SYMPOSIA PAPER

Equivalence and Convention

Neil Dewar

Faculty of Philosophy, University of Cambridge, Cambridge, UK
Email: nad42@cam.ac.uk

Abstract

The goal of this article is to analyze the role of convention in interpreting physical theories—in particular, how the distinction between the conventional and the nonconventional interacts with judgments of equivalence. We will begin with a discussion of what, if anything, distinguishes those statements of a theory that might be dubbed "conventions." This will lead us to consider the conventions that are not themselves part of a theory's content but are rather *applied* to the theory in interpreting it. Finally, we will consider the idea that what conventions to adopt might, itself, be regarded as a matter of convention.

1 Conventions within a theory

A major component of the logical-empiricist program—associated especially with the work of Carnap—was the project of analyzing a theory into its "factual" and "conventional" (or synthetic and analytic) components. However, this project has fallen into disfavor: the idea of a hard-and-fast distinction between the factual and the conventional is widely held to have been dealt a decisive blow by Quine (1951). Quine's argument may be summarized as based on two compelling observations. The first is that any attempt to explicate the analytic–synthetic distinction only leads us in a circle of tightly interconnected concepts (of meaning, synonymy, etc.). The second is that we cannot distinguish in any robust fashion between those parts of a theory that are immune to revision and those that are subject to empirical input. If an observation runs contrary to a theory's predictions, then the theory must be modified *somewhere*, but there is no matter of fact about where the modification must land. No part of a theory, says Quine, is not, in principle, available for modification in light of recalcitrant evidence.

And yet there do, in fact, seem to be clear cases of conventions in physical theories: consider the statement of a gauge condition, or a commitment to a particular system of units.[1] So we have a puzzle. On the one hand, Quine's analysis appears to provide

---

[1] I hope that what I say here might apply outside of physics, too, but for the sake of not exposing my ignorance, I will confine myself to explicit discussion only of examples from physics.

very convincing *general* reasons for being skeptical that the conventions in a theory can be singled out. But on the other, we seem to have at least some clear examples where it is possible to do this.

The purpose of this first section is to see if we can, after all, find some properties that are characteristic of conventions; later, we will consider how to reconcile this with Quine's critique. As a starting point, we will follow Quine (1936) by taking *definitions* to be paradigmatic examples of conventions. Let us take a definition to be a statement that fixes the meaning of some newly introduced term by assigning it the meaning already associated with some complex of existing terms. Examples might include "let the kinetic energy of a body be $\frac{1}{2}mv^2$," or "let tan $x = \sin x/\cos x$." If one takes a theory T and augments it with definitions, then the resulting theory is said to be a *definitional extension* of T.

However, not all conventions are definitions: not all conventions introduce new vocabulary, and even where they do, they need not fix the meaning of that vocabulary uniquely. For example, suppose we introduce the gauge potential to electromagnetism by stipulating that it must obey the condition $F_{ab} = \partial_{[a}A_{b]}$. This condition is plausibly understood as a convention governing the use of this new symbol. But it does not provide a *definition* of A_a: it does not fix the meaning of A_a; it merely constrains it.

Thus, we must generalize beyond definitions. An important feature of definitional extensions is that they are *conservative* extensions. Recall that a theory T^+ is a conservative extension of T if for any sentence ϕ in the language of T, $T \models \phi$ if and only if (iff) $T^+ \models \phi$. On the face of it, conservativeness is a plausible condition to require of conventions: surely a mere convention should not, by itself, add content. Moreover, one finds that conservativeness is often presupposed in philosophical analysis of conventions—for example, in Gödel's critique of Carnap's conventionalism about mathematics (Gödel 1995).[2] So perhaps we should demand that supplementing a theory with a convention yields a conservative extension of that theory.

This still cannot be right as it stands, though. Once again, consider the theory of electromagnetism formulated in terms of potentials. Imposing a gauge condition is surely a convention. Yet doing so will, in general, have nontrivial consequences. For example, if the Lorenz gauge condition $\partial_a A^a = 0$ is imposed, then Maxwell's equations may be reexpressed as $\partial_a \partial^a A^b = J^b$ (where J^b is the current). This is not a condition that can be derived from the ungauged version of the theory; hence, the gauged theory is not a conservative extension of the ungauged theory. More generally, conservativeness will trivialize in cases where we are not introducing new vocabulary: if the theories T and T^+ are both in the same language, then T^+ is a conservative extension of T just in the case that the two theories are logically equivalent. So it would seem that strict conservativeness is too demanding a condition to impose on putative conventions.

