useful* in determining whether two theories are theoretically equivalent only if we have explicit knowledge of their representation maps (and hence a third theory already adequate to represent the systems which they represent). Sometimes this is precisely our actual situation: at present, for instance, general relativity is our best theory of gravitating systems and we can interpret the various forms of nonrelativistic gravity via their respective (approximate and restricted) embeddings in general relativity. That gives us the resources to apply the TT definition; the devil may be in the details but in general I would expect a fairly unequivocal result that formal equivalence suffices for theoretical equivalence in these cases.

Often, though, and particularly at the cutting edge of physics where the theories under discussion are the state of the art, we lack this explicit knowledge; even when we do have it, often we seem to be able to judge theoretical equivalence without it. (Discussions of equivalent formulations of Newtonian gravity rarely appeal to general relativity.) Here we need to look at the details of how the respective theories are introduced and used to judge theoretical equivalence (just as questions of linguistic translation are normally answered thus, and not through explicit appeal to theories of meaning).

The most straightforward case (and by far the most common) is when a new theory is introduced explicitly via its mathematical equivalence to an already-understood theory — as when a student of general relativity already familiar with its tensor formulation is introduced to the frame-bundle formulation, or when a student of quantum mechanics who learned the Schrödinger representation is taught the Heisenberg representation. Here the new theory is presented formally and its interpretation is entirely via its equivalence to the ordinarily-understood theory: the student knows how to use the new theory to represent because they know how to use the old theory to represent and they can pull that representation through the formal equivalence. More formally, the student has an interpreted theory (X_1, ρ_1), an uninterpreted theory X_2, a formal equivalence s, and then *defines* the interpretation of X_2 as $\rho_2 \equiv \rho_1 \cdot s$. Thus understood, formal equivalence immediately entails theoretical equivalence.

(It is constructive to contrast the language-first approach to the introduction of a new theory formulation. That introduction is normally not purely mathematical but accompanied by some kind of linguistic description: 'in this theory, gravity is not a force but a manifestation of the curvature of spacetime', for instance. The TT account regards this as heuristic: a helpful predicate precisification that aids understanding, but not an interpretation. The language-first account treats it as interpretation, so that it becomes an open question whether that interpretation is consistent with the (also linguistic) interpretation of the original theory or simply empirically equivalent.)

As usual I have left unspecified just what kind of mathematical map counts as a 'formal equivalence', but at least in this context it can be very broad. As an extreme example, consider some standard presentation of N-particle Newtonian mechanics; the class of models of this system has the cardinality of the continuum, so let ϕ be a quite arbitrary 1:1 map from those models into the real line $\mathbf{R}$. Then we can treat $\mathbf{R}$ as a model for N-particle mechanics, interpreting it via the map ϕ: to work out what system, say, 701.2117 represents, just map it back to the already-known theory.

Examples of this kind are sometimes taken as *reductios* of functionalist or structuralist approaches: how can a single real number represent a constellation of moving particles? But really they just demonstrate the pitfalls of supposing that an *uninterpreted* mathematical model represents anything at all. In this case it is clear that our understanding of the content of the new theory is, and always will be, derivative on our understanding of the old theory. (The situation is analogous to a new 'language' whose syntax and semantics are given by some complicated cryptographic operation on sentences of English.) There is nothing wrong with our new representation, except that it is largely useless.

What makes a formal equivalence useful? I see no reason to expect a systematic answer, but there are clear tells: for instance, can questions we want to ask about specific models be mapped in any tractable way to mathematical features of the new system or are we always going via the equivalence? And can we see how to relate the new theory to other theories (such as special or limiting cases of our previous theory) or is that again something that can be understood only through the old theory? Given a sufficiently useful theory, in due course we can probably come to understand its content directly, rather than derivatively — but this can be a matter of degree. (In much the same way, one initially understands a language through knowing how to translate it into an already-known language, but in due course comes to understand it directly — though the analogy is imperfect, since new languages are in practice never given an interpretation definitionally through their translation.)

A subtler case of equivalence arises in the context of discovery, where different groups of physicists develop alternative theories for a given class of systems and then a formal equivalence is discovered between them. The classic case is Dirac's demonstration of the equivalence between Schrödinger's wave mechanics and Heisenberg's matrix mechanics. Here, physicists were trying to develop a new theory and initially interpreted their attempts through its relation to an old theory (classical mechanics), something clear from the central role that classical concepts played in early attempts to make sense of the theory by Bohr and others. I think it's reasonable to say that, at the time of Dirac's result, neither Schrödinger nor Heisenberg had any particularly developed understanding of their respective theories *except*, Lewis-style, through their relationship to classical mechanics. And since Dirac's formal equivalence also preserved all the facts about the quantum-classical correspondence — in particular, the expectation values assigned to all classical quantities — then it would make sense for physicists of that time to treat that result as showing the theoretical, and not merely formal, equivalence of the two theories.

