Book Review

The Logic in Philosophy of Science, by Hans Halvorson. Cambridge: Cambridge University Press, 2019. Pp. viii + 296.

In 2012, Hans Halvorson published an important and widely discussed article entitled 'What Scientific Theories Could Not Be' in *Philosophy of Science*. It was a broadside against the 'semantic view of theories', as developed by philosophers such as Patrick Suppes, Bas van Fraassen, Frederick Suppe, and Lisa Lloyd. The semantic view was a response to what those authors called the 'received view', or sometimes the 'syntactic view of theories', as developed by mid-century Logical Empiricists. Roughly speaking, according to the semantic view, a scientific theory can be identified with a collection of models. The key claim of Halvorson's paper was that this view cannot be correct, because it does not support an adequate account of when two theories are equivalent, that is, when two putatively distinct theories are nonetheless 'the same'. To defend this claim, he presented three candidate criteria of equivalence available to the defender of the semantic view, and then ' showed that these criteria render clearly inadequate verdicts for various simple examples.

Halvorson's paper has spawned several new literatures. One such literature has developed a program esquissed in the paper's conclusion, according to which a scientific theory should be understood as a 'structured' set of models. That is, one might try to save something like the semantic view of theories by taking a theory to consist not just in its models, but also in relationships between those models. Work in the past decade that has explored the idea of representing a theory as, for instance, a category of models or as a topological space of models might be seen as steps in this direction.

A second literature – and one which Halvorson himself has engaged with in much more depth – has approached the problem raised in his 2012 paper from the opposite direction. By Halvorson's lights, the semantic view fell because it could not adequately individuate theories. If one wants an account of theories that can succeed where the semantic view failed, then one should first explore the range of plausible and compelling criteria of equivalence available for theories, particularly in first-order logic; and then one may use that inquiry to help answer questions about what the structure of a theory

must be, given that it is faithfully preserved by certain transformations. This book represents a kind of capstone to Halvorson's work on this programme over the past decade.

Halvorson does not exactly defend, or even systematically develop, a 'syntactic view' of theories to counter the semantic view that he has criticized. But he does defend the view that mathematical logic, syntax and all, is an essential tool for studying theories—including, but not only, scientific theories. And he shows, through a systematic development of metalogic, that there is a deep link between logical syntax and the mathematical theory of semantics. On the one hand, he argues, logical semantics is *itself* a mathematical theory, and thus, contra many defenders of the semantic view, any philosophical problems of scientific representation or world-theory relations that plague the syntactic view arise in just the same way for the semantic view. There are some who will think that the real lesson of the downfall of Logical Empiricism beginning in the 1960s, of which the debate about the syntactic view and the semantic view was just a part, was that mathematical approaches in philosophy of science are misguided. Such readers will not be convinced by what Halvorson writes here. But for those who maintain that the real issue was what mathematical resources are the most fruitful to use, Halvorson's book is deeply enlightening and rewarding.

Much of the book is devoted to technical results in (first-order) logic. After a brief Introduction, with an impressionistic history of the ways in which technical results in logic during the late nineteenth and twentieth centuries have influenced major turns in analytic philosophy during the past century, Halvorson proceeds to offer seven chapters – about 85% of the text – developing mathematical aspects of metatheory, with an emphasis on results and ideas that have been influential in philosophy of science. It is not quite right to say that the treatment is idiosyncratic, because it reflects how a significant number of contemporary mathematicians think about this subject; yet still, the perspective is likely one that philosophers who are not independently familiar with the mathematical logic literature will not recognize. It is also selective in the topics presented: for instance, while Beth's theorem and Svenonius's theorem are given detailed treatments, and Łos's theorem on ultraproducts is stated, Gödel's incompleteness theorems are not mentioned, much less proved. This is entirely appropriate, given Halvorson's goals, but it means that the book is probably most naturally used as a supplemental text for a graduate-level logic course or a source for further reading after such a course, and not as the primary text.

After a standard overview of propositional logic in chapter 1, Halvorson proceeds in chapter 2 to introduce the Elementary Theory of the Category of Sets. This chapter serves as both a quick introduction to basic ideas in Category Theory, which remains a gap in many philosophers' training; and a translation of ideas from set theory that may be familiar to technically-minded philosophers into a possibly unfamiliar language. The treatment is

somewhat terse and may be heavy going for someone encountering the subject for the first time, but it is also complete and pedagogically oriented, and suitable for a reader who has taken graduate-level logic in a philosophy department. Chapter 3, meanwhile, builds on these two chapters by first showing how propositional theories may be associated with Boolean algebras (and vice versa); and then stating and proving the celebrated Stone Duality Theorem, establishing a kind of equivalence (namely, categorical duality) between the category of Boolean algebras and the category of Stone spaces. Although this result is well known in some circles, it is both mathematically deep and deserving of broader appreciation within philosophy. Halvorson's treatment is clear and enlightening.