However, we have been neglecting an important feature of the ungauged theory of potentials: the fact that it contains *surplus structure*. A consequence of this is that assessing what counts as "new content" is harder than it might seem. True, the gauged theory lets us derive new equations, but it does not let us derive any new *gauge-invariant* equations. Because it is only the gauge-invariant content that is to be

---

[2] See Warren (2020) and Marschall (2021) for more discussion of Gödel's argument.

considered robustly physical, adding the Lorenz gauge condition does not add to the physical content of the ungauged theory. Let us say, then, that an extension T^+ of a theory T is *conservative up to invariance* if for every *invariant* sentence ϕ in the language of T, if $T^+ \models \phi$, then $T \models \phi$. The notion of invariance will be understood as the preservation of truth-value across equivalent models of T: ϕ is invariant if for any model M of T, ϕ is true in M iff it is true in all models equivalent to M.

Thus, I propose the following condition for something being classified as a convention: a statement $\phi \in T$ may be regarded as a convention, relative to some theory $T' \subseteq T$, just in the case that $T' \cup \{\phi\}$ is a conservative extension of T' up to invariance. Thus, different standards of equivalence will yield different verdicts on what constitutes a convention. For example, if the standard of equivalence adopted is that of empirical equivalence, then one obtains the classic logical-empiricist position: any claim added to a theory that does not affect its predictive outputs is a mere convention. In the next section, we will devote more attention to the issue of what the criterion of equivalence should be.

Moreover, whether ϕ is a convention is also relative to both the theory T in which ϕ is embedded and the subtheory T' that one takes to capture the nonconventional or factual content of T. Of course, if one has a particular method for determining T' from T, one need not specify the two theories separately. One natural choice, for example, would be to simply let $T' = T \backslash \{\phi\}$. But this means that ϕ might be a convention when regarded as a sentence of T but not when regarded as a sentence of some theory logically equivalent to T. Alternatively, one could follow Carnap and identify the theory T' with the Ramsey sentence of T so that the conventional content of T is given by the Carnap sentence $(T' \rightarrow T)$.[3]

However, I think we are better off not committing to one particular way of extracting a theory T' from the theory T. This is because recognizing that conventionality is relative to a choice of T' gives us a way to make sense of the observation with which we began: namely, that we seem to have clear examples of conventions in science, despite Quine's critique. The resolution is that it does make sense to take a given theory (and a given standard of equivalence) and argue that *adding* a statement to that theory amounts to adding a mere convention. This is, indeed, precisely what we did in the case of gauge conditions in electromagnetism. What does not make sense is seeking a general prescription for how to identify the conventional components of a given theory—unless, like Carnap, one gives a general prescription for identifying T' from T. Absent such a prescription, identifying a convention is a one-way process. One can identify that adding such and such a claim to a theory would be merely to add a convention, but one cannot say that such and such a claim that has *already been added* to a theory is a convention. This ties in nicely with a remark of Putnam's: "Quine has suggested that the distinction between truths by stipulation and truths by experiment is one which can be drawn only at the moving frontier of science. Conventionality is not 'a lingering trait' of the statements introduced as truths by stipulation" (Putnam 1962, 371).

---

[3] Note that a theory is always a conservative extension of its Ramsey sentence (Button and Walsh 2018, proposition 3.5).

2 Conventions about theories

Assessing whether a statement is a convention, then, depends (in part) on determining which models of the theory T' are equivalent to one another. The relevant sense of equivalence here is that of physical or theoretical equivalence—that is, whether the models in question depict the same state of the world or not. Gauge equivalence provides one example. More generally, symmetries are a good example of the kind of phenomenon at play here. For example, suppose one formulates a theory of N Newtonian particles using coordinates. One could then impose the condition that the center of mass of the system is to be at rest. This is plausibly regarded as a convention, but only if we regard models related by a boost to be physically equivalent to one another. If they are not so equivalent, then "the center of mass is at rest" is a hypothesis, not a convention. So it seems that to settle the question of whether a theoretical statement is a convention, we need to address questions such as whether symmetry-related models are physically equivalent or not.

