Metaphysical Equivalence

How formal concepts of theory equivalence inform metaphysical debates.

by

Timo Schobinger

Submitted in partial fulfilment to the requirements for the degree of Masters of Philosophical Studies (MPhil Stud)

to the

Department of Philosophy
UCL

Declaration

I, Timo Schobinger, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Date: 14/3/2025

Abstract

Concepts of theoretical equivalence aim to capture what it means that theories “say the same thing” or “have the same” content. In the context of formal theories, there are precise formal concepts that explicate ways in which formalized theories are equivalent. Based on notions of definition and translation, these concept formulate what it means that the structure of a theory is preserved in another, either in its entirety or in parts. These concepts vary in strength and differ on which parts of the theoretical structure is understood as invariant between equivalent theories.

This thesis investigates how formal concepts of theoretical equivalence can be used to investigate disagreements in metaphysics. The thesis does three things. It first motivates the role for a concept of metaphysical equivalence, i.e. a concept of theory equivalence applicable to metaphysics. Second, it presents formal notions of theory equivalence that characterise relations between formal theories. Third, it argues that formal equivalence constrains the metaphysical commitments of theories – realist interpretation cannot rely on the assumption that theories have a metaphysically privileged linguistic expression. Formal equivalence relation here are the basis for explicating a realist understanding of the language independence of ontological commitments.

Impact Statement

Concepts of theoretical equivalence are central for understanding in which sense formulations of mathematical or scientific theories express the same theory. Concepts of theoretical equivalence are developed in the context of formalized theories and sharpen our understanding of mathematical and scientific theories and practices.

Investigating the relation between formal concepts of theory equivalence and metaphysical considerations deepens our understanding of both theories and relations between theories. In particular, it advances our understanding of how theories and their metaphysical interpretation depend on the languages and other formal frameworks used to express the theories. The perspective taken in this research connects formal concepts with the metaphysical interpretation of scientific, mathematical, and philosophical theories. It thereby advances both our philosophical understanding of formal concepts of equivalence and of how metaphysics has to understand theories that are formulated in different languages.

Concepts of metaphysical equivalence allow for a better understanding of the distinction between those disagreements in metaphysics that are "purely verbal" and those that are substantive. While this discussion arises in the context of metaphysics, it formulates the fundamental question what it is to disagree with another if each party to the disagreement uses their own language or theoretical framework. While the distinction between purely verbal and substantive disagreements has been a central issue in contemporary (meta-)metaphysics, there are few attempts to consider different candidate notions of metaphysical equivalence on the basis of concepts of formal theory equivalence. This research takes into account both, formal concepts and metaphysical perspectives.

Contents

1 Philosophical Motivation and Methodology

1.1 Introduction

This chapter is a philosophical introduction to the study of metaphysically relevant equivalence relations between theories. This introduction has two roles – to motivate the project pursued and to sketch its methodological framework.

Concerning the first purpose, this introduction establishes that philosophical debates rely on some conception of theoretical equivalence. This holds particularly in the case of metaphysics, where the alleged equivalence of particular proposals gave raise to the question whether at least some metaphysical disagreements are at all substantive. I argue that philosophical discussions, in particular in metaphysics, rely on some concept of theoretical equivalence by emphasizing the role of this concept for the distinction between substantive and "purely verbal" disagreements. For every disagreement one might ask whether it is purely verbal; however, in philosophy, the notion of verbal disagreement is discussed in the context of contemporary (meta-)metaphysics and the philosophy of science.

In contemporary metametaphysics, deflationary responses are developed on the claim that some disputes in metaphysics are purely. I argue that we cannot understand the distinction between substantive and merely verbal disagreements without reference to some concept of theoretical equivalence.

A corresponding question is discussed in the philosophy of science concerning the metaphysical interpretation of scientific theories. The disagreement between scientific realists and anti-realists concerns the limits to metaphysical interpretations of scientific theories – which differences in the theoretical posits of competing theories ought to be understood as expressing competing descriptions of the world. If theories are equivalent, given a notion of theoretical equivalence adequate to this context, their potential differences are to be understood as not corresponding to a difference in the target system of the theories in question. Concept of theoretical equivalence describe limits to the metaphysical interpretation of scientific theories. In giving an answer to what is to be preserved between equivalent theories, particular equivalence concepts correspond to positions in the debate between realists and anti-realists (Halvorson, 2019, sect. 8.4).

Considering questions concerning the “good” and “bad” (or maybe “explanatory” and “idle”) parts of theories generalizes to other discussions concerning metaphysics. Everyone who presents some kind of metaphysical (or scientific) theory, from empiricists to strong realists in the line of Sider (2011), relies on a concept of equivalence. Equivalence concepts distinguish features of theories that are important, theoretically or in a metaphysical interpretation, from those features that ought to be understood as accidents of the presentation of theories. *Any* disagreement ultimately relies on an understanding of which features of a theories can make a difference to their content and therefore on an understanding of their equivalence conditions. In this thesis, I will focus on the constraints on the metaphysical interpretation of theories imposed by formal concepts of theory equivalence.

The second goal in this chapter is to sketch the general methodology for this project. I introduce the idea that meta-theory, i.e. treating theories as objects of investigation, can be used to elucidate philosophical (first-order) debates. Based on this general approach, I discuss its application in this thesis to metaphysics. Here, I aim to establish the scope of the investigation, and to outline preliminary assumptions about theories and the meta-theoretical tools used to investigate them. Additionally, I provide an outlook on the general limitations of the metatheoretic approach and how this impacts the results of this thesis. I return to questions concerning the scope of the investigation and the interpretation of its results throughout the thesis.

1.2 Comparing theories

This section focuses on two phenomena that motivate the investigation of theories as objects of comparison. The first observation is that disagreements can be “purely verbal”. Disagreements of this kind arise if participants in a debate interpret central theoretical vocabulary in different ways; they “talk past each other”. Getting a better understanding of this phenomenon is particularly pressing for discussions in metaphysics, as even the interpretation of quantifiers may not be shared (see for example Hirsch, 2002). Second, competing theories, in particular in metaphysics, often are not even superficially formulated in a shared language. These theories might not mutually inconsistent in a narrow sense: there is no sentence entailed by one theory such that the competing theory entails the negation of that sentence. This raises the need for a way to compare theories that are expressed using different languages, for example by investigating translations between the sentences of competing theories.

A list of potentially equivalent theories

Here is a list of cases in which the equivalence of theories are asserted:

- Notational variants of theories that systematically substitute the symbols for terms or logical connectives (e.g. systematically replacing $\wedge$ with & or swapping every occurrence of $\vee$ for $\rightarrow$ and vice versa) (French, 2019).
- Theories that are formulated using different sets of logical connectives (but see McSweeney, 2019).
- First-order Peano-Arithmetic formulated in the signatures $\Sigma = \langle 0, s, +, \cdot \rangle$ and in $\Sigma_< = \langle 0, s, +, \cdot, < \rangle$, i.e. with and without an ordering predicate such as $<$ in the language; $<$ is definable in terms of addition $m < n$ iff$_{\text{def}}$ $\exists x(x \neq 0 \wedge m + x = n)$.
- Peano-Arithmetic in relational formulation with the signature $\Sigma_R = \langle Z, S, A, M \rangle$ and the usual functional formulation in $\Sigma = \langle 0, s, +, \cdot \rangle$. The relational formulation has no constants or function terms in its signature. Instead, its signature has a predicate Z for "is zero", a two-place predicate for successor relation S, and three place predicates A and M read as "z is the sum of x and y" and "z is the product of x and y". Additional axioms guarantee the existence and uniqueness for these predicates that are required for a definitional equivalence of the formulations.
- Perdurantism and endurantism as theories about how concrete objects persist over time (cf. Hirsch, 2002).
- Systems of anti-Humean modal metaphysics with different modal primitives ("laws of nature", "dispositions", "potentialities", "forces", "necessity") might agree on the description of modal reality, e.g. by proving the axioms of the same formal system of modal logic.[1]

[1] See for example Vetter, 2015, appendix A for showing that a modal logic based on a potentiality operator proves axioms of the system $S4$.

- Mereological universalism and mereological nihilism (chapter 3, cf. Warren, 2015).

- Some theories of geometry (in particular: of Euclidean geometry) can be formulated in different ways, e.g. in terms of only points, only lines, and both points and lines (chapter 4, see in particular Schwabhäuser et al., 1983, (Propositions 4.59, 4.89), Barrett and Halvorson, 2017a).

- Different formulations (Newtonian, Hamiltonian, and Lagrangian) of classical dynamics (chapter 5, see Barrett, 2019, Coffey, 2014, Kaveh, 2023).

None of the examples is entirely uncontroversial. As I will discuss throughout the thesis, for each of the examples, the theories are equivalent in some explications of "equivalent", but not in others. The central question then is whether there is a privileged concept of theoretical equivalence that is generally adequate for formalized theories. A privileged concept of this sort could then be used as concept for the identity of theories across languages (see for example the proposal in Szczerba (1977, pp. 128, 135) to understand theories as closed not only under logical consequence, but also under "the rules of definition", see also section 2.3.1).

Moreover, there are questions on whether the specific formal equivalence relations that hold between these theories ought to guide their metaphysical interpretation. If one holds that different formulations of classical logic are equivalent for the purpose of mathematics, this might not satisfy every metaphysician (cf. McSweeney, 2019). Strong realists about logic will challenge the idea that different presentations of classical logic are metaphysically equivalent. Their assertion of a (fundamental) logical structure of the world may require that some set of connectives is privileged as it matches this fundamental structure. In this thesis, I argue that realists need to account for much weaker formal equivalence relations, which put constraints on what a realist interpretation of theories can achieve.

I will return to these examples throughout this thesis to motivate questions concerning concepts of theoretical equivalence and the metaphysical assumptions that are connected to particular equivalence concepts. However, I cannot discuss most of the examples in detail.

The exceptions here are the cases of mereology, geometry, and classical dynamics. I cover mereology in chapter 3, as example for a precise statement of an equivalence claim and its limitations in a restricted settings. In the discussion of geometry in chapter 4, I then focus on the consequences of a multi-sorted framework. My focus in both cases will be on how concepts of theoretical equivalence impact our understanding of ontological commitments of theories as language independent. In chapter 5, I present how realists use fundamentality considerations to challenge the adequacy of formal equivalence concepts for interpretative purposes (cf. Coffey, 2014). Here, I argue that if two theories are formally equivalent to the degree that they express (translations of) each others theorems, realists cannot employ purely metaphysical considerations to assert that one of them is more fundamental.

1.2.1 Purely verbal disagreements

In this subsection, I argue that it is often important to distinguish between substantive and purely verbal disagreements. I then argue that this distinction presupposes a notion of theoretical equivalence. The suspicion that purely verbal disagreements arise in metaphysics therefore motivates a concern how directly thinking about theoretical equivalence is useful in metaphysics.

In disagreements, it is central to figure out when people say the same thing or talk past each other. Disagreements can result not only from substantive differences between the positions expressed, but also from differences in the use of language by each party to express their position. In the most pointed case, the case of "purely verbal" disagreements, people could agree entirely about the substance of their views, but would not notice this agreement as their statements appear to be mutually exclusive.[2] To motivate the investigation of concepts of metaphysical equivalence, I focus here on deflationary strategies that analyse some metaphysical disputes as purely verbal and thereby not substantive. There are other ways in which disagreements might not be substantive. A discourse might not be truth-apt – the truth-values of its sentences might be relative to a context, e.g. to the speaker in the case of disagreements about taste, or its sentences might not have a meaning, as they are stated external to a linguistic framework that gives meaning to its terms (see for example de Sa, 2007, 2015; Carnap, 1950). This thesis is concerned with comparisons between theories formulated in different languages and therefore touches on the internal-external distinction, but otherwise will not be concerned with other ways in which disagreements in metaphysics might not be substantive. The distinction between substantive and "purely verbal" disagreements has important consequences for how we ought to understand the purpose of a debate and how to adequately respond to competing

[2] I will return to the case in which there is both a substantive disagreement and people talking past each other in the next section.

proposals. When disagreements are substantial, the participants of a debate in fact have different views about what is at stake. They address the same question, but give mutually exclusive answers.[3] Substantive disagreements warrant theoretical engagement with the question and supporting one's answer by evidence pertinent to the question. By presenting evidence and other epistemic considerations, a substantive theoretical disagreement can be resolved by a justified adoption of a theory over competing proposals.