The subsequent development of quantum mechanics has arguably[23] seen it transcend classical mechanics: modern quantum theory makes extensive use of non-classical concepts like laser, Bose-Einstein condensate, or quasi-particle, and classical mechanics is understood as an approximation to or limiting case of quantum mechanics. But that development took place after the establishment of the formal equivalence of the Schrödinger and Heisenberg pictures, and physicists have learned better how to use quantum theory to represent even as they have freely moved from one to the other and back according to which better fits a given application.

The most difficult — if perhaps somewhat implausible — case of equivalence would arise if two distinct physics communities possess two mature and well-understood theories which, upon coming into contact with one another, they discovered to be formally equivalent. Does that formal equivalence establish a theoretical equivalence? From the TT perspective, in principle the answer is clear: it does so iff the formal equivalence map commutes appropriately with representation. But of course that answer is accessible to either community only if they already possess explicitly the interpretation maps of both theories, which is unlikely[24] to be the case.

In that case, the two communities face a mathematized version of Quine's problem of radical translation: the relation of formal equivalence between the theories is a hypothesis of theoretical equivalence, and the test of that hypothesis is whether the two theories are used in the same way to model physical systems. If the two communities share some other theory — say, if their two theories are two versions of General Relativity, and they share Newtonian gravity — then agreement on the relation between Newtonian gravity and their two theories (modulo formal equivalence) is strong evidence of a theoretical equivalence. So — assuming they share a natural language — is agreement on a predicate precisification of the theories. But neither is conclusive: there is no definitive check for theoretical equivalence, any more than there is a definitive check for translation.

In fact, the presence in the example of formally equivalent but distinct theories is unnecessary. Exactly the same issue would arise if the two communities had formally *identical* theories. Absent sameness of application, that is no guarantee of theoretical equivalence, as examples like the harmonic oscillator and Brownian motion make clear. Indeed, I am not certain that cases like this do not occur in actuality. There are quite deep differences between the way General Relativity is used by physicists in the particle-physics tradition (usually in string theory and string-theory-related approaches to quantum gravity) and by physicists in the mathematical-relativity community, and I don't think it would be unreasonable to describe those communities as espousing slightly different theories (although there is probably semantic indeterminacy between that hy-

[23] This is somewhat controversial, and related to one's stance on the measurement problem; it presupposes what I have elsewhere (Wallace 2016) called the 'Decoherent View'.

[24] Unlikely, but not impossible: the two theories might be distinct lower-energy approximations to a third theory that is common to the two communities, in which case establishing theoretical equivalence will be straightforward.

pothesis and the hypothesis that one or other community is slightly mistaken about a shared theory). But that difference is not a difference of formalism: the communities have somewhat different preferences as to which formalisms of general relativity are most useful, but they agree that the various formalisms are fully equivalent and that the question of which to use is pragmatic rather than factive.

8 Conclusions

> Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and to read the letters in which it is composed. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures, without which it is humanly impossible to understand a single word of it: without these, one goes wander- ing about in a dark labyrinth. (Galileo[25])

Since Galileo it has become a cliché to say that mathematics is the language of physics. Taken literally it is false: languages have quantifiers and compositional semantics; topological spaces and Banach algebras do not, and the fact that we use language to describe topological spaces and Banach algebras does not make them languages themselves, any more than the use of language to describe a hippopotamus makes hippos into languages. But mathematics *is* the representational tool of physics, as natural language is the representational tool for everyday life.

Trying to understand that representational tool, to understand what mathematical representation is, leads us to the apparent paradox that we cannot say how a thing is represented without already knowing how to represent it. It is tempting to escape that paradox by making mathematical representation derivative on linguistic representation, but this fails to take seriously the core of truth in the cliché: it demotes mathematics from the 'language' of physics to a mere formal auxiliary of natural language. More importantly, it offers no true escape: the way a language describes a system cannot itself be described without already possessing a way to describe that system. The apparent new paradox of mathematical representation is just a recurrence of the old truth that the world is not known to us in itself but only through one or another representative tool. Appreciating these lessons clears the way to develop an account of how theories in physics represent, and when they are representationally equivalent, that is not derivative on but is closely parallel to our accounts of representation and translation in natural language, and which therefore can draw deeply on the lessons we have learned in philosophy of language about the subtleties of developing those accounts.