This is a debate with a sizeable (and still-growing) literature.[4] In the past, I have tended to regard this debate as one with a determinate answer: namely, that symmetry-related models are indeed physically equivalent. Now, however, I am inclined to take a somewhat different attitude. It seems to me to be better to say that this, too, is an issue of what conventions to adopt. Unlike the conventions we have considered so far, however, these conventions will not be conventions *within* a theory; rather, they are conventions *about* the theory. This is because one cannot have a convention within the theory that stipulates that a pair of models are equivalent to one another: merely consider the question of how one might try to indicate, within the theory of electromagnetism, that gauge-equivalent potentials are equivalent. A statement of the form $A_a = A_a + \partial_a \lambda$ accompanied by the assertion that λ can be any scalar field, for example, will make the theory inconsistent.

From this perspective, the question to ask is not "Are symmetry-related models physically equivalent?" but rather, "What are the pragmatic advantages or disadvantages of treating symmetry-related models as physically equivalent?" Treating them as equivalent has various pragmatic advantages: it avoids concerns about underdetermination; it lets one freely choose whichever model is most calculationally convenient; and for local symmetries (i.e., gauge symmetries), it will make it possible to have a well-posed initial-value problem.

Let us say, then, that the decision to treat certain models as equivalent to one another is a semantic convention. However, this is not the only kind of semantic convention that is important. There are also conventions concerning how the theory relates to the world. Clearly, these kinds of conventions play some kind of important role—not least, in determining relationships of equivalence. Van Fraassen (2014) observes that the equation describing heat diffusion is formally identical to that describing gas diffusion, and hence the difference between them must be a matter of their physical interpretation. In a similar vein, Sklar (1982) notes that the statements "all lions have stripes" and "all tigers have stripes" are formally intertranslatable—but, again, would typically receive different interpretations.

---

[4] See Brading and Castellani (2003) and references therein.

So, it seems, it is not enough to delineate a theory's internal standards of synonymy: one must also describe that theory's relationship to the world. This idea is pervasive in recent philosophy-of-physics literature. For example, Maudlin (2018) argues that any theory must specify a physical ontology, not just the mathematical representation of that ontology. For another, De Haro and Butterfield (2018) make use of "interpretation maps," which "map from our theories and models, to 'meanings' and to 'the world'" (322). Indeed, one might even think that providing this kind of interpretation isn't even needed *in addition to* the sort of "internal" interpretational work I outlined earlier; one could argue that once it has been specified what the mathematical structures represent, that will determine when two mathematical structures represent the *same* thing. Wilhelm's contribution to this symposium makes just such a claim, as do Coffey (2014) and Teitel (2021)—the latter of whom, incidentally, also describes interpretations as "mappings from representational vehicles to contents" (4125).

Now, it's surely true in some sense that an interpretation consists of a mapping from representations to contents. And it is a tempting idealization to suppose that we have a box of representations, on the one hand, and an array of contents, on the other, and the business of interpretation is a matter of correlating the one to the other—like a child with a sticker-book (Price 2011) or a museum curator appending labels to the exhibits (Quine 1969). However, I'm a bit concerned about this picture, for two reasons.

First, it suggests that *any* mapping from vehicles to contents counts—at least in principle—as an interpretation.[5] Indeed, Teitel explicitly argues that we need to take account of "trivial semantic conventionality": the "familiar platitude that any representational vehicle can in principle be used to represent the world as being just about any way whatsoever," that is, that any association between vehicles and contents is an admissible interpretation. Of course, we're free to give the term *interpretation* a wide scope like this. But I think it is a mistake to abstract away so far from the kinds of interpretations *we* could give. The vast, vast majority of such interpretations are not, in any relevant sense, available to us. Only those interpretations—those mappings from representations to contents—that admit of specification by finite means are the sorts of interpretations that we could, in fact, articulate. One might say that this is why trivial semantic conventionality says merely that we have such interpretational lassitude *in principle*. But this doesn't seem right: it's not for lack of time, or resources, or ingenuity that our capacities to specify interpretations are so circumscribed. Make those as generous as you wish, and we will still only be able to articulate an infinitesimal fraction of the possible associations between words and contents. To consider our capacities "in principle" is to suppose those capacities to be arbitrarily large; it is not to suppose them infinite.