In the case of purely verbal disagreements, the situation is different. As the disagreement concerns not primarily the content of the view, the task is to understand the difference in language use, to argue for a particular interpretation of terms, and to dissolve the disagreement in the course of doing so. Purely verbal disagreements can be resolved by developing a shared understanding of the terms involved and finding a way of expressing one's theory that is acceptable for all parties involved or, instead, allows for a formulation of a substantive disagreement. The focus changes from a theoretical question to the practical question which language to adopt for one's purpose – even if this purpose is finding an acceptable way of expressing one's theory.[4][5][6]

Some contemporary metaphysicians suggest that some metaphysical debates involve purely verbal disagreements. Hirsch (2002) holds this view about theories of persistence: perdurantists and endurantists appear to disagree on whether objects have temporal parts. But this disagreement, so Hirsch, results from their respective use of metaphysical vocabulary such as "exists" or "object". Moreover, he claims that both parties to the disagreement would accept the theory of their counterpart in the debate if that theory is translated into their language. Hirsch therefore claims that perdurantists and endurantists agree on the substance of their theory, but not about the language used to formulate that theory. Their disagreement is therefore metaphysically shallow and can be resolved by adopting either the perdurantist's or

[3] For sake of exposition, this simplifies disagreements as involving inconsistency between theories, but there there might also be substantive disagreements if a theory is a proper subtheory of a competing theory.

[4] For the idea of external or practical question in metaphysics and deciding between languages for practical considerations see (Carnap, 1950; Hirsch, 2002).

[5] While purely verbal disagreements are not substantive with respect to metaphysics, they can still be important: in many cases, there are practical and maybe even moral reason to adopt a way of speaking or to make revisions in the understanding of terms in a way that serves our purposes. In particular, the meaning of terms used to describe social ontology itself is politically consequential (cf. Haslanger, 2000). Discussions that are framed as *conceptual engineering* makes this this understanding of a type of (philosophical) disagreement explicit. Here, the disagreement in place might be understood as substantially concerned with the choice of words.

[6] For metaphysical debates as metalinguistic negotiation see Belleri, 2017; Sambrotta, 2019.

the endurantist's language on the basis of practical considerations.[7]

1.2.2 Disagreement and shared language

In the following, I argue that disputes concerning metaphysics require the comparison between theories formulated in different languages. In metaphysics in particular, one cannot rely on a shared background theory that would provide the basis of a fixed interpretation of a language in which a disagreement then could be formulated.

Like one may worry that diverging interpretations of metaphysical terms obfuscate substantive agreements, one might also worry whether there can be substantive disagreement in metaphysics in the first place.

As I noted for the case of purely verbal disagreement, substantive disagreement requires that the languages in which theories are formulated are interpreted in a way that leads to an inconsistency of these interpretations. Mutually inconsistent theories require that these theories have consequences that are incompatible, and, presumably, that this incompatibility can be stated in some language.

One might worry that metaphysical theories allow for shared interpretations of their sentences, a necessary condition for mutual inconsistency, only in cases in which they are equivalent. Theories in metaphysics potentially extend to every term of the language, and even to the logic used (cf. Williamson, 2013). Accordingly, every term of the theory will be a theoretical term and its interpretation will not be shared between competing theories

In other words, the worry is that the interpretation of metaphysical terms is only fixed by the entire theory so that non-equivalent theories will involve giving different meaning to metaphysical terms in a way that prevents competing theories from being incompatible.

This worry is particularly pressing for metaphysics, as theories here are often formulated with only a limited vocabulary. The interpretation of this limited vocabulary is not fixed, as even the interpretation of quantifiers are something that is meant to be determined by the theories themselves: Hirsch (2002), for example, argues that the meaning of "exists", "object", and identity predicates can vary between theories. Central concepts of ontology therefore would be not invariant between competing metaphysical theories.[8] To recover a disagreement between theories formulated in

[7] Hirsch (2009, p. 240) prefers endurantism on the basis of claiming that ordinary English is an endurantist language.

[8] See also Carnap's view that the meaning of metaphysical vocabulary is internal to a linguistic

different interpreted languages, one would therefore need to translate between the languages in a way that maps theoretical terms of one theory to corresponding expressions of the other language. The intention behind a translation of this kind is to match expressions to preserve the interpretation of the theoretical terms in a way that allows to formulate the disagreement (or lack thereof).

On the other hand, there remains the question which metaphysical theory to adopt, or at least which linguistic framework to use (Carnap, 1950). If one takes disagreement about metaphysics as substantive, one needs to be able to compare theories formulated in different languages, with potentially diverging interpretations of their central terms. This means that the comparison of these theories cannot simply rely on a holistic understanding of the interpretation of the theoretical terms internal to each theory, but has to investigate the relation between the theories that determine the meaning of their technical terms (cf. Potter, 1998).

If one thinks that this choice between theories or linguistic frameworks is rational, at least in a practical sense (e.g. Carnap, 1963, p. 982), one needs at least some way of comparing theories (or frameworks, respectively). If comparisons between theories are at all possible, it requires a sense in which theories are equivalent or non-equivalent for the purpose at hand. So even if one considers disagreements in metaphysics as practical disagreements about the adoption of a certain language, there is a need to determine when theories are equally well suited for one's purposes. But to do so, one needs to have some way of comparing theories that are formulated in different languages. A shift towards a metalinguistic perspective can be used to formulate the disagreement. A shared metatheoretic language allows one to posit the question whether one should prefer language $\mathcal{L}_1$ over language $\mathcal{L}_2$. But this reformulation of the question still requires an answer (in a metatheory) to the question whether the background theories that fix diverging interpretations of the terms of $\mathcal{L}_1$ and $\mathcal{L}_2$ are competing theories after all.

So there is a general need for a comparison of theories formulated using different languages in a way that cannot rely on a fixed interpretation for most of its vocabulary. Even if we grant a classical interpretation of the propositional logical connectives, the possibility of varying interpretations of quantifiers forces metaphysics to consider whether theories formulated in different languages can be equivalent or be genuinely competing accounts.

framework in Carnap (1950).

1.3 Meta-theory and equivalence relations

Both phenomena introduced in the previous section show the need for some way of comparing theories with respect to their theoretical and metaphysical content – i.e. comparing “what they say about the world”. In this section, I develop this idea further to motivate comparisons between theories from a meta-theoretical perspective. The approach here is to explain why taking theories as objects of an inter-theoretic comparison is a promising way of capturing whether a disagreement is substantive. On that basis, I introduce notions of theoretical equivalence in general and give a preliminary exposition of how this applies to metaphysics.

It seems to be a natural first step to determine how theories, as linguistic or mathematical structures, relate to the content they express.[9] This leads to investigating the semantics of theories in general, and of the theories in question in particular, to determine whether the competing theories describe the same structure. One would therefore ask whether theories “express the same proposition”. Doing so introduces a concept of equivalence directly in terms of an understanding of theoretical content as the proposition expressed by a theory. But this is not as useful as one might hope: it requires a prior understanding of propositions, or the introduction of a concept of propositions as theoretical objects used to investigate relations between theories. While there are explanations of what propositions might be, this remains ultimately a contentious issue in philosophy (contrast, for example, the view that propositions are sets of possible worlds, cf. Stalnaker, 1976, with the view that understands propositions as structured entities King, 2019. See also McGrath and Frank, 2020 and King et al., 2014). In chapter 5, I discuss interpretative equivalence (Coffey, 2014) as a semantic approach to theory equivalence and provide more extensive methodological considerations to reject it.

On the other hand, theories are reasonably well understood objects, if one treats them as sets of sentences or collections of models. Explaining theory equivalence in terms of the shared propositions they express would attempt to elucidate relations between theories using a philosophically more contentious concept.

Instead, one might investigate relations between theories *as objects or structures* such as mappings (or translations) between their underlying languages that satisfy certain conditions. An investigation of this kind might still refer to propositions, but “proposition” would not be antecedently understood and then relied upon in

[9] I am here talking about the content of the theories as an abstraction to not presume a referentialist understanding for the semantics of metaphysical theories.

the comparison of theories, but a theoretical term introduced for the purpose of determining the equivalence of theories. Similarly, one might introduce "proposition" and "content" as a expressive devices that allow to state that theories are equivalent as a result of an account of theoretical equivalence. This would have the form of an abstraction principle: theories T_1 and T_2 express the same proposition if and only if they are (theoretically or metaphysically) equivalent. In this case, T_1 and T_2 can be said to have the same (theoretical or metaphysical) content.

So how could one argue for the deflation of a metaphysical disagreement? To argue that the disagreement between endurantists and perdurantists is purely verbal, Hirsch (2002) proposes that their theories are sufficiently intertranslatable. Each would assent to their opponents theory if translated into their own language. Given a systematic translation of the endurantist's language into their own, a perdurantist would assent to the sentences (in their language) that correspond to the endurantist's theory, and vice versa. Accordingly, they both would be able to understand their counterpart's theory in their own terms and come to accept that this theory is not in fact incompatible with their own. Moreover, they would be able to speak in their counterpart's language and be able to understand themselves as giving truthful descriptions of the metaphysics of persistence by mentally replacing each sentence they speak with the corresponding sentence in their own language (Hirsch, 2002).

Translating their counterpart's language into their own therefore allows the metaphysicians to connect the metaphysical interpretation of their counterpart's terms with their own interpretation of terms.

A committed realist metaphysician could still insist that the offered translation is inadequate. They might assent to each translated sentence, but claim that the resulting theory is not their own theory, as it fails to capture something they hold to be philosophically relevant. Metaphysicians, for example, will disagree on the choice of the "fundamental ideology", i.e. the choice of privileged concepts that are supposed to capture fundamental features of the world. Even under a shared choice of "fundamental ideology", they might still disagree on which sentences of their theory ought to be understood to express laws of metaphysics, or how non-fundamental features are grounded in something more fundamental.

Similarly, a metaphysician might hold that the theory resulting from this translation is in fact their own theory, but deny that the theory of their counterpart is equivalent to their own. After all, this theory is formulated using different primitive concepts and fails to capture something the metaphysician might care about (see also Potter, 1998). I will return to challenges from genuinely metaphysical considerations, in particular in chapter 5.

The suggestion has therefore to be a normative one: if their theories are equivalent, a metaphysician *should* accept their counterpart's theory if it is presented under an *adequate* translation, independently of their actual inclinations.

But this shifts the focus from the metaphysician, their assertions, and their mental states to the concept of *adequate translations* between theories: the question is then whether theories stand in the correct relations to each other such that someone would have to rationally accept both theories. At this point, one might directly investigate the relations between theories, dropping the reference to rational acceptance.

Accordingly, one might hold the following general view: theories T_1 and T_2 are equivalent if there are *adequate translations* $t_1 : \mathcal{L}_1 \to \mathcal{L}_2$ and $t_2 : \mathcal{L}_2 \to \mathcal{L}_1$ between the languages of T_1 and T_2 such that $T_2 \vdash t_1(\phi)$ for all sentences ϕ of $\mathcal{L}_1$ for which it is the case that $T_1 \vdash \phi$, and vice versa. In this variant of Hirsch's proposal, theories are equivalent if a translation of each theorem of the respective theories can be proven in the other theory.

This is only a minimal account of theoretical equivalence. It is meant to demonstrate what concepts of theoretical equivalence might accomplish for (meta-)metaphysics. I have not yet discussed what makes translations adequate for the purpose of metaphysics. In chapter 5, I will argue that translations that preserve theoremhood between theories put pressure on realists' demand for a fundamental formulation.

1.3.1 Theories as objects of investigation

The preliminary idea of asking which theories are equivalent raises immediate questions: what are the conditions for an adequate translation? What other kinds of relations between theories can we leverage to investigate whether theories say the same thing about the world?

Questions like these are central to what is sometimes called "metatheory" (Halvorson, 2019). Metatheory, as theory about theories, takes theories as its objects of investigation. Metatheory describes properties of theories, often formal properties such as consistency or completeness, and relations between different theories. We might be interested in metatheory to determine whether a theory stands in philosophically interesting relations to other theories – e.g. whether it is reducible to another theory, whether theories overlap, whether they are genuine alternatives, or even if they are in some sense incomparable (see for example Niebergall, 2000a, on formal concepts of intertheoretic reduction). The approach is more familiar in foundational projects, for example in mathematics and physics. In these projects, the challenge is to prove correlata of the axioms of less fundamental theories in candidates for a fundamental theory of the field.