[25] Taken from Drake (1957, pp.237-8).

Along the way we get a clearer picture of what is involved in understanding a theory of physics, beyond a grasp of its formal mathematical structure. That understanding does not solely or even primarily consist in a verbal description of what the theory says about the world (those verbal descriptions are often heuristically and pedagogically useful but do not constitute interpretation). It consists in understanding how the theory is to be applied and used, which normally includes some learned facility in modelling, approximating, and calculating. If one does not know at least something about how a theory can be used in this way, one does not understand it. The point is of course very familiar in physics pedagogy (pretty much any textbook contains the exhortation that you have to do the problems and the calculations to properly understand the theory) and is recognized by historians of science (Kuhn's notion of a 'disciplinary matrix' incorporates it) but is not always appreciated by philosophers.[26]

I suspect the account of representation I advocate here will seem antimetaphysical, even antiphilosophical. Jacobs (2024) gives a very clear version of this concern:

> On any account of interpretation compatible with [math]-first realism, however, philosophy of physics becomes a radically different enterprise. If nativism [Jacobs' term for, roughly, the position I espouse here] is true, mathematical models acquire content by the way in which they are used in practice. The aspirant-interpreter should therefore engage in a project of radical translation by studying the way in which physicists use their theories — or 'go native' and join the ranks of scientists themselves'.

But things aren't as radical as that. It is already widely recognized that philosophers of physics need to know some physics, after all. If '"go native" and join the ranks of scientists' just means 'take some physics classes', perhaps the demand might seem less onerous (though it does also mean that to learn a physics theory it is insufficient to just read the chapter that lays out the formalism, grow impatient with how imprecisely the chapter explains what the formalism means, and then skip to the philosophy literature!)

Yet again, there are familiar parallels in language. A modern philosopher can learn a great deal from reading (say) Plato in a good scholarly translation; indeed, all I have learned from Plato is through translation. But Plato scholars will universally say — and I believe them — that the deepest understanding of Plato requires the ability to read his writings in the original Greek. Similarly, much of what is philosophically important about a theory of physics can be learned through a verbal description of that theory, and even more if that verbal description is combined with an understanding of the theory's mathematical formalism. But the deepest understanding of the theory cannot be gained thus: physical theories are not given in natural language and cannot losslessly be

[26] I have heard a prominent philosopher of physics repeatedly claim that the education of a physicist is concerned with how to use a theory to calculate and model, and that this has exactly nothing to do with understanding the theory's foundations.

translated or interpreted that way. To fully understand a physical theory it is necessary to learn how it actually used to model physical systems and how those models interface with other aspects of our physics. It is insufficient to learn the abstract formalism: one must learn to represent.

Acknowledgements

This paper has benefitted enormously from conversations with Neil Dewar, Caspar Jacobs, Eleanor Knox, James Ladyman, Kerry McKenzie, Tushar Menon, Simon Saunders, Jim Weatherall, and the participants in two workshops at UC Irvine in 2023.

References

Allori, V., S. Goldstein, R. Tumulka, and N. Zanghì (2008). On the common structure of Bohmian mechanics and the Ghirardi-Rimini-Weber theory. *British Journal for the Philosophy of Science 59*, 353–389.

Arntzenius, F. and C. Dorr (2012). Calculus as geometry. In *Space, Time and Stuff*. Oxford University Press.

Ayer, A. (1936). The principle of verifiability. *Mind 45*, 199–203.

Bacciagaluppi, G. and J. Ismael (2015). Review of the emergent multiverse. *Philosophy of Science 82*, 129–148.

Barbour, J. (2001). *The Discovery of Dynamics*. Oxford: Oxford University Press.

Bartels, A. (2006). Defending the structural concept of representation. *Theoria 21*, 7–19.

Bueno, O. (2010). Models and scientific representations. In P. D. Magnus and J. Busch (Eds.), *New Waves in Philosophy of Science*, pp. 94–111. Hampshire: Palgrave Macmillan.

Bueno, O. and S. French (2011). How theories represent. *British Journal for the Philosophy of Science 62*, 857–894.

Callender, C. (1999). Reducing thermodynamics to statistical mechanics: The case of entropy. *The Journal of Philosophy 96*, 348–373.

Callender, C. and J. Cohen (2006). There is no special problem about scientific representation. *Theoria 21*, 67–85.