Second, it seems to imply the wrong direction of explanation. In this picture, what makes something an interpretation is that it is such a mapping. So to interpret a theory is just to "give" such a mapping: to specify what propositions correspond to what sentences or, more generally, what contents correspond to what representational vehicles. This, I think, is misleading because it suggests that when we interpret a theory, we put it into contact with some realm of semantic objects of which we

---

[5] North (this symposium) discusses a concern like this in more detail.

already have a grasp. Now, in some cases the practice of interpretation may involve something like this. In translating a theory from a foreign language, for example, we might indicate what terms in the other language correspond to what terms in the home language—and hence, assuming the home language is understood, what the semantic content of the foreign terms is. But I submit that the sense of interpretation we are interested in as philosophers of science simply is not this kind of thing. We do not start out with a grasp of those propositions the theory of general relativity might be trying to say, then interpret that theory by putting its sentences into correspondence with those propositions. Rather, it is through the articulation and application of general relativity itself that we come to be in a position to articulate the propositions that the sentences of general relativity express.

What can we replace this picture with, then? Unfortunately, I do not have a good answer to this question. However, I do want to suggest that the essence of interpreting a theory lies in making sense of the use and application of that theory—in other words, in characterizing its empirical content. Indeed, I am minded to say that so far as interpreting a theory on its own goes, specifying the empirical content is *all* there is to do. On the face of it, one might worry that this is inconsistent with realism, but I think that this worry is misplaced. We need to distinguish two things. On the one hand, there is the claim that all there is to the *content* of a theory is its empirical content: that a theory "says nothing more" than the set of its empirical consequences. That is indeed a strong (and likely unworkable) form of empiricism. However, one can deny this claim—and so endorse the basic realist commitment to theoretical content beyond empirical content—without thinking that there is anything more to the *activity* of interpretation than the specification of empirical content, with the theoretical content then following automatically in its wake. If, following such a specification, somebody says, "Well, that's all well and good, but I don't only want to know how to interpret the empirical part of the theory; I want you to also tell me what the theoretical part is saying," we have no choice but to simply *repeat* the theoretical part itself.

However, there is an important missing piece here. Now suppose that somebody proffers a different theory, to which they have assigned the *same* empirical content as that which has been associated with our theory—so, in other words, the two theories are empirically equivalent. As already noted, we are not identifying the content of the theory with its empirical content, so we are not immediately forced to conclude that these two theories have the same content. However, we have not said anything that gives us the capacity to determine whether these two theories do, in fact, have the same content.

In other words, what is needed is appropriate *criteria of equivalence*. This is what I meant when I said that the specification of empirical content is all there is to interpretation when we are considering a theory *on its own*. When we consider a theory in relation to other theories, there is further work to be done, and that work consists of the specification of equivalence criteria. Again, this is a question for which I have previously tended to prefer a particular answer: by and large, the answer that we should try to adopt liberal criteria of equivalence, such as intertranslatability or categorical equivalence.

However, this stance gives rise to the following dialectical problem. Many of those tempted by liberal criteria of equivalence are attracted by something like the

following thought: there should not be questions that possess a definite answer but where that answer could not be determined, even in principle, by empirical inquiry. (In a slogan: no disagreement without the possibility of empirical resolution.) Yet the question "Are these two theories equivalent?" does not appear to be one that could be settled by empirical inquiry. Certainly, if it were to be maintained that two intertranslatable theories were distinct, then it does not seem that we could point to empirical evidence that would refute their position.

Of course, this difficulty is familiar: it recalls the problem that the principle of verification—that any meaningful assertion must be empirically verifiable—does not seem to itself be verifiable. Carnap (1936) famously suggests escaping this difficulty by denying that the principle in question *is* an assertion: "it is preferable to formulate the principle of empiricism not in the form of an assertion … but rather in the form of a proposal or requirement" (33). The principle defines what it is for a language to be an empiricist language and indicates the user's belief that such a language is more scientifically appropriate than the alternative. More generally, Carnap's *principle of tolerance* holds that "it is not our business to set up prohibitions, but to arrive at conventions" (Carnap 1937, p. 51); what convention to use will be determined by pragmatic, not factual, considerations.