Treating theories as objects of investigation is meant to give a more secure ground for discussing object-theoretic questions and questions concerning the philosophical understanding of these theories.[10]

A metatheoretic approach takes theories primarily as mathematical structures relative to a language that allow for investigating their features. This allows for the use of clearly defined formal concepts, but also raises the question of what these idealized notions can tell us about particular examples. Coffey (2014), for example, holds that formal notions of theory equivalence cannot answer the question whether theories have shared metaphysical commitments. I will return to Coffey's challenge and the connected idea of metaphysical interpretation in chapter 5.

In this thesis, I focus on theories that are formulated using classical first-order logic, both single- and many-sorted. This simplifies the exposition, in particular as I will not be concerned with comparing theories that are formulated using different logical frameworks. Translating between different logics is a deep formal and philosophical issue and warrants its own investigation.[11] This has the additional advantage that I have to be less concerned with the question whether to understand theories syntactically, i.e. as set of sentences closed under logical consequence, or semantically as collection of models, as the semantic and syntactic views coincide in virtue of the soundness and completeness of first-order logic. This is an important feature of the investigation in this thesis: one can compare semantic and syntactic concepts of equivalence while relying on the fact that theories under investigation are held constant between the perspectives. Extending the view to theories formulated using different logics requires a more general perspective that investigates how different logics can be adequately accounted for by inter-theoretic relations.

[10] The metatheoretical perspective resembles attempts by philosophers of science in the first half of the twentieth century, e.g. in understanding physicalism as a claim about the reducibility of theories in the special sciences to physical theories. Halvorson (2019) and others attempt to bring focus on the metatheoretical perspective in the philosophy of science. They propose to use new formal concepts to metatheory (in particular: concepts developed in category theory), and apply metatheory to philosophical questions such as reducibility of different fields, scientific realism and anti-realism, and questions concerning the interpretation of modality in the context of scientific theories (see Halvorson, 2019, in particular ch. 8).

[11] For issues in comparing different logics see for example Potter, 1998; Woods, 2018; Wigglesworth, 2017; Mossakowski et al., 2009; see Williamson (2013) for the assertion that "laws of metaphysics" can be identified with principles of logical systems, specifically of higher order modal logic.

1.3.2 Theoretical equivalence and metaphysical equivalence

"Theoretical equivalence" in the following is to meant to designate any equivalence relation on theories as (maybe mathematical) objects or structures. This characterization is overly general. It includes both philosophically and scientifically useful concepts of equivalence, such as empirical equivalence or definitional equivalence, but also less interesting or scientifically useful intertheoretic equivalence concepts: all theories are equivalent, all consistent theories are equivalent, all theories published in books with a red cover are equivalent.

The question whether two theories are equivalent usually is the question whether they are in fact two presentations of the same mathematical or scientific content. Under this view, a theory would be understood not as dependent on a particular language, but for example as the class of theoretically equivalent formulations in different languages (cf. Coffey, 2014; Szczerba, 1977).

In section 1.2, I noted that concepts of theoretical equivalence and the content of a theory are related.[12] In an important sense, concepts of equivalence determine limits of our theoretical investigation. By asserting that two theories are theoretically equivalent, one asserts that additional distinctions introduced by these theories (such as the particular notation used to formulate the theory) have no influence for the content for the theory (Halvorson, 2019, sect. 8.4). The paradigm example for this assertion is an empiricist who holds that empirical equivalence is the only concept of theoretical equivalence applicable for scientific theories: preservation of observational statements then exhaust the theoretical content of a theory, as theoretical content would be identified with observational content.[13] This idea generalizes to other concepts of theoretical equivalence. Anything not preserved by the relevant concept of equivalence is beyond the limits of our potential knowledge of the world. As structure of a theory, it is at best required for expressive purposes. Concepts of theoretical equivalence would then correspond, depending on their strength and what they preserve, to a range of realist or anti-realist positions in the philosophy of science (Halvorson, 2019, p. 8.4). This immediately means that some concepts of theoretical equivalence are not useful for scientific or philosophically purposes. While one might be concerned what the primitives are in which the "book of the world" is written (Sider, 2011, cf.), so that logically equivalence might not be a sufficiently strong concept of equivalence, it is presumably insubstantial which font is used for

[12] There appears to be a connection between a concept of equivalence and the content of a theory that is shared by all members of its equivalence class; but this connection needs further investigation.

[13] E.g. as the subset of sentences entailed by the theory that is "purely observational".

writing it.

Metaphysical equivalence is meant to capture a concept of theoretical equivalence that is adequate for the context of metaphysical investigations. At this stage of the thesis, I want to broadly characterize metaphysical equivalence as the strictest "worldly" concept of theoretical equivalence.[14]

This means that metaphysically equivalent theories are supposed to be equivalent with respect to what matters in the context of metaphysics: the features that are preserved between metaphysically equivalent theories are to be supposed features that correspond to "something in the world". This characterization still vastly underdetermines equivalence in the context of metaphysics. It is precisely a question for metaphysics to determine what our best theories about the world ought to capture and what our best theories can capture, i.e. which distinctions we should take to be meaningful (McSweeney, 2016; Miller, 2005).

Concepts of theoretical equivalence therefore need to be investigated with metaphysical considerations in mind. As candidates for a concept of metaphysical equivalence, concepts of theoretical equivalence preserve different features that are potentially significant in the context of some, or potentially all, metaphysical debates. Miller (2017) for example holds that a privileged concept of metaphysical equivalence has to account for hyperintensionality in order to track grounding relations.

Understanding (potential) metaphysical assumptions reflected in different concepts of equivalence then helps us to reflect both which distinctions metaphysicians can refer to in their debates and whether the assumptions about the target of the debate might differ between partisans of a particular disagreement.

This can be made plausible by pointing out the role for different ends of the spectrum of plausible equivalence concepts in the context of metaphysics. Strong or narrow concepts of theoretical equivalence are realist about more parts of theories, whereas wider concepts of equivalence allow that one ignores some parts of the theory as not corresponding to some worldly feature. We have strong reasons to hold it implausible that something like typographical details (e.g. the font used) are important; similar that the use of particular symbols for logical or non-logical symbols (French, 2019). But already at the point of choosing the set of logical connectives for classical logic, strong realists about a logical structure of the world would need to claim that the choice of logical symbols matters: a theory expressed using only $\neg$ and $\wedge$ would not be metaphysically equivalent to a theory that differs only in its use of $\neg$ and $\vee$ as their

[14] This is analogous to the characterization of metaphysical modality as the widest "worldly" modality.

logical symbols, even if these theories are definitionally equivalent.On the other side of the spectrum, it appears implausible to hold that all, or maybe only all consistent theories are metaphysically equivalent.[15]

At this point, there appears to be an argumentative impasse: while I argued that everyone who presents a theory has to accept some concept of theoretical equivalence, the prospect of coming to an agreement about which concept of equivalence is adequate for metaphysical theories in general, or even for certain debates, does not seem to be on the horizon. After all, everyone who presents a view will have some motivation to object to their favourite theory being equivalent with the theory of their opponent. As Putnam observes, the object-level dispute repeats on the metalevel (cf. 1987, p. 76).

In this thesis, I take the following approach. I think that a concept of theory equivalence at least as strong as mutual faithful interpretability and at most as strong as definitional equivalence is a good candidate for a scientifically and mathematically useful concept of theory equivalence. A useful concept should account for equivalent theories in different languages, while preserving what the theories identify as theorems. In chapter 2, I present candidate notions that satisfy these criteria. Equivalence notions that satisfy these requirements are used in the examples in chapters 3-5.

Ultimately, I will argue that these concepts should be acceptable for realists, even if they provide reasons to reject the assumption that theories have a metaphysically privileged formulation. As a consequence, these concepts are promising candidates for a concept of metaphysical equivalence, but present a revised understanding of what ontological commitments of a theory can be.

1.4 Outlook

The rest of this thesis proceeds as follows: Chapter 2 introduces formal concepts of theory equivalence and provides a preliminary discussion on whether they are promising in the context of metaphysics. Chapter 3 and 4 provide case studies of pairs of theories that are often discussed as examples of equivalent theories: mereological universalism and nihilism (3), and theories of geometry in languages that have only point variables or only line variables(4). Chapter 5 focuses on the relation

[15] Unfortunately, I do not know how to provide an argument for this point. It would appear that this would deny that conceptual distinctions made by theories track anything in the world, i.e. that there ultimately is a complete disconnect between our mental and linguistic representations and whatever metaphysical reality would be.

between theoretical equivalence and interpretation. This final chapter starts with a rejection of an alternative approach that explains theoretical equivalence in terms of interpretation. On that basis, the second part of the chapter argues that formal equivalence puts strong constraints on interpretations, but also provides a better realist understanding of what can achieved by interpreting theories.

2 Formal Concepts of Theoretical Equivalence

2.1 Theories and language

There are two common views on the nature of (formalized) theories. According to the syntactic view, theories are *sets of sentences closed under logical consequence or derivability*. Sometimes, theories are identified with the respective axiomatizations of such a set, in particular if one does not know whether different sets of axioms have the same deductive closure (for example Barrett and Halvorson, 2016a, p. 468). According to the semantic view, theories are *sets of models*. There is not always an agreement what models are; here, they are understood as models in the model-theoretic sense, i.e. as set theoretic constructs or tuples of a set (the domain) and relations on that set that interpret sentences (Barrett and Halvorson, 2016a, p. 468, Halvorson, 2019, pp. 172ff, see also Enderton, 2001, pp. 155ff).

This raises the question for the relation between the semantic notion of interpretation and the syntactic view. Just as the syntactic view does not presuppose that the theory is presented as an axiomatized theory, the semantic view does not presuppose that one can in advance identify a set of sentences that characterizes the collection of models. In practice, theories would be commonly presented in terms of sentences, either purely in the form of axioms, or in the form of axioms characterizing a set of models. Theories therefore remain something that requires an expression in some language (Glymour, 2013, cf. Halvorson, 2013). The linguistic demands on presenting a structure might be minimal and, for example, may not require specifying particular predicates. But it requires at least some abstract description on the predicates used for specifying relations that obtain in a structure, e.g. by specifying the places of a relation and thereby the arity of a predicate expressing the relation.

For theories formulated in classical first-order logic, and other logical systems that are sound and complete, the semantic and the syntactic view are coextensive with regards to the identity conditions for theories.[1] In the rest of this thesis, I will therefore use both notions as the identity concept for theories. In general, the coincidence of syntactic and semantic notions of theories in the concept allows for the application of both syntactic and semantic metatheoretic perspectives in the investigation of the same theory, as the theory can be presented in both ways. Additionally, there are concepts of theoretical equivalence that are either formulated in semantic or syntactic terms, or have formulations in both; it will be important to discuss their relation. One cannot assume from the outset that there are naturally corresponding pairs of semantic and syntactic equivalence notions, but in the context of classical first order logic, at least the concept of logical equivalence has both semantic and syntactic formulations that coincide due to the soundness and completeness results .

The identity of a theory is relative to the language in which it is formulated, and in particular to the non-logical symbols or *signature*.[2] These are particularly important as whatever theoretical content can be attributed to the logical system is often assumed to be neutral between theories with respect to their theoretical content and metaphysical assumptions. One cannot, however, assume the neutrality of logic outright in the context of metaphysics (and maybe not in mathematics or science either). If the choice of a logical system corresponds to the choice of a metaphysical theory, as Williamson (2013) suggests, discussions of metaphysical equivalence have to account for theories that use different background logics. The dependence on language arises both in the semantic and the syntactic view. On the syntactic side, theories are sets of sentences. We therefore have to (recursively) define what sequences of expressions in the signature form a sentence. On the semantic side, models are defined as structures that interpret the sentences of a language.

For the purpose of this thesis, I will generally assume that we compare theories formulated in shared background theories of classical first-order logic, either single- or many-sorted. This serves two purposes: on the one hand, it limits the scope of the discussion and allows for a better discussion of issues and limitations that arise in the first-order context. On the other hand, it keeps the role of the background logic for the theoretical and metaphysical content of a theory constant. In the context of this thesis, I understand many-sorted logic as a genuinely first order due to two features: first, it has syntactic features of single-sorted first order logic. In particular, many-sorted logic

[1] With the exception of these theories, in the semantic sense, that consist of a class of models that cannot be described by a set of sentences.