From here, Halvorson moves from propositional logic to full first-order logic. Chapter 4 introduces what he calls 'syntactic metalogic'—a subject that in other settings might be referred to simply as 'metalogic', though in this case it is important for Halvorson's purposes that much of what he will do can be re-done from a 'semantic' perspective. (The upshot will turn out to be that syntactic and semantic aspects of metalogic are both essential to a full mathematical treatment of theories; and that it is particularly important to appreciate the relationships between them, as emblemized by the Stone Duality Theorem for propositional theories.) Much of this chapter covers standard material such as grammar and deduction rules. But Halvorson also covers material that may be passed over quickly in a standard course, such as reconstruals, translations, extensions, and definitions, ultimately leading to a discussion of 'definitional equivalence' as a candidate criterion for the equivalence of theories. Here we see how Halvorson's emphases reflect the origins of the book in the debate over whether the semantic view of theories can support an adequate notion of equivalence of theories.

Chapter 6 revisits much of the material in chapter 4, now from a 'semantic' perspective—that is, through the lens of model theory. Halvorson states and proves several classic results – soundness and completeness for first-order logic; the downward Löwenheim-Skolem theorem; and so on – and then proceeds to define the category of models associated with a first-order theory, which he uses to analyse notions such as translation and conservativity from a more semantic perspective. The chapter concludes with a detailed and very enlightening discussion of different notions of implicit definability, including a discussion of the relationship between Beth's theorem and Svenonius's theorem, both of which concern senses in which implicit definability implies explicit definability in first-order theories, though with important subtleties that are often overlooked.

One thing I wish Halvorson had included with his discussion of implicit definability is some commentary on the status of definability in second-order theories. Beth's theorem is known to fail in that context, and thus it was not clear to me how much of Halvorson's discussion should be understood to carry over to mathematical theories 'in the wild', which are often 'overtly

second-order' (as he puts it on p. 94). Halvorson deals with this issue briefly in chapter 4, arguing that second-order theories can be formalized in set theory, which is a first-order theory; but it was not clear to me how much that fact helps with interpreting particular theorems that are asked to carry significant philosophical weight, and which are known to fail for second-order theories. (the Löwenheim-Skolem theorem is another example that comes to mind.) Halvorson briefly discusses second-order logic in his discussion of Ramsey sentences in chapter 8, but does not return to reflect on the questions I have just raised. In any case, given the emphasis on translation, interpretability, equivalence, and definability in the book, it might have been helpful to discuss the sense of interpretation at issue when one formalizes a second-order theory within (first-order) set theory, and to ask how this bears on the generality of the philosophical morals based on first-order logic.

Chapters 5 and 7 extend the material in chapters 4 and 6, respectively, to many-sorted logic. The centrepiece of these chapters is Halvorson's discussion of 'Morita equivalence' (also sometimes called 'generalized definitional equivalence'), which is a generalization of definitional equivalence based on a notion of translation between theories with different sorts. Halvorson has done important work on Morita equivalence, mostly in collaboration with Thomas Barrett, and this book contains the most complete and up-to-date discussion of the topic. As Halvorson observes, philosophical treatments of first-order logic almost exclusively focus on the single-sorted case; he attributes this to the fact that Quine famously argued that many-sorted logic can be 'reduced' to single-sorted logic. But Halvorson believes this is problematic, both because the sense in which many-sorted logic may be reduced to first-order logic can only be made precise using Morita equivalence, and because Quine's inference from 'every multi-sorted theory is equivalent to a single-sorted theory', which is true even though Quine did not quite establish it, to 'multi-sorted logic is dispensable' is tendentious. I tend to agree with Halvorson, and I think the examples he adduces – such as the (Morita but not definitional) equivalence of geometry formulated using points and geometry formulated using lines – show the power and importance of multi-sorted logic *even for understanding single-sorted theories.*

Although chapters 1 through 7 are mostly expositions of technical material in logic, Halvorson sometimes pauses to connect what he is doing with more philosophical themes. For instance, in section 4.4, he discusses empirical equivalence in the context of the Logical Empiricist's program; section 5.3 introduces and evaluates Quine's arguments on many-sorted logic; section 5.4 includes an extended example addressing 'quantifier variance', which is the view that certain arguments over whether mereological composites 'exist', above and beyond whether their parts exist, are purely 'verbal'; and section 5.5 discusses symmetries, as inspired by discussions of that topic in recent philosophy of physics. One particularly important point that Halvorson takes pains to make comes in a 'Philosophical Moral' announced

on p. 174, at the end of section 6.3, which is that a Sigma-structure, for some first-order signature Sigma, is not a 'set-theoretic structure', much less something that can stand in some relationship of 'isomorphism' or 'partial isomorphism' with the 'world'. Here he sharpens and extends his influential critique of the semantic view of theories, discussed above.

These intermezzos are valuable and insightful gems, though sometimes difficult to find within the text. But the most sustained engagement with philosophical arguments in the book comes in chapter 8, where Halvorson devotes several pages each to a range of major debates in twentieth and early twenty-first century analytic philosophy, from mental state functionalism to Putnam's model-theoretic argument to scientific realism. Several original and important arguments appear here. For instance, Halvorson mounts a sustained critique of Ramsification, whether in the context of functionalism, structural realism, or the meaning of theoretical terms. As he argues, it is not clear what is gained by moving from a theory to its Ramsey sentence, and he diagnoses 'the impulse to Ramsify' as 'no other than the . . . impulse to use uninterpreted mathematical symbols to represent physical reality', which he finds wrong-headed (p. 252).