By analogy, we can take the same conventionalist or tolerant attitude toward intertheoretic equivalence (just as we earlier took a conventionalist attitude toward intratheoretic equivalence). Claims of theoretical equivalence, then, are to be understood as recommendations rather than reports. The relevant claim is that our scientific purposes are better served by regarding intertranslatable or categorically equivalent theories as equivalent than by regarding such theories as distinct.

3 Conventions about conventions

However, all this raises difficulties. I am now advocating that the decision to adopt liberal standards of equivalence—that is, to regard the differences between intertranslatable theories as merely notational or conventional—is itself a convention. As already discussed, one reason for doing this is by appeal to Carnap's principle of tolerance. But one might worry that adopting the principle of tolerance might itself *already* commit one to a more liberal standard of equivalence because, after all, doesn't the principle of tolerance argue that when the choice between two theories may be regarded as a convention, it should be so regarded?

In other words, we seem to have a tension between two "levels" at which tolerance might be applied. At the level of comparing theories, the principle of tolerance seemingly instructs us to regard the choice between those theories as conventional—in other words, to regard the two theories as equivalent. But at the level of comparing criteria for theoretical equivalence, the principle seemingly instructs us to regard the choice between liberal and illiberal criteria as a matter of convention, contradicting the earlier instruction to take the side of the liberal criteria! In other words, if we are tolerant about the choices between theories, that appears to commit us to *intolerance* about the choices between criteria of equivalence. This seems an uncomfortable position.

The key to dissolving this tension lies in looking more carefully at what happens if we do indeed adopt the principle of tolerance at both levels. As has just been

discussed, at the level of comparing theoretical criteria, the principle of tolerance requires regarding the disagreement as merely conventional. This means that we cannot declare advocates of a stricter criterion of theoretical equivalence to be wrong. However, we are permitted to regard them as *unwise*: it is consistent with the principle of tolerance to say that an illiberal criterion of equivalence is pragmatically inferior. From this perspective, the lower-level principle of tolerance—the one that is marshaled in support of liberal criteria of equivalence—amounts to a claim that being tolerant will bring pragmatic benefits; the higher-level principle of tolerance takes these benefits as reasons to adopt the lower-level principle (as a convention).

Indeed, this is generally taken to be Carnap's own view of the matter: the principle of tolerance is merely a proposal, not a factive assertion. Coffa (1991) does make the case to the contrary, although even he admits that Carnap's "official" position "would probably have been to say that the principle is not true but is only a proposal" (314). Goldfarb (1997) makes a convincing case that interpreting Carnap as a "semantic factualist" is not only in tension with Carnap's own writings (especially in the period after *Logical Syntax*) but would also undermine the Carnapian project more generally.[6]

This also gives us the resources to address an objection that might arise on the basis of trivial semantic conventionality. Recall that this is the thesis that any representation can be used to represent any content. A version of trivial semantic conventionality could be invoked to argue that any two representations *may* be regarded as equivalent. And if that's so, then it might seem that the principle of tolerance will insist that they *should* be regarded as equivalent. This will then collapse the contents of all representations into one another—surely a reductio of this kind of view!

However, we can resist the pressure toward collapse if we semantically ascend—that is, if we consider the pragmatic benefits of a tolerant framework rather than an intolerant one (using tolerance at the higher level to explain why it is pragmatic benefits that are the relevant ones to consider). I said earlier that the lower-level principle of tolerance then becomes the observation that being tolerant tends to bring pragmatic benefits. Such benefits might include the fact that a more tolerant framework will permit more inferences (because we can "export" inferences between different theories); the fact, closely related, that we can switch between theoretical methods as the need may arise; and the time saved in not debating which theoretical method is true.

Nevertheless, such benefits are defeasible. If we identify theories too freely, then we may encounter pragmatic disadvantages that outweigh these considerations. I take it to be close to self-evident that there are pragmatic disadvantages to identifying theories that are not empirically equivalent. Even where two theories are empirically equivalent, there might be pragmatic arguments against identifying them. But if two representations are empirically equivalent *and formally intertranslatable*, then it is hard to see what the disadvantages to identifying them might be.[7] So we have reason—pragmatic reason, but reason nonetheless—to draw the line there.