[2] I assume that theories compared are formulated on the basis of the same syntactic rules for well-formulated formulae and sentences and differ only with respect to their non-logical vocabulary.

does not have quantifiers that range over properties, i.e. it syntactically does not allow for quantification into the predicate position. Second, and related, a conservative definitional extension of new sorts does not allow for an implicit quantification over properties either. Conservative extensions introducing new sort terms are never "type-raising", in particular do not allow for the definition of a sort corresponding to a powerset.[3]

While there are arguments that many-sorted languages with finitely many sorts can be given a paraphrase in a corresponding single-sorted language (Quine, 1956), others deny that these perspectives are interchangeable, or at least hold that their equivalence needs to be established independently (Barrett and Halvorson, 2017b). This becomes relevant if one means to compare a pair of first order theories in which only one of the theory employs sort terms, as there is no obvious choice of a background perspective for this comparison – do we ought to collapse sort distinctions in favour of single-sorted quantifiers, or do we assert that single-sorted languages have implicit sort terms? While taking single-sorted languages to be a special case of a sorted logic may simplify the comparison between single- and multi-sorted theories, it also might insufficiently capture underlying metaphysical assumptions of single-sorted[4] languages. Note that the dependence of the identity of theories on their language is a demanding notion; from simple changes in the signatures used to formulate theories (e.g. simple substitutions of symbols) to more fundamental differences in the languages used, the same content appears to be expressible in different languages (French, 2019). The question of theoretical equivalence is therefore to find general criteria for which theories ought to be understood as formulations of the same underlying content. This immediately raises the question whether it is adequate to understand theories as language dependent in the way sketched in this section. If there can be equivalent presentations of the same content, in a way suggested by an adequate equivalence notion, it might be better to identify a theory across its presentation in different languages. I will return to this point in section 2.3.1 after a presentation of formal concepts of theoretical equivalence. In particular, I will discuss whether theories should be understood as having formulations in different languages, in light of mathematical and scientific practice.

[3] See Florio (2023) for the claim that the expressive power of multi-sorted languages do not extend that of single-sorted languages.

[4] Or "unsorted", which hints at the potential tension between the perspectives.

2.1.1 Logical equivalence

Logical equivalence is the strictest concept of equivalence I will discuss in the context of this thesis. In the context of formal theories, it typically provides the concept of theory identity.

Theories T_1 and T_2 with a shared signature Σ and a shared language $\mathcal{L}$ are *logically equivalent* if and only if they have the same consequence set, i.e. $Cn(T_1) = Cn(T_2)$, where the consequence set $Cn(T)$ of a theory T is the deductive closure of a theory in a language, i.e. $Cn(T) = \{\phi \mid T \vdash \phi\}$ for sentences $\phi \in \mathcal{L}$. Given a concept of derivability or proof expressed by $\vdash$, this means the following: for every sentence ϕ over a signature Σ, $T_1 \vdash \phi$ if and only if $T_2 \vdash \phi$.[5] Equivalently (for sound and complete logical systems such as classical first order logic), one can define logical equivalence with the corresponding semantic (here: model-theoretic) notion of logical consequence expressed by $\vDash$. Theories T_1, T_2 of a shared signature Σ are logically equivalent if they have the same models, i.e. $\mathcal{M} \vDash T_1$ if and only if $\mathcal{M} \vDash T_2$ for all models $\mathcal{M}$. Here $\mathcal{M} \vDash T$ means that $\mathcal{M} \vDash \phi$ for all sentences ϕ s.t. $T \vDash \phi$. Via soundness and completeness, this illustrates the direct connection between the semantic and the syntactic view concerning theories.

As I take in this thesis theories to be their logical closure (i.e. $T = Cn(T)$, such that for any sentence $\phi \in \text{Sent}_{\mathcal{L}}$, if $T \vdash \phi$ then $\phi \in T$) logically equivalent theories are *the same theory* in the sense of identity.[6] As they are identical, logically equivalent theories ought to be substitutable in all extensional contexts.

Why is this logical equivalence a relevant notion, distinct from identity of theories?

Methodologically, I assume that a theory as object of investigations has the same metaphysical commitments under different axiomatizations of the set of sentences. But this raises the question whether logical equivalence is adequate as criterion for theory identity.

In many cases, theories are by some axiomatization. Given two sets of axioms, it is not trivial whether they are axiomatizations of the same theory. To establish that they are, one needs to prove that their theories, i.e. the deductive closures of the axiomatizations, are the same – one needs to prove the logical equivalence of the axiomatizations.

As an identity concept, logical equivalence ought to be sufficient for the substi-

[5] Unless noted otherwise, I take $\vdash$ to refer in this thesis to derivability in a deductive system of classical logic, e.g. in a Hilbert-style axiomatic system.

[6] At least for the purpose of this thesis; some contexts will demand stronger notions of theoretical identity, in particular if one takes differences between axiomatizations to be important.

tutability in all extensional metatheoretic contexts, but it cannot be a satisfying concept of equivalence in intensional contexts.

Because I focus on the role of equivalence concepts for the comparison of the metaphysical commitments of theories, the presentation of a theory should not matter. Therefore I will be able to focus on (weaker) extensional equivalence concepts for theories. Alternatives to this position, e.g. the idea that the choice of an axiomatization makes a differences for the metaphysical commitments of a theory, entail a stricter conception for the identity of theories. These alternatives therefore do not simply claim that the way a theory is presented is relevant for its metaphysical commitments.

This is to say that I identify theories with sets of sentences closed under logical consequences instead of axiomatizations of these sentences, thereby taking an extensional perspective on theories. As I will return to at the end of this chapter, this view takes a stance, at least for the time being, on which theories are substitutable in all extensional contexts. Throughout the thesis, however, I want to make the case that both theoretical content and metaphysical commitments are shared between theories that are equivalent under equivalence notions weaker than logical equivalence, i.e. weaker than identity of sets of sentences.

Logical equivalence is relative to a logical system, i.e. a choice of logical operators (connectives, quantifiers, other logical symbols) and a concept of derivability and proof (if the logic is given syntactically), or a concept of logical consequence (if the logic is characterized semantically).[7] In the context of theory equivalence and individuation, this can be an obstacle if differences in the logical background theory are understood as differences with respect to their theoretical content.

The choice of a logical system will already be important if one takes theories to be sets of sentences, but might be even more pronounced if one understands theories as *structured* with respect to relations of inference or derivability, including the choice of axioms (cf. Halvorson, 2019, p. 272). As mentioned in chapter 1, there are strong realists concerning the choice of logical connectives matters. They will therefore challenge the equivalence of logics that have different (primitive) logical connectives, even if they are otherwise formulations of the same logic (e.g. classical propositional logic formulated with alternatively $\wedge$ and $\neg$, the usual presentation with $\wedge, \vee, \rightarrow, \leftrightarrow$ $, \neg$, or a full set of (binary) truth-functional connectives).[8] In the following, I will set

[7] I assume here both a distinction between the logical vocabulary and signature, as well as formation rules for well-formulated formulas and sentences of the language.

[8] I will not expand on the concept of equivalence between different formulations of the same logic. One could for example treat the different background logics as definitionally equivalent in the sense described below.

aside considerations about logical realism. I will assume a presentation of classical logic in terms of negation, material conditional, and the universal quantifier ($\neg, \rightarrow, \forall$), but will understand the other usual connectives and the existential quantifier as defined expressions.

Furthermore, I will not discuss comparisons between different logical systems, with one exception: in section 3.3.1, I compare the metaphysical assumptions of plural logic with the metaphysical theory of mereological universalism.

I take this to be a necessary simplification to limit the scope of this thesis that needs to be revisited.[9]

2.2 Weaker concepts of theoretical equivalence

As I have addressed in the previous section, logical equivalence can only address comparisons between theories formulated in the same language. Logical equivalence corresponds to the usual understanding of the identity of theories and has the same dependence on the implementation of theory. This leads to artefacts of the implementation in a particular language – to present the content of a theory, one needs to choose *some* language (cf. Visser, 2015, pp. 2f).

The immediate question is how one can compare theories formulated in different languages in order to distinguish the artefacts of the language from what remains invariant between different presentations of the same content.In the following, I will introduce a series of formal equivalence concepts for theories discussed in the literature that do not presuppose a shared language. Starting first with the concept of definitional extension and definitional equivalence, I will then turn to a series of concepts of theory equivalence that are based on the concept of *relative interpretation*. My focus here is on the motivation of these concepts in a formal context; I will turn to their philosophical discussion in the following section and throughout the subsequent chapters.

The case is more complicated than presented here, as one might need work with different kinds of deductive systems (or be limited to one deductive system), and has to account for theories in different signatures, which means that one needs to resort to some sort of schematic formulation. On discussions for concepts of equivalence between logics see for example Meadows, 2021, Woods, 2018, Pelletier and Urquhart, 2003.

[9] In particular with the view endorsed by Williamson (2013) that metaphysics is supposed to find the most general laws that hold for the world, which in his view have the form of a system of necessitarian higher-order modal logic.

2.2.1 Definitional equivalence

A theory T^+ with signature Σ^+ is a (conservative) *definitional extension* of T with Σ if and only if for every symbol s in $\Sigma^+ - \Sigma$, there is an explicit definition of s, δ_s, in terms of Σ such that $T^+ = T \cup \{\delta_s \mid s \in \Sigma^+ - \Sigma\}$.[10] An explicit definition (of an n-place predicate symbol $R(x_1, \ldots, x_n)$ is a sentence of the form $\forall x_1 \ldots x_n[R(x_1, \ldots, x_n) \leftrightarrow \phi]$, where all free variables (if any) in ϕ are among the $x_1, \ldots, x_n$. Theories T_1 and T_2 are *definitionally equivalent* if they have a *common definitional extension* (CDE), i.e. if they can be definitionally extended to the same theory.

Take theories T_1 and T_2 with disjoint signatures Σ_1 and Σ_2, formulated in the languages $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively.[11]

These theories are definitionally equivalent if and only if there are sets of explicit definitions D_1 and D_2 that conservatively extend each of the original theories into a theory with a shared signature Σ^+.[12] This results in theories $T_1^+ = T_1 \cup D_1$ and $T_2^+ = T_2 \cup D_2$ that have the same deductive closure ($T^+ = Cn(T_1^+) = Cn(T_2^+)$) and are therefore logically equivalent (and identical, see above) (Glymour, 1970, p. 297). The extended theory T^+ is a common definitional extension of T_1 and T_2.

Definitional equivalence allows to compare theories that do not have a shared language. Because conservative definitional extensions do not allow the proof of additional sentences in the original language, definitional equivalence is a good candidate for theoretical equivalence: a common definitional extension means that there is a theory that proves all theorems of both original theories, but does not prove any additional sentence in either of their languages.

2.2.2 Interpretations

Definitional equivalence coincides with the notion of the *synonymy* between theories. This alternative formulation is based on a concept of *relative interpretation* between theories that provides the basis for a range of equivalence concepts that are weaker than logical equivalence. I will now introduce these concepts and briefly motivate

[10] To be precise, T^+ is the consequence set $T^+ = Cn(T \cup \{\delta_s \mid s \in \Sigma^+ - \Sigma\})$, but I will use the imprecise notation for better readability.

[11] My presentation of formal equivalence concepts generally assumes that the signatures of the theories are disjoint. The independence of the signatures can be ensured by requiring that all elements of both signatures are renamed. In section 2.3, I will return to a philosophical motivation for treating the signatures as independent.

[12] Halvorson (2019, p. 123) requires that *admissibility conditions* for the explicit definition of functions and constants, i.e. conditions of existence and uniqueness, are derivable in the original theories. Admissibility conditions guarantee that the explicit definitions are conservative.

each of these notions from a formal perspective. My presentation of interpretations and equivalence concepts defined in terms of interpretations is based on Friedman and Visser (2014, in particular appendix A), Visser (2015), and Button and Walsh (2018, ch. 5).

Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be single-sorted first-order languages with relational signatures Σ_1 and Σ_2.

An *m-dimensional translation* $\tau : \mathcal{L}_1 \rightarrow \mathcal{L}_2$ is given by two things:

- a domain formula $\delta(\bar{x})$ of $\mathcal{L}_2$;
- for each n-place relation $R(x_1, \ldots, x_n) \in \Sigma_1$ a formula $\tau(R(\bar{x}_1, \ldots, \bar{x}_n))$ of $\mathcal{L}_2$, where the $\bar{x}_i$ are pairwise disjoint sequences of m variables, such that predicate logic proves $\tau(R(\bar{x_1}), \ldots, \bar{x}_n)) \rightarrow (\delta(\bar{x}_1) \wedge \cdots \wedge \delta(\bar{x}_n))$. This includes the identity-relation, for which we assume that $\tau(x = y)$ is an equivalence relation $E(\bar{x}, \bar{y})$, which may be the identity relation =.