Carroll, S. (2003). *Spacetime and Geometry: an Introduction to General Relativity*. San Francisco, CA.: Addison Wesley.

Coffey, K. (2014). Theoretical equivalence as interpretative equivalence. *British Journal for the Philosophy of Science 65*, 821–844.

Davidson, D. (1973). Radical interpretation. *Dialectica 27*, 313–328.

Davidson, D. (2001). *Inquiries into Truth and Interpretation* (2nd ed.). Oxford University Press.

Demopoulos, W. and M. Friedman (1985). Bertrand Russell's *The Analysis of Matter*: its historical context and contemporary interest. *Philosophy of Science 52*, 621–639.

Dewar, N. (2023). Interpretation and equivalence; or, equivalence and interpretation. *Synthese 201*, 119.

Drake, S. (1957). *Discoveries and Opinions of Galileo, translated with an Introduction and Notes*. New York: Doubleday and Co.

Esfeld, M., D.-A. Deckert, and A. Oldofredi (2017). What is matter? the fundamental ontology of atomism and structural realism. In A. Ijjas and B. Loewer (Eds.), *Philosophy of Cosmology*. Oxford: Oxford University Press.

Esfeld, M., D. Lazarovici, V. Lam, and M. Hubert (2017). The physics and metaphysics of primitive stuff. *British Journal for the Philosophy of Science 68*, 133–161.

Field, H. H. (1980). *Science Without Numbers: A Defence of Nominalism*. Princeton, NJ: Princeton University Press.

Fodor, J. A. (1975). *The Language of Thought*. Hassocks: Harvester Press.

French, S. (2003). A model-theoretic account of representation (or, i don't know much about art... but i know it involves isomorphism). *Philosophy of Science 70*, 1472–83.

French, S. (2014). *The Structure of the World: Metaphysics & Representation*. Oxford: Oxford University Press.

French, S. and J. Ladyman (1999). Reinflating the semantic approach. *International Studies in the Philosophy of Science 13*, 103–21.

French, S. and J. Ladyman (2003). Remodelling structural realism: Quantum physics and the metaphysics of structure. *Synthese 136*, 31–56.

French, S. and J. Saatsi (2006). Realism about structure: the semantic view and nonlinguistic representations. *Philosophy of Science 73*, 548–559.

Frigg, R. (2002). Models and representation: why structures are not enough. Measurement in Physics and Economics Project Discussion Paper Series, DP MEAS 25/02.

Frigg, R. and J. Nguyen (2020). *Modelling Nature: An Opinionated Guide to Scientific Representation*. Switzerland: Springer Nature.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hanfling, O. (1981). *Essential Readings in Logical Positivism*. Oxford: Basil Blackwell.

Jacobs, C. (2024). Stating maths-first realism: Some tentative proposals. https://philsci-archive.pitt.edu/23142/.

Kaplan, J. (2016). Lectures on AdS/CFT from the bottom up. http://sites.krieger.jhu.edu/jared-kaplan/files/2016/05/AdSCFTCourseNotesCurrentPublic.pdf.

Knox, E. and D. Wallace (2023). Functionalism fit for physics. https://philsci-archive.pitt.edu/22655/.

Kuhlmann, M. (2010). Why conceptual rigour matters to philosophy: on the ontological significance of algebraic quantum field theory. *Foundations of Physics 40*, 1625001637.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kuhn, T. S. (1970). Postscript-1969. In *The Structure of Scientific Revolutions* (2nd enlarged ed.)., pp. 174–210. Chicago: University of Chicago Press.

Ladyman, J. (1998). What is structural realism? *Studies in History and Philosophy of Science 29*, 409–424.

Ladyman, J. and D. Ross (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy 67*, 427–446. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1974). Radical interpretation. *Synthese 23*, 331–44. Reprinted in David Lewis, *Philosophical Papers*, Volume I (Oxford University Press, Oxford, 1983).

Lewis, D. (1983). Languages and language. In K. Gunderson (Ed.), *Philosophical Papers*, Volume 1. Oxford: Oxford.

Martens, N. C. M. (2022). Machian comparativism about mass. *British Journal for the Philosophy of Science 73*, 325–349.

Maudlin, T. (2007). Completeness, supervenience and ontology. *Journal of Physics A 40*, 3151.

Maudlin, T. (2011). Three roads to objective probability. In C. Beisbart and S. Hartmann (Eds.), *Probabilities in Physics*, pp. 293–320. Oxford: Oxford University Press.