---

[6] Thanks to a referee for pressing me to consider Carnap's own views more carefully.

[7] More subtly, it might be that some of the pragmatic advantages enumerated earlier will only apply when the theories are intertranslatable. For example, it is not clear to me how one might "export" a conclusion from one theory to another without knowing how to express that claim in terms digestible by the second theory.

References

Brading, Katherine, and Elena Castellani, eds. 2003. *Symmetries in Physics: Philosophical Reflections.* Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511535369

Button, Tim, and Sean Walsh. 2018. *Philosophy and Model Theory.* Oxford: Oxford University Press. doi: 10.1093/oso/9780198790396.001.0001

Carnap, Rudolf. 1936. "Testability and Meaning." *Philosophy of Science* 3 (4):419–71. doi: 10.1086/286432

Carnap, Rudolf. 1937. *Logical Syntax of Language.* London: Routledge. doi: 10.4324/9781315823010

Coffa, Jose Alberto. 1991. *The Semantic Tradition from Kant to Carnap: To the Vienna Station.* Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139172240

Coffey, Kevin. 2014. "Theoretical Equivalence as Interpretative Equivalence." *British Journal for the Philosophy of Science* 65 (4):821–44. doi: 10.1093/bjps/axt034

De Haro, Sebastian, and Jeremy Butterfield. 2018. "A Schema for Duality, Illustrated by Bosonization." In *Foundations of Mathematics and Physics One Century after Hilbert: New Perspectives*, edited by Joseph Kouneiher, 305–76. Cham, Switzerland: Springer. doi: 10.1007/978-3-319-64813-2_12

Gödel, Kurt. 1995. *Collected Works: Volume III: Unpublished Essays and Lectures*, edited by Solomon Feferman, John W. Dawson Jr., Warren Goldfarb, Charles Parsons, and Robert N. Solovay. Oxford: Oxford University Press.

Goldfarb, Warren. 1997. "Semantics in Carnap: A Rejoinder to Alberto Coffa." *Philosophical Topics* 25 (2):51–66. doi: 10.5840/philtopics19972529

Marschall, Benjamin. (2021). "Carnap and the Ontology of Mathematics." PhD diss., University of Cambridge.

Maudlin, Tim. 2018. "Ontological Clarity via Canonical Presentation: Electromagnetism and the Aharonov–Bohm Effect." *Entropy* 20 (6):465. doi: 10.3390/e20060465

Price, Huw. 2011. *Naturalism Without Mirrors.* Oxford: Oxford University Press.

Putnam, Hilary. 1962. "The Analytic and the Synthetic." In *Scientific Explanation, Space and Time*, edited by Herbert Feigl and Grover Maxwell, 358–97. Minneapolis: University of Minnesota Press.

Quine, Willard Van Orman. 1936. "Truth by Convention." In *Philosophical Essays for Alfred North Whitehead*, edited by Otis H. Lee, 90–124. London: Longmans, Green & Co.

Quine, Willard Van Orman. 1951. "Two Dogmas of Empiricism." *Philosophical Review* 60 (1):20–43. doi: 10.2307/2181906

Quine, Willard Van Orman. 1969. "Ontological Relativity." *In Ontological Relativity and Other Essays, 26–68.* New York: Columbia University Press. doi: 10.7312/quin92204

Sklar, Lawrence. 1982. "Saving the Noumena." *Philosophical Topics* 13 (1):89–110. doi: 10.5840/philtopics19821315

Teitel, Trevor. 2021. "What Theoretical Equivalence Could Not Be." *Philosophical Studies* 178 (12):4119–49. doi: 10.1007/s11098-021-01639-8

van Fraassen, Bas C. 2014. "One or Two Gentle Remarks about Hans Halvorson's Critique of the Semantic View." *Philosophy of Science* 81 (2):276–83. doi: 10.1086/675645

Warren, Jared. 2020. *Shadows of Syntax: Revitalizing Logical and Mathematical Conventionalism.* Oxford: Oxford University Press. doi: 10.1093/oso/9780190086152.001.0001