A translation τ maps $\mathcal{L}_1$ formulas to $\mathcal{L}_2$ formulas as follows:

- $\tau(R(x_1, \ldots, x_n)) := \tau(R)(\bar{x}_1, \ldots, \bar{x}_n)$;
- $\tau(\cdot)$ commutes with the propositional connectives;
- $\tau(\forall x \phi) := \forall \bar{x}(\delta(\bar{x}) \rightarrow (\tau(\phi)))$;
- $\tau(\exists x \phi) := \exists \bar{x}(\delta(\bar{x}) \wedge (\tau(\phi)))$.

My presentation of m-dimensional translation is slightly simplified, as it is limited to relational signatures and does not consider the translation of constants and function-symbols. But the extension to non-relational signatures is natural (see for example Button and Walsh, 2018, pp. 115ff).

An 1-dimensional translation τ for which it is the case that $\tau(x = y) := (x = y)$ interprets identity absolutely or *preserves identity*. Identity preserving 1-dimensional translations are a special case that support particular notions of equivalence. Generally, an m-dimensional translation τ preserves identity if $\tau(\bar{x} = \bar{y}) := \bigwedge_{i \leq m} (\delta(x_i) \wedge \delta(y_i) \wedge x_i = y_i)$.

Translations determine the construction of quotient structures, as presented in Visser (2015, A.2): Assume a k-dimensional translation $\tau : \mathcal{L}_1 \rightarrow \mathcal{L}_2$ and a model $\mathcal{M}$ with domain M and signature Σ_1. Let $N = \{\bar{m} \in M^k \mid \mathcal{M} \models \delta(\bar{m})\}$. Assume that N is not empty. Let E be the equivalence relation on N defined by $\tau(=)$. Then τ determines a model $\mathcal{N}$ with N/E and with $\mathcal{N} \models \tau(R([m_1]_E, \ldots, [m_n]_E))$ if and only if $\mathcal{M} \models R(m_1, \ldots, m_n)$.

Let T_1, T_2 be theories in the first order languages $\mathcal{L}_1$ and $\mathcal{L}_2$. A translation $\tau : \mathcal{L}_1 \to \mathcal{L}_2$ supports a *relative interpretation* of T_1 in T_2 if for all sentences $\phi \in T_1$, if $T_1 \vdash \phi$ then $T_2 \vdash \tau(\phi)$. We say that T_1 is *relative interpretable* in T_2.[13][14]

This means that T_2 interprets T_1 if and only if there is a translation τ such that T_2 proves the τ-translations of all theorems of T_1. Tran-Hoang (2021, p. 119) holds that interpretation therefore can be intuitively understood as a signature-neutral notion of the subtheory relation. A translation between the languages allows to assert that every theorem of T_1 is also a theorem of T_2, albeit under the guise of a translation. An interpretation $\tau : T_1 \to T_2$ is *faithful* if and only if it also preserves the non-theorems of T_1, i.e. $T_1 \vdash \phi$ iff $T_2 \vdash \tau(\phi)$.

Note that the notion of an m-dimensional interpretation is general in two important ways: the theories do not have to quantify over the same objects, and it allows that structures can be constructed on equivalence classes. This feature of interpretation allows for the definition of *quotient structures* by translating identity into any equivalence relation of the interpreting theory, i.e. we have cases in which $\tau(x = y) := Exy$ (for 1-dimensional interpretations, or (in general), $\tau(x = y) := E\bar{x}\bar{y}$ for an equivalence relation E.

In terms of the models of theories, an interpretation $\tau : T_1 \to T_2$ allows for the uniform construction of models of T_1 as quotient structures of models of T_2 and thereby to embed a model of T_1 in an isomorphic substructure of a model of T_2 (cf. Button and Walsh, 2018, sect. 5.3). The model induced in this way by τ from a model $\mathcal{M}$ of T_2 is sometimes called an "internal model" $\tilde{\tau}(\mathcal{M})$ of T_1 (cf. Friedman and Visser, 2014; Visser, 2015). For a concept of theory equivalence, it is important to consider whether these isomorphisms between the structures, i.e. the internal models under pairs of interpretations, are definable in terms of the theories, or whether they are only expressible in the metatheory.

This leads to the first notion of theoretical equivalence in terms of interpretations, in which no additional conditions are put on the interpretations.

Theories T_1 and T_2 are *mutually interpretable* if and only if there are interpretations $\tau_1 : \mathcal{L}_1 \to \mathcal{L}_2$ and $\tau_2 : \mathcal{L}_2 \to \mathcal{L}_1$. Mutual interpretability is an equivalence relation (Friedman and Visser, 2014, p. 4).

Mutual interpretability is weaker than definitional equivalence: Although defini-

[13] I will sometimes identify an interpretation and its supporting translation. This is not precise, but unproblematic for my purposes. See Appendix A of Friedman and Visser, 2014.

[14] If one does not assume that the theories share a background logic, relative interpretation needs to account for the interpretation of operators, cf. Meadows, 2021.

tional equivalence entails mutual interpretability, Barrett and Halvorson (2022) note that there are theories that are not definitionally equivalent (or Morita equivalent), but are mutually (faithfully) interpretable. In such a case, there are translations that embed one theory into the other, and vice versa, but the theories are not definitionally equivalent and do not have the same models (see Button and Walsh, 2018, sect. 5.3).

Translations compose, as do the interpretations they support: if $\tau_1 : \mathcal{L}_1 \longrightarrow \mathcal{L}_2$ and $\tau_2 : \mathcal{L}_2 \longrightarrow \mathcal{L}_3$ are translations, $\tau_1 \circ \tau_2 : \mathcal{L}_1 \longrightarrow \mathcal{L}_3$ is a translation.

Composing translations allows us to formulate more demanding notions of equivalence. It allows us to to formulate that the translations preserve features of the theories in ways that are *definable by the theories* (cf. Button and Walsh, 2018, p. 5.4). Doing so shifts the perspective. By composing translations, one can formulate conditions for equivalence that a composition of translations that returns to the original language needs to satisfy.

One can, for example, demand that the theories each need to recognise that the translations preserve the theorems, i.e. that the composition of two interpretations between equivalent theories has to return to sentences that are logically equivalent from the perspective of the theory.

That gives the criterion that the internal models generated by the translations are elementary equivalent, i.e. that they satisfy the same sentences.

Theories T_1 and T_2 are *elementary equivalent* iff there are translations $\tau_1 : T_1 \longrightarrow T_2$ and $\tau_2 : T_2 \longrightarrow T_1$ such that the following conditions hold:[15]

$$T_1 \vdash \phi \leftrightarrow \tau_2\tau_1(\phi), \text{ for all } \textit{sentences } \phi \in \mathcal{L}_1 \tag{2.1}$$

$$T_2 \vdash \phi \leftrightarrow \tau_1\tau_2(\psi), \text{ for all } \textit{sentences } \psi \in \mathcal{L}_2 \tag{2.2}$$

In this case, the internal models of T_1 and T_2 under τ_1 and τ_2 are elementary equivalent.

A stronger condition is to demand that these internal models are not only elementary equivalent, but *isomorphic*. Friedman and Visser (2014, p. 4) calls theories for which there are translations under which the internal models are isomorphic "*iso-congruent*". Iso-congruence is attractive, but means that the recognition that the models are isomorphic is not generally available in terms of the theories.

A stronger requirement is that of *bi-interpretability*, which asserts that there is a *definable isomorphism* between the models of the theories (Friedman and Visser, 2014, p. 4). In the syntactic perspective, this means that the conditions for isomorphisms of models can be proved in the languages under the pairs of interpretations, see Visser, 2015, A.7.2. In contrast to iso-congruence, this means that the theories have to verify that the

[15] cf. Friedman and Visser, 2014, p. 4.

internal models are isomorphic under the translations. This makes bi-interpretability attractive as notion for theory equivalence, as the theories are indistinguishable of each other from the perspective of either theory.

For the case of 1-dimensional identity preserving bi-interpretation, Button (2022, p. 17) states bi-interpretability by observing that composing the translations that support bi-interpretability lead to an self-embedding.

An identity preserving translation $* : T \longrightarrow T$ is a *self-embedding* iff there is some one-place term θ such that the following conditions hold:

1. $T \vdash \forall x \delta_*(\theta(x))$
2. $T \vdash \forall y(\delta_*(y) \longrightarrow \exists! x \theta(x) = y)$
3. $T \vdash R(x_1, \ldots, x_n) \leftrightarrow *(R(\theta(x_1), \ldots, \theta(x_n)))$ for every R in Σ_T.

A self-embedding translation expresses that the theory T is true for its restriction of the quantifiers to a definable condition θ.

Theories T_1 and T_2 are bi-interpretable in a way that preserves identity if there are interpretations $\tau_1 : T_1 \longrightarrow T_2$ and $\tau_2 : T_2 \longrightarrow T_1$ such that their compositions $\tau_1\tau_2$ and $\tau_2\tau_1$ are self-embeddings.

The case of m-place bi-interpretability generalizes the idea to a definable isomorphism and of self-embedding in general. The notion of m-place bi-interpretability is here usually given semantically in terms of uniformly definable isomorphisms between the models of the theories. Friedman and Visser (2014, A.7.2) presents a syntactic definition of an isomorphism between interpretations. Using the notion of an identity interpretation $\tau_{id} : T_1 \longrightarrow T_1$ that maps T_1 to itself, these conditions express that for theories T_1 and T_2, there are interpretations $\tau_1 : T_1 \longrightarrow T_2$ and $\tau_2 : T_2 \longrightarrow T_1$ such that T_1 proves that there is an isomorphism between τ_{id} and $\tau_2\tau_1$ – that $\tau_2\tau_1$ is a self-embedding.

The strongest criterion presented in terms of interpretation I will mention here is *synonymy*. Theories T_1 and T_2 are *synonymous* iff there are translations $\tau_1 : T_1 \longrightarrow T_2$ and $\tau_2 : T_2 \longrightarrow T_1$ such that the following conditions hold :

$$\text{if } \phi \in T_1, \text{ then } T_1 \vdash \phi \leftrightarrow \tau_2(\tau_1(\phi)), \text{ for all } \textit{formulas } \phi \in \mathcal{L}_1 \tag{2.3}$$

$$\text{if } \psi \in T_2, \text{ then } T_2 \vdash \psi \leftrightarrow \tau_1(\tau_2(\psi)), \text{ for all } \textit{formulas } \psi \in \mathcal{L}_2 \tag{2.4}$$

The criterion of synonymy asserts that there is a pair of translations such that their composition map each formula of the language to a logically equivalent formula – the composition of interpretations returns the same theory (cf. Button, 2022, p. 17).

Friedman and Visser (2014) prove that synonymy is strictly stronger than bi-interpretability, but if at least one of the theories allows for the coding of arbitrary sequences, then identity preserving bi-interpretability also implies their synonymy.

Note that synonymy and definitional equivalence are different presentations of the same criterion (cf. Button and Walsh, 2018, p. 108 and note 2). The definitions provided for a definitional extension directly induce the required translations; clauses for the translation of predicates are explicit definitions for a definitional extension.

2.2.3 Multi-sorted languages

The equivalence concepts presented in the previous sections assume that the languages in question are single sorted. In this section, I present two concepts of theory equivalence for multi-sorted languages. First, I present multi-sorted languages, i.e. languages that include sort-terms in their signature. I then introduce *Morita equivalence*, as a generalization of definitional equivalence, and multi-sorted bi-interpretability.

A multi-sorted signature is a signature $\Sigma = \langle S, \mathbf{R}, \mathbf{F}, \mathbf{C} \rangle$ that includes a (non-empty) set S of sort-terms σ_i, in addition to the sets of predicates, function terms, and constant terms R, F, and C.[16] Each n-place relation in R has an arity of $\sigma_1, \ldots, \sigma_n$, where $\sigma_i \in S$ are not necessarily distinct sort-terms for singular terms. Every function term in F has an arity $\sigma_1, \ldots, \sigma_n \to \sigma$ of sort terms $\sigma_1, \ldots, \sigma_n, \sigma \in S$. Every constant term in C has a sort $\sigma \in S$. Additionally, each variable has a sort term $\sigma \in \Sigma$, and we assume an identity predicate $=_\sigma$ for every $\sigma \in S$.[17] Consequently, quantifiers also need to be sorted, so that there are the quantifiers $\forall_\sigma, \exists_\sigma$ for each $\sigma \in \Sigma$.