Maudlin, T. (2013). The nature of the quantum state. In A. Ney and D. Albert (Eds.), *The Wave Function*, pp. 126–153. Oxford University Press.

Maudlin, T. (2018a). Ontological clarity via canonical presentation: Electromagnetism and the aharonov–bohm effect. *Entropy 20*, 465.

Maudlin, T. (2018b). Ontological clarity via canonical presentation: Electromagnetism and the Aharonov–Bohm effect. *Entropy 20*, 465. https://doi.org/10.3390/e20060465.

McKenzie, K. (2014). Priority and particle physics: Ontic structural realism as a fundamentality thesis. *British Journal for the Philosophy of Science 65*, 353–380.

McKenzie, K. (2024). Structuralism as a stance. *Philosophy of Physics 2*, 1.

Misner, C. W., K. S. Thorne, and J. A. Wheeler (1973). *Gravitation*. New York: W.H. Freeman and Company.

Newman, M. (1928). Mr Russell's "causal theory of perception". *Mind 37*, 137–148.

Nguyen, J. and R. Frigg (2021). Mathematics is not the only language in the book of nature. *Synthese 198 (Suppl. 24)*, S5941–S5962.

Pitt, D. (2022). Mental Representation. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University.

Quine, W. (1969). Natural kinds. In *Ontological Relativity and Other Essays*, pp. 114–138. New York: Columbia University Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review 60*, 20–43.

Redhead, M. (2001). The intelligibility of the universe. In A. O'Hear (Ed.), *Philosophy at the New Millenium*, pp. 73–90. Cambridge: Cambridge University Press.

Rickles, D. (2016). *The Philosophy of Physics*. Cambridge: Polity Press.

Rosenstock, S., T. W. Barrett, and J. O. Weatherall (2015). On Einstein algebras and relativistic spacetimes. *Studies in the History and Philosophy of Modern Physics 52*, 309–316.

Rovelli, C. (2004). *Quantum Gravity*. Cambridge: Cambridge University Press.

Saunders, S. (2003). Structural realism, again. *Synthese 136*, 127–133.

Saunders, S. (2016). The emergence of individuals in physics. In A. Guay and T. Pradeu (Eds.), *Individuals Across the Sciences*, pp. 165–192. New York: Oxford University Press.

Schlick, M. (1932). Positivismus und realismus. *Erkenntnis 3*, 1–31. Reprinted and translated in M.Schlick, *Philosophical Papers, volume II (1925-1936*, edited by H. L. Mulder and B. van de Velde-Schlick (Dordrecht:Reidel, 1979).

Schlick, M. (1979). *Philosophical Papers, volume II (1925-1936)*. Dordrecht: D.Reidel. edited by H. L. Mulder and D. van de Velde-Schlick.

Sider, T. (2011). *Writing the Book of the World*. Oxford: Oxford University Press.

Sider, T. (2023). 3D in high-D. *Journal of Philosophy*, forthcoming; preprint at https://tedsider.org/papers/3D_in_high-D.pdf.

van Fraassen, B. (2008). *Scientific Representation.* Oxford: Oxford University Press.

Waismann, F. (1967/1979). *Wittgenstein und der Wiener Kreis.* Oxford: Basil Blackwell. Translated by J.Schulte and B.McGinnis as *Wittgenstein and the Vienna Circle.*

Wallace, D. (2016). Interpreting the quantum mechanics of cosmology. http://philsci-archive.pitt.edu/15295/.

Wallace, D. (2018). The case for black hole thermodynamics, part II: statistical mechanics. *Studies in the History and Philosophy of Modern Physics 66*, 103–118.

Wallace, D. (2022a). Quantum gravity at low energies. *Studies in History and Philosophy of Science 94*, 31–46.

Wallace, D. (2022b). Stating structural realism: mathematics-first approaches to physics and metaphysics. In J. Hawthorne (Ed.), *Philosophical Perspectives Volume 36: Metaphysics*, pp. 345–378. Wiley-Blackwell.

Weatherall, J. O. (2019). Theoretical equivalence in physics. *Philosophy Compass 14.* In two parts; https://doi.org/10.1111/phc3.12591 (part 1), https://doi.org/10.1111/phc3.12591 (part 2).

Wilhelm, I. (2023). The propositional account of effective theories. *Philosophy of Science*, forthcoming; DOI: 10.1017/psa.2023.117.

Zee, A. (2013). *Einstein Gravity in a Nutshell.* Princeton, NJ: Princeton University Press.