He also mounts a compelling critique of philosophers, especially philosophers of physics, who worry about 'counting possibilities', in the sense of trying to identify whether distinct set theoretic structures that one identifies as 'models' of a physical theory do or do not represent distinct 'possible worlds'. He argues that such distinctions are generally not invariant under equivalence of the category of those models. One might object that defining that category involves specifying a class of transformations as 'symmetries' or, more generally, isomorphisms of models, which is simply another way of putting what is at issue in these debates about counting possibilities. But as he remarks, this is an important shift in perspective, and moreover, properly to understand 'symmetry' in the sense in which it is often invoked in such discussions, one needs to pay careful attention to an object language / metalanguage distinction that is completely lost when one considers only whether certain models should count as 'the same' or 'different'.

But perhaps the most significant argument of the book, and one of the only ones that runs consistently throughout the entire manuscript (as opposed to appearing as an application of some idea or other), is that one can gain insight into classical debates about realism and anti-realism, both scientific and metaphysical, by recasting them as debates about what criteria of equivalence to adopt for theories. (Thomas Barrett previously made a similar point about scientific theories in his dissertation, but Halvorson pushes it much further.) So, on this view, realists are people who adopt very strong criteria of equivalence, and thereby take theories that disagree in 'minor' ways to make different assertions about the world; whereas anti-realists are those who claim that very different-seeming theories are nonetheless equivalent.

As an example, he attributes to the anti-realist Putnam the view that *all consistent theories are equivalent*, as a way of recovering the moral of the famous Löwenheim-Skolem model-theoretic argument; whereas some realists – for instance, he notes Ted Sider – might claim that there is a privileged language in which the true theory of the world must be expressed, and thereby maintain that no theory not in this language, or perhaps even no theory without the correct axioms in this language, could be equivalent to the true theory. For his own part, Halvorson seems to prefer a criterion of equivalence somewhere in the middle – perhaps Morita equivalence – and thus, it seems to follow, adopts a position intermediate between radical metaphysical anti-realism and radical realism.

I think this is a valuable idea, and that it mirrors an important idea in mathematics (reflected throughout the book) that a fruitful way of studying the structure of something is to look at what transformations 'preserve' that structure. Thus, if one wants to understand what kind of structure one wishes to attribute to the world, understanding the transformations that generate equivalent descriptions of the world is a natural way to proceed. But I worry that something is missing from Halvorson's account. In particular, he directly associates views on equivalence with views on realism only in extreme cases. But he does not say much about what form of realism might be most naturally associated with, say, definitional equivalence, Morita equivalence, or categorical equivalence, all of which he seems to take to be more plausible (and moderate) criteria. And thus, although it seems that certain views about realism / anti-realism support certain criteria of theoretical equivalence, it is not clear how to go in the other direction, or even what is at stake for realism debates in the choice between the different formal criteria of equivalence at hand. In the end, I wondered whether refocusing the debate in this way would actually change anything for philosophers concerned with realism and anti-realism—or whether, rather, debates about equivalence would ultimately reduce to whether one wants to be a realist about things that are not preserved under certain equivalence relations.

There was also an important issue that I worried was left behind in this discussion, and also elsewhere in the book, concerning how equivalence is supposed to be related to semantics and pragmatics—semantics, here, not in the sense of logical semantics, but rather construed as the interpretation of our theories as assertions about the world. One might have thought that two theories could be equivalent only if they were not empirically distinguishable. One might respond that if two theories are equivalent in some appropriate logical sense, and one of them makes (only) empirically true statements, then there is *ipso facto* a world-theory semantics that makes the second theory also make (only) empirically true statements. But it does not follow that this interpretation of the second theory would be adequate on other grounds, for instance because it did not reflect the intended interpretation of that theory as understood by its advocates. On the other hand, if one insists

that equivalence of theories somehow respect intended interpretation, any relationship between realism and criteria of equivalence would need to take a detour through an account of how theories represent the world. I suspect Halvorson would not care about this concern, but I would very much have liked to see him address it and explain why he does not care.

Indeed, although Halvorson is very effective at arguing that the semantic view of theories does *not* solve the problem of theory-world relations, and at times seems to imply that a more syntactic approach to theories is superior in this regard, he does not attempt to articulate an account of how theories represent the world. On the one hand, this is hardly a criticism: theory-world relations, and more generally word-world relations, are a huge and difficult subject that would extend far beyond the scope of this book. On the other hand, I often found myself wondering how to interpret Halvorson's arguments about the interpretation of logical calculi, the semantic view of theories, and the significance of criteria of equivalence for metaphysics without some hint of how he took theories to say anything about the world.

Department of Logic and Philosophy of Science, JAMES OWEN WEATHERALL
University of California, Irvine
weatherj@uci.edu
doi: 10.1093/mind/fzaa020