Informally, sorted quantifiers can be read as ranging only over objects of a specific sort, which can be made precise in the model theory as having separate domain for each quantifier.

Complex expressions of a language with signature Σ – Σ-terms, Σ-formulas, Σ-sentences – are recursively defined in the usual way (cf. Barrett and Halvorson, 2017a, pp. 1045f).

The possibility of languages with sort-terms raises the question whether languages for single sorted first-order logic are a special case of a multi-sorted language with a unique implicit sort symbol. This question has metaphysical consequences, as it

[16] For the ease of presentation, I will assume in the rest of this chapter that the signature is simply a set of predicate-terms, function-terms, constant-terms, and sort-terms, i.e. treat the signature as "flattened" set of symbols.

[17] I will omit subscripts on variables if their type is made clear by the context.

asks whether there is a uniform concept of existence expressed by quantifiers.[18] I will argue in chapter 4 that theoretical equivalence deflates that question, as ontological commitments should be understood as independent of the choice of particular base sorts.

For many-sorted logics, the concepts of theory equivalence presented earlier are inadequate. The notion of a definitional extension does not include the possibility of defining additional sort terms; similarly, the notion of an interpretation did not include clauses for translation of sort terms.

In the following, I present two concepts of theory equivalence that provide ways of handling sort-terms. Barrett and Halvorson (2016b) propose that *Morita equivalence* is an adequate concept of theoretical equivalence. This concept is meant to extend definitional equivalence to many-sorted languages by allowing for the explicit definition of new sort terms.

Let Σ, Σ^+ be signatures such that $\Sigma \subset \Sigma^+$. Following the presentation of Barrett and Halvorson (2017a, pp. 1047f), I now present how symbols in $\Sigma^+ - \Sigma$ can be explicitly defined in terms of Σ (see also Barrett and Halvorson, 2016b).

In the context of multi-sorted languages, an explicit definition of a predicate symbol R of arity $\sigma_1, \dots, \sigma_n$ is a Σ^+ sentence of the form

$$\forall_{\sigma_1} x_1 \dots \forall_{\sigma_n} x_n (R(x_1, \dots, x_n) \leftrightarrow \phi(x_1, \dots, x_n)) \tag{2.5}$$

where $\phi(x_1, \dots, x_n)$ is a Σ-formula.

Explicit definition of a predicate symbol in the multi-sorted context therefore differs from the single-sorted case only in accounting for the sort specific arity of a predicate and the corresponding use of sorted variables and quantifiers.

An explicit definition of a function symbol of a function symbol $f \in \Sigma^+ - \Sigma$ of arity $\sigma_1, \dots, \sigma_n \longrightarrow \sigma$ is a Σ^+ sentence of the form

$$\forall_{\sigma_1} x_1 \dots \forall_{\sigma_n} x_n \forall_{\sigma} y (f(x_1, \dots, x_n) = y \leftrightarrow \phi(x_1, \dots, x_n, y)) \tag{2.6}$$

where $\phi(x_1, \dots, x_n, y)$ is a Σ-formula. The definition of a function symbol f has the Σ-sentence $\forall_{\sigma_1} x_1 \dots \forall_{\sigma_n} x_n \exists_{\sigma} ! y \phi(x_1, \dots, x_n, y)$ as admissibility condition; as in the single sorted case, any admissibility condition required by an explicit definition needs to be provable in the original theory.

An explicit definition of a constant symbol $c \in \Sigma^+ - \Sigma$ of sort σ is a Σ^+ sentence of the form

$$\forall_{\sigma} x (x = c \leftrightarrow \phi(x)) \tag{2.7}$$

[18] See for example Quine, 1956

where $\phi(x)$ is a Σ-formula. The definition of a constant symbol has the Σ-sentence $\exists_\sigma !x\phi(x)$ as admissibility condition.

So far, these clauses only adapted the clauses for the definitional extension of a theory by predicates, function-symbols, and constants to the multi-sorted context. Additionally, multi-sorted logic allows for the explicit definition of additional sort-terms (Barrett and Halvorson, 2017a, pp. 1047f).

I present here how to explicitly define new sort terms of a signature Σ^+ in terms of a signature $\Sigma \subset \Sigma^+$.

There are four cases discussed by Barrett and Halvorson (2017a): the definition of a new sort-term σ as product sort, coproduct sort, subsort, and as quotient sort. These definitions require the definition of additional function terms of $\Sigma^+ - \Sigma$ that relate σ to sorts of Σ.

I will now present these types of definition in turn.

To define σ as *product sort*, one simultaneously defines two function symbols $\pi_1, \pi_2 \in \Sigma^+ - \Sigma$ for the projection functions, where π_1 has arity $\sigma \longrightarrow \sigma_1$, π_2 has arity $\sigma \longrightarrow \sigma_1$, and $\pi_1, \pi_2 \in \Sigma$.[19] A definition of σ, π_1, and π_2 is a Σ^+ sentence

$$\forall_{\sigma_1} x \forall_{\sigma_2} y \exists_{\sigma=1} z(\pi_1(z) = x \wedge \pi_2(z) = y) \tag{2.8}$$

Objects of a product sort σ may be thought of as ordered pairs of objects of sorts σ_1, σ_2 (Barrett and Halvorson, 2017a, p. 1047)

An explicit definition of σ as *coproduct sort* requires two function symbols $\rho_1, \rho_2 \in \Sigma^+ - \Sigma$, where ρ_1 has arity $\sigma_1 \longrightarrow \sigma$, ρ_2 has arity $\sigma_2 \longrightarrow \sigma$, an $\rho_1, \rho_2 \in \Sigma$. A definition of σ, ρ_1, and ρ_2 is a Σ^+ sentence

$$\forall_\sigma z(\exists_{\sigma_1=1} x(\rho_1(x) = z) \vee \exists_{\sigma_2=1} y(\rho_2(y) = z)) \wedge \forall_{\sigma_1} x \forall_{\sigma_2} y \neg(\rho_1(x) = \rho_2(y)) \tag{2.9}$$

A coproduct sort σ may be understood informally as disjoint union of elements of sort σ_1 and σ_2 (Barrett and Halvorson, 2017a, p. 1047).

An explicit definition of σ as *subsort* uses a "canonical inclusion", a function symbol $i \in \Sigma^+ - \Sigma$ with arity $\sigma \longrightarrow \sigma_1$, where $\sigma_1 \in \Sigma$. An explicit definition of σ as subsort and i is a Σ^+-sentence

$$\forall_{\sigma_1} x(\phi(x) \leftrightarrow \exists_\sigma z(i(z) = x)) \wedge \forall_\sigma z_1 \forall_\sigma z_2(i(z_1) = i(z_2) \longrightarrow z_1 = z_2) \tag{2.10}$$

with the Σ-sentence $\exists_{\sigma_1}\phi(x)$ as its admissibility condition. The objects of sort σ are the objects of sort σ_1 satisfying the condition ϕ (Barrett and Halvorson, 2016b, p. 564).

[19] Definition of product sorts naturally extends to the case of products of more than two sorts $\sigma_i \in \Sigma$, given projections π_i (see Tsementzis, 2017, p. 1184).

The last case is an explicit definition of $\sigma \in \Sigma^+ - \Sigma$ as *quotient sort*, which requires a function symbol $\epsilon \in \Sigma^+ - \Sigma$ with arity $\sigma_1 \to \sigma$, where $\sigma_1 \in \Sigma$. The quotient sort σ and the function ϵ are defined by a Σ^+ sentence

$$\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2 (\epsilon(x_1) = \epsilon(x_2) \leftrightarrow \phi(x_1, x_2)) \wedge \forall_\sigma z \exists_{\sigma_1} x (\epsilon(x) = z) \tag{2.11}$$

where $\phi(x_1, x_2)$ is a Σ-formula. A definition of a quotient sort requires that $\phi(x_1, x_2)$ is an equivalence relation. This means that the following Σ-sentences are the admissibility conditions for a definition of a quotient sort.

$$\begin{aligned}
&\forall_{\sigma_1} x \phi(x, x)\\
&\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2 (\phi(x_1, x_2) \leftrightarrow \phi(x_2, x_1))\\
&\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2 \forall_{\sigma_1} x_3 (((\phi(x_1, x_2) \wedge \phi(x_2, x_3)) \to \phi(x_1, x_3))
\end{aligned}$$

Defining σ as quotient sort defines it as the equivalence classes of objects of sort σ_1 under the equivalence relation $\phi(x_1, x_2)$.

Morita extensions of a theory are extensions of a theory by explicitly defining new predicates, function symbols, constants, *or sort terms* in one of the ways presented above.

A theory T^+ with signature Σ^+ is a *Morita extension* of T with signature Σ if and only if $T^+ = T \cup \{\delta_s \mid s \in \Sigma^+ - \Sigma\}$ such that for every symbol s in $\Sigma^+ - \Sigma$, there is an explicit definition of s, δ_s, in terms of Σ and two conditions hold. First, if $\sigma \in \Sigma^+ - \Sigma$ is a sort symbol and $f \in \Sigma^+ - \Sigma$ is a function symbol that is used in the definition of σ, then $\delta_f = \delta_\sigma$. Second, if α_s is an admissibility condition for a definition δ_s, then $T \vdash \alpha_s$.

Theories T_1^0 and $T_2 0$ are *Morita equivalent* if they can be stepwise extended, by chains of Morita extensions, to logically equivalent theories T_1^n and T_2^m, i.e. if there are theories $T_1^1, \ldots, T_1^n$ and $T_2^1, \ldots, T_2^m$ such that

- each theory T_1^{i+1} is a Morita extension of T_1^i,
- each theory T_2^{i+1} is a Morita extension of T_2^i,
- T_1^n and T_2^m are logically equivalent Σ-theories with $\Sigma_1 \cup \Sigma_2 \subset \Sigma$.

Each T_i^{k+1} in the signature Σ_i^{k+1} extends T_i^k by explicit definitions of the symbols in $\Sigma_i^{k+1} - \Sigma_i^k$ (Barrett and Halvorson, 2017a, pp. 1048f). It is necessary to define Morita equivalence in terms of stepwise Morita extensions, as the introduction of additional sort-terms requires that the signature contains sort terms defined in intermediate steps. Morita extensions are conservative over their base theories, which makes

Morita equivalence to the multi-sorted version of definitional equivalence (see chapter 4, (Barrett and Halvorson, 2016b, Theorem 4.4)).

Interpretations of multi-sorted theories

As for the single-sorted case, one can define translations and supported relations between multi-sorted languages. Doing so requires to translate sort terms using multiple domain formulas.

A *translation* $\tau : \mathcal{L}_1 \to \mathcal{L}_2$ is given by two things:

- for every sort symbol $\sigma \in \Sigma_{\mathcal{L}_1}$, a domain formula $\delta_\sigma(\bar{x})$ of $\mathcal{L}_2$;
- for each n-place relation symbol $R(x_1, \ldots, x_n) \in \Sigma_1$, where x_i is a variable of sort σ_i, a formula $\tau(R)(\bar{x}_1, \ldots, \bar{x}_n))$ of $\mathcal{L}_2$, where the arity of $\bar{x}_i$ is given by the domain formula $\delta_{\sigma_i}(\bar{x})$ for σ_i.

Translations of complex formulas are given as in the single-sorted case, i.e. translation commutes with the logical connectives and quantifiers are relativised to the domain formula of the adequate sort (McEldowney, 2020, p. 402).

Interpretations between theories are defined as in the single-sorted case: As in the single-sorted case, a theory T_1 in T_2 Let T_1, T_2 be theories in the multi-sorted languages $\mathcal{L}_1$ and $\mathcal{L}_2$. A translation $\tau : \mathcal{L}_1 \to \mathcal{L}_2$ supports an *interpretation* of T_1 in T_2 if for all sentences $\phi \in T_1$, if $T_1 \vdash \phi$ then $T_2 \vdash \tau(\phi)$.

Interpretations between multi-sorted theories allows for the definition of the equivalence relations presented in the previous section. This means that one can define all of

- faithful interpretations
- mutual interpretability
- elementary equivalence
- iso-congruence
- bi-interpretability
- and synonymy.

of theories in multi-sorted languages (McEldowney, 2020, pp. 402ff).

I will not define these notions separately for the multi-sorted case. Instead, let me note that if m-dimensional translation (section 2.2.2 is supplemented with domain-formulas for every sort-term in the signature, as presented above, it already provides the resources for translation between multi-sorted languages. The concept of m-dimensional translations, which are not necessarily identity preserving, allows for the translation of a sort term $\sigma \in \Sigma_1$ via a domain formula $\delta_\sigma(x_1, \dots, x_n)$ of arity $\sigma_1, \dots, \sigma_n$ into a construction of sorts $\sigma_1, \dots, \sigma_n \in \Sigma_2$. The extension of m-dimensional translation from single-sorted languages to the multi-sorted case is therefore immediate.

McEldowney (2020) proves an important result concerning the relation between bi-interpretability and Morita equivalence.

Theories T_1, T_2 with signatures Σ_1, Σ_2 are Morita equivalent if and only if they are bi-interpretable, as long as there at least two objects of any particulars sort $\sigma_1 \in \Sigma_1$ and $\sigma_2 \in \Sigma_2$ according to the theories, i.e. if $T_1 \vdash \exists_{\sigma_1} x \exists_{\sigma_1} y(x \neq y)$ and $T_2 \vdash \exists_{\sigma_2} x \exists_{\sigma_2} y(x \neq y)$ (cf. McEldowney, 2020, p. 413).

2.3 Preliminary philosophical discussion

This section is meant to address and motivate a few philosophical questions that arise from this first introduction of formal equivalence notions. These questions fall in three, related, categories. The first set of questions surrounds the language dependence of logical equivalence and therefore theory identity, as introduced earlier. I will dedicate the final section 2.3.1 to support the claim that in a substantial sense, theories are language independent. The second set of question discusses the role of definitions, interpretability, and reduction, and how they capture central ideas about metaphysics. Third, and related, the question arises how formal equivalence notions are related to the meta-theoretic tools used to express equivalence conditions and interpret theories.

The first set of questions surround the language dependence of the notion of theory identity Theories are logically equivalent only if they are formulated in the same language. The identity of a theory therefore depends on its language, strictly speaking. But theories are rarely formulated in the same (formal) language.[20] Comparisons between theories therefore assumes that the descriptive content of theories can be compared even between theories formulated in different languages.

[20] This extends to formulations of theories in the varieties of English (and other natural languages) that are used in the sciences, as they are enriched by theoretical terms and specific definitions of other expressions that are not necessarily shared within a field of research.

An overlap between the languages, by a partially shared language and maybe base theory, is not necessarily helpful for that purpose. The focus of this idealized discussion about formalized (and often axiomatized) theories provides both philosophical and technical reasons for treating the languages as independent. In the axiomatic setting, the meaning of non-logical expressions is determined by the theory itself. For languages that share expressions, we therefore cannot generally assume that the same expression has a shared interpretation. This idealized context generalizes to other fields; in particular in the context of mathematics and metaphysics, where most terms, including quantifiers and identity predicates, can be understood as theoretical terms (cf. Potter, 1998). Accordingly, we should avoid potential conflicts and replace the expressions in at least one of the signatures. Otherwise, one needs to explain explicitly how one's equivalence concepts account for overlapping signatures.[21] Because we cannot assume that theories under comparison share their language, we need ways to express in which sense the content of a theories is independent from its signature (Visser, 2015, p. 2).

Definitional equivalence (or synonymy) here provides an important step towards the independence of a theory from its language.

This brings us to the second set of questions.

It is a common practice in mathematics, sciences, and philosophy to explicitly define new expressions on the backdrop of an existing theory. Defined expressions allow for a concise use of more complex expressions. This invites the understanding that definitions and their introduced terms are purely metatheoretic devices to abbreviate expressions that are otherwise too cumbersome. As metatheoretic abbreviations, definitions contribute to the ease of use of a theory, but neither changes the language nor the theory. Even the staunchest realist would not object against that abbreviating use of definitions.

However, the notion of a definitional extension, as defined above, genuinely extends the language of a theory by new expressions – even if they are explicitly defined in terms of complex expressions in the old language. For a lack of a shared language, the definitional extension is not logically equivalent to the initial theory. But as long as the definitions are conservative, i.e. do not allow for the proof of new theorems in the original language, the use of explicitly defined terms does not change the content of the theories – any theorem that is newly provable in the extended language is a reformulation of a theorem of the original theory.[22]

[21] See Lefever and Székely (2019) for a discussion of definitional equivalence for non-disjoint languages).

[22] This means that the extended theory proves that this theorem is logically equivalent to a theorem of

This is reflected in the semantic perspective: An n-place relation R is definable in an $\mathcal{L}$-structure $\mathcal{M}$ if and only if there is an $\mathcal{L}$-formula $\xi(x_1, \dots, x_n)$ with free variables $x_1, \dots, x_n$ such that $R = \{\langle x_1, \dots, x_n \rangle \mid \mathcal{M} \vDash \xi(x_1, \dots, x_n)\}$. A definitional extension of a theory does therefore not allow for new properties to be defined in its models; properties and relations that are definable in the Σ^+ structures of a definitionally extended theory T^+ are already definable in the original Σ structures.

It is therefore plausible to understand conservative extensions of a theory T to be equivalent to T in every theoretical respect. However, we have already seen metaphysicians that deny this, as they assert that the choice of vocabulary is metaphysically significant. As the extended theory has additional vocabulary, these strong realists would deny that they are both interchangeable from a metaphysical perspective – the expressions of one theory are formulated with fewer, and potentially more fundamental, expression and cut closer to the metaphorical "joints of nature". Terms defined using these primitives might then correspond to non-fundamental features of the world. For discussions of the relation between non-fundamental theory and definability see Hicks and Schaffer (2017) and Dewar (2023, p. 19).

The desideratum of metaphysical reduction, in that sense, asks to identify a privileged language and theory even if we have already established that the theories share their theorems under a translation. I think that this request is misguided, and a more deflationary realism is appropriate. However, I need the rest of this thesis (in particular chapter 5) to develop this idea.

In the philosophy of science, definitional equivalence is presented as central notion by Glymour (1970). A central task for this notion, parallel to metaphysical reduction, is to capture is an initial idea of how a theory can be defined by a more fundamental theory, which thereby can be understood as reducing the extended theory. While interpretability is more often understood as a more central concept for understanding the reduction of theories, definitional extension and equivalence remains a key concept for modelling and understanding both the role of definitions and reduction in the context of scientific theories. Both definitional equivalence and interpretation then play a central role in the philosophy of science and philosophy of mathematics for explaining inter-theoretic reduction.

Interpretations are the central notion for intertheoretic comparisons, as they show how a strictly stronger theory can express (and prove) the theorems of a weaker theory. This motivates the suggestion for using interpretability to understand theory

the original language: for every sentence $\phi \in T^+$, there is a sentence $\psi \in T$ such that $T^+ \vdash \phi \leftrightarrow \psi$.

reduction, where the aim is to interpret some theory in a more fundamental theory (see for example Niebergall, 2000b, Visser, 2006). A paradigmatic example for this idea is the interpretation of Peano arithmetic in Zermelo-Fraenkel set theory (see Button and Walsh, 2018, pp. 116ff).

Interpreting Peano arithmetic in set theory also demonstrates how interpretations allow for the constructions of inner model of the interpreted theory. Interpreting PA in ZFC demonstrates how one can construct, in models of ZFC, models that isomorphic to models of PA.

This discussion of interpretation segues to the third set of questions, which arise around the question which features of theories are preserved under different conditions on interpretation.

This becomes clear in the difference between mutual faithful interpretability, iso-congruence, and bi-interpretability. Mutual faithful interpretability preserves theoremhood between equivalent theories. This means that one can define theories of one theory in the models of the other. However, the theories do not necessarily describe the same structures. Iso-congruence of theories instead means that the theories have isomorphic models; in that sense, the theories describe the same structure (Visser, 2015, p. 4). But this fact is only accessible from the perspective of the meta-theory, by constructing the relevant inner models. Bi-interpretability of theories instead entails that the theories have definably isomorphic models. The theories themselves witness that the composition of the respective interpretation return to a self-embedding (cf. Barrett and Halvorson, 2022, pp. 3ff). Bi-interpretability of theories thereby provides a strong argument that the theory itself does not distinguish between potential metaphysical assumptions underlying each model.

For this reason McSweeney (2016) argues in favour of definitional equivalence as an *epistemic* concept of metaphysical equivalence. Definitionally equivalent theories have a common definitional extension that can be defined using conservative explicit definitions. These explicit definitions are given in their respective languages. From the perspective of each theory in question, one can therefore verify that the theories have the same models, up to isomorphism. Assertions about their equivalence is therefore something that is already expressible in the theories themselves, which makes definitional equivalence (and bi-interpretability) useful for an epistemic access to metaphysical equivalence.

But as I noted above, synonymy is a particular restrictive concept of theory equivalence. In particular, it requires that identity is interpreted absolutely, i.e. that the identity relation of one theory is mapped to the identity relation of the other. For ontology in particular, this means both that synonymous theories have the same models, and that the number of objects is an invariant of theories. But there are good candidates for theories that are equivalent, at least for all mathematical purposes, even if they have different commitments about the cardinality of their models. A realist here has good reason not to defer to mathematical practice and insist that these theories are not equivalent for metaphysical reasons. Here is the sketch of one problem that arises from not interpreting identity absolutely.

Not identity preserving interpretations map the identity relation = of one theory T to an equivalence relation $E(\bar{x}, \bar{y})$ over constructs $\bar{x}$ from the objects of the second theory T'. This equivalence relation, whatever it may be, is therefore understood as "identity for the purpose of interpreting the theory T". Here, a problem may arise due to the fact that the constructs $\bar{x}$ have an internal structure, and are therefore not in general substitutable in all contexts of T' – only in the subtheory that interprets T. Accordingly, it can occur that $T' \vdash \phi(\bar{x})$, but $T' \nvdash \phi(\bar{y})$, for some formula ϕ, even if $E(\bar{x}, \bar{y})$ The equivalence relation on the constructs therefore does not satisfy Leibniz's Law; there are "identical" objects that are discernible.

I think that this reason to reject weaker and non-identity preserving translations should be further investigated, in particular how assumptions concerning the language dependence of theories constrain ontology. This provides part of the philosophical motivation to revisit paradigmatic cases of mereology and geometry in which there are philosophical (chapter 3) and mathematical reasons (chapter 4) to hold that the cardinality of models is not a metaphysical invariant of theories.

2.3.1 Identity of theories

I started this chapter by presenting the assertion of the syntactic view that theories are sets of sentences closed under logical consequence. A theory therefore depends on its formulation in a particular language: it is both dependent on its *language*, but independent of its axiomatization. Candidates for a concept of theory equivalence raise doubts concerning the adequacy of this notion of theory identity. These concepts can be seen as presenting alternatives for the identity of a theory *qua theory*. While one needs some way of presenting one's theory, some details of the implementation of a theory will not be relevant (see Visser, 2015, in particular pp. 1f; cf. French, 2019).

First-order Peano arithmetic formulated in the signature $\langle 0, S, +, \cdot, < \rangle$ does not meaningfully differ, in a mathematical sense, from a formulation that lacks $<$ as primitive predicate. In fact, the formulations are definitionally equivalent, for example by $m < n$ iff$_{\text{def}}\exists x(x \neq 0 \land m + x = n)$. On the other hand, PA and $PA_<$ are different theories, even if they are both recognizably theories of first order Peano arithmetic.

This raises the question whether our understanding of theories is better reflected if we identify theories under a different concept of theoretical equivalence, maybe one that is supported by the practices of researchers. Instead of identifying theories with their formulations in a language, one would speak of a theory as having alternative *formulations*, potentially in different languages, that are equivalent under a concept of equivalence weaker than logical equivalence. The weaker concept of theoretical equivalence would then be understood as the identity concept for theories that, which takes theories as closed under both logical consequence and the rules for adequate translation or definition (cf. Szczerba, 1977).[23]

However, we already encountered realistically inclined metaphysicians who will disagree with the idea that this equivalence, which might be adequate for mathematicians, is also informative for metaphysical purposes. It appears that the demands posed by a metaphysician can be stronger than the standards of equivalence used by researchers of a particular scientific or mathematical field working with these theories. In section 2.1 , I already mentioned that one might need to draw even more fine grained distinctions between theories, beyond logical equivalence, for reasons not depending on metaphysical considerations. For example, it might be important for theories in the empirical sciences whether the description of a phenomenon is entailed by more general assertions or laws, or a is explained only in an ad-hoc manner. From the perspective of formalized theories, this would motivate the investigation of competing presentation of a theory *in the same language*. This cold provide reason to take different axiomatizations of the same consequence set as distinct theories for investigating their theoretical properties and to assess their scientific merit. Metaphysicians likewise draw a wide range of distinctions that are finer grained than definitional and even logical equivalence. Investigating formalized theories using different equivalence concepts therefore would require a more fine grained concept of identity for theories, e.g. taking alternative axiomatizations of the same set of sentences as distinct theories.

How we choose to identify a theory is therefore relative to a particular purpose. I agree with Szczerba (1977) that logical equivalence is too strict for many applications. Definitional equivalence seems to be a first cautious way of recognising that a theory has formulations in different languages. Ultimately, the question appears to be termi-

[23] Note that different fields of research might differ with respect to the concept of theoretical equivalence for the purposes of their own practice.

nological. Working with theories requires working with formulations of the theory in different languages. Argumentative steps that transfer between different formulations (or theories) need to be justified in reference to the concept of equivalence that pertains to the discussion.

I think, however, that metaphysicians interested in a realist understanding of theories should be interested in these concepts, as they provide an explication of the language independence of the content of a theory. If taken serious, these concepts both provide a substantial argument against the assumption that a realist understanding of theories require a privileged formulation. This realist position, which I start to develop in the following chapters, takes not only theoretical content, but also metaphysical assumptions to be independent of details of their "implementation" (Visser, 2015). In this thesis, I therefore focus the main discussion on consequence of theoretical equivalence that is not stronger than logical equivalence. I will continue to use logical equivalence as identity concept for theories, but argue that weaker concepts preserve the content of theories. In particular, equivalent theories are expressible in different languages. I will argue that a demand for stricter equivalence concepts, for example on the basis of metaphysical considerations, fails due to constraints on metaphysical interpretation imposed by formal equivalence that preserves theoremhood (cf. chapter 5).

3 Mereology

3.1 Introduction

In this chapter, I turn to an example that illustrates how formal equivalence notions can inform metaphysics.

The disagreement between mereological universalists and nihilists concerning the *special composition question*, is one of the paradigmatic examples for a debate in metaphysics that is putatively shallow. Mereological (toy) examples are central in Putnam's elaboration of his meta-metaphysical views (Putnam, 1987b, Putnam, 1987a, pp. 113ff, but see Miller, 2014 for a defence of the debate as substantive).

In the meta-metaphysical context, the debate is usually framed as the question whether the disagreement between nihilists and universalists is substantive. But mereological languages and theories often provide a framework for other metaphysical debates, e.g. debates concerning ordinary objects concerned with the problem of the many (Unger, 1980) and material constitution (Wiggins, 1968), or persistence (Hirsch, 2002). The debate between endurantists and perdurantists asks how concrete objects persist; it can be expressed in mereological terms as the question whether they persist by having temporal parts (perdurantism) or are temporally extended simples (endurantism). Hirsch asserts that the language of the endurantist and perdurantist can be intertranslated in a way that preserves theoremhood and thereby shows that these are not competing theories of persistence (Hirsch, 2002). These translations cover mereological vocabulary; in particular, they translate the perdurantist's language that centrally quantifies about (temporal) parts into a mereology-free language, and vice versa. Our understanding of whether different theories of mereology are equivalent in the extreme case of nihilism and universalism thereby may impact our understanding of other theories in metaphysics.

In this chapter, I am primarily concerned with the limitations of formal equivalence concepts for mereology. There is a general understanding that there are more or less obvious ways to describe *mereological nihilism* and *mereological universalism* as equivalent. Ignoring equivalence results in the first place amounts to a failure of realists, whether or not they hold that the theories are metaphysically equivalent, or even that the question of their metaphysical equivalence is not well posed (cf. Putnam, 1987b, p. 76, see also Button, 2013, ch. 19).

However, the case is less simple than often presented. Theories in mereology are commonly expressed in a first-order language and asserted to be genuinely first order theories (but see Varzi and Cotnoir, 2021, pp. 233ff). Demonstrating that nihilism and universalism are mutually interpretable, however, instead requires that nihilism is formulated in a language that has more resources than first-order logic (see for example Warren, 2015). A metaphysical understanding of the equivalence result needs to account for this limitation.

The rest of the chapter has the following structure. I first present the mereological theories of nihilism and universalism as expressed in single-sorted first-order languages. On that basis, I present the claim (and proof) of Halvorson (2019, Example 5.4.4) that universalism and nihilism provide equivalent descriptions of models with finitely many atoms. I will argue that this result is insufficient to show the general equivalence of universalism and nihilism. Describing the general equivalence of nihilism and universalism requires that the nihilist's language has the expressive resources to describe collections of objects. I outline how formulating mereological nihilism in plural logic shows that the theories are equivalent Warren (cf. 2015) and argue that pluralism incurs metaphysical commitments corresponding to the universalist's commitment to composite objects.

3.2 The theories

The special composition question is the question "under what conditions does composition occur"?(Varzi and Cotnoir, 2021, p. 174)[1]

The first difficulty is to determine the background framework in which answers to that question, in particular nihilism and universalism, are formulated. There are two things to note here: first, there is a question about the languages used for presenting the theories. In particular, a question arises whether universalists need to describe collections of objects, and whether a nihilist's language contains mereological expressions with a minimal axiomatization of the parthood relation.

To have an explicit disagreement, one might require that the nihilist's theory con-

[1] See also van Inwagen, 1987, 1990.

tains a sentence that is meant to express the negation of the universalist's claim. But one could equally understand the nihilist as denying the applicability of mereological terms, i.e. to reject that a correct and metaphysically transparent description of any situation requires mereological terms (cf. Putnam, 1987b, p. 71). In that case, the language of the nihilist would not contain any mereological vocabulary. An explicit disagreement can still be formulated: given at least two simples, nihilists and universalists provide different answers to the unrestricted question "How many things are there here?" These answers, and the disagreement, can therefore be formulated using first order logic.

For the task of interpreting universalism in nihilism, the different ways of presenting nihilism are equivalent. At this point, we cannot assume that the expressions of the languages have the same meaning, in particular that mereological terms and quantifiers have a shared meaning between the universalist's and nihilist's language (Hirsch, 2002, cf. see also Halvorson, 2019, p. 147). This is reflected in the demand that the languages are disjoint, which is the general assumption for formal notions of interpretability, see section 2.3. If universalism is interpretable in the nihilist's theory, the translation that supports the interpretation will not directly map the mereological predicates of the universalist to mereological predicates of the nihilist's language.

3.2.1 Core Mereology

Here, I introduce the theory of *Core Mereology M* that axiomatizes a minimal parthood predicate. Core Mereology is a neutral base theory of mereology that puts little demands on the parthood relation. In particular, it does not has any commitments on which objects stand in the parthood relation (Varzi, 2019, section 2.2).

Core Mereology M is formulated the first order language $\mathcal{L}$ with signature $\Sigma = \langle P \rangle$, with the sole binary predicate Pxy that may be read as "x is part of y".

Core Mereology in $\mathcal{L}$ is the theory

$$M = \{\forall x Pxx, \forall x \forall y \forall z (Pxy \wedge Pyz \to Pxz), \forall x \forall y (Pxy \wedge Pyx \to x = y)\} \qquad (3.1)$$

Core mereology only states that the parthood relation is reflexive, transitive, and antisymmetric, i.e. it demands that it is a weak partial order. On the basis of an axiomatization of the parthood relation, different theories of mereology are formulated, which may strongly differ in their principles governing the composition and decomposition of objects (Varzi, 2019, sect. 2.2).

We can use M to define *general fusion predicates* $F_\phi x$, by a schema for defining fusions of objects that satisfy a condition ϕ of $\mathcal{L}$ (Varzi and Cotnoir, 2021, p. 26). This will be useful as abbreviation for giving a concise expression of universalism. A possible axiom schema has the form

$$F_{\phi}z :\equiv \forall x(\phi(x) \to Pxz) \land \forall y(\forall x(\phi(x) \to Pxy) \to Pzy) \tag{3.2}$$

where $\phi(x)$ is an $\mathcal{L}$-formula and x occurs free in ϕ (Varzi and Cotnoir, 2021, p. 26).[2] The schema defines F_{ϕ} as the property instantiated by the minimal fusion of all objects that are ϕ. We can therefore read $F_{\phi}z$ as "z is the fusion of the ϕs". As abbreviation, that allows us to talk about fusions. Officially, however, we do not expand our language. Which conditions determine a fusion is the special composition question. I will now turn to two the two trivial answers this question, universalism and nihilism.

3.2.2 Universalism

Universalists hold that objects compose under every condition.[3] Their response to the special composition question is therefore to assert universal or unrestricted composition.

There is a difficulty for expressing universalism in languages that do not contain expressive resources for talking about collections of objects. To formulate universalism in a first order setting without set theory or a similar background theory, one therefore needs to rely on an axiom schema that asserts that if any objects satisfy an expressible condition ϕ, the corresponding fusion exists, i.e. its fusion predicate $F\phi$ is instantiated.

Using the abbreviation of fusion predicates, axiom schema for unrestricted composition has the following form:

$$\exists x\phi(x) \to \exists z F_{\phi}z \tag{3.3}$$

This abbreviates the schema

$$\exists x\phi(x) \to \exists z(\forall x(\phi(x) \to Pxz) \land \forall y(\forall x(\phi(x) \to Pxy) \to Pzy)) \tag{3.4}$$

The universalist theory U in $\mathcal{L}$ then is the theory

$$U = M \cup \{A_{\phi} \mid \exists x\phi x \to \exists z F\phi z\}, \tag{3.5}$$

where ϕ is an $\mathcal{L}$-formula.[4] This means that we understand universalism as the

[2] Varzi and Cotnoir (2021, pp. 160ff) present and discuss other ways of defining mereological fusions. In this thesis, the differences between these definitions do not matter. I will here understand the fusion predicates as mere abbreviations and not extend the language.

[3] There is a disagreement on whether this includes the empty condition, i.e. whether there is a null-object (Varzi and Cotnoir, 2021, p. 227). I will here assume that there is no null object.

[4] Officially, $U = M \cup \{A_{\phi} | \exists x\phi(x) \to \exists z \forall x(\phi(x) \to Pxz) \land \forall y(\forall x(\phi(x) \to Pxy) \to Pzy)\}$.

extension of core mereology by every instance of the axiom schema of unrestricted composition for every condition expressible in the language.

As Varzi and Cotnoir (2021, p. 176, fn 25) note, this captures only fusions of pluralities that are specifiable in the language. This is a genuine limitation, as this restricts universalists to countably many fusions, that arises from the use of a first order framework (see also Varzi and Cotnoir, 2021, sect. 6.1). The fact that every fusion has a characteristic condition ϕ is the basis for a translation of universalism in a nihilist language.

Adding the axiom schema has the consequence that the every finite domain of a model for universalism has the cardinality of the powerset of its atoms minus one, i.e. $2^n - 1$, where n is the number of objects without proper parts (Varzi and Cotnoir, 2021, p. 179). For infinite models, the axiom schema only determines the existence of countably many fusions, namely those definable by formulas of the language. I will return to these limitations when I discuss in which sense universalism and nihilism are equivalent.

3.2.3 Nihilism

There are two different ways of expressing nihilism. The first approach is to use a theory of mereology that includes an axiom that expresses the nihilist's claim that there are no (non-trivial) fusions. This can be given by an $\mathcal{L}$-sentence that asserts that nothing has a proper part.

$$\forall x \forall y (Pxy \rightarrow x = y) \tag{3.6}$$

Let $M \subset \mathcal{L}$ again be the theory of core mereology. Mereological nihilism is then the theory $N = M \cup \{\forall x \forall y (Pxy \rightarrow x = y)\}$.

The nihilist's theory N is the explicit rejection that there are composite objects that extends a theory of mereology. It is formulated using the same (uninterpreted) language $\mathcal{L}$ used to formulate the universalist's theory U.

As mentioned above, the second approach is to understand nihilism as a rejection of composite objects by using a language without any parthood predicate (or other mereological terms) (Putnam, 1987b, p. 71). In the context of simply comparing only universalism and nihilism, this amounts to comparing the universalist's U with the empty theory.

In other contexts, and how claims about the equivalence between universalism and nihilism are usually motivated, the theories in question are universalist and nihilist extensions of a base theory. Here, theories of a particular structures, or type of structures, are formulated in languages respectively with or without mereological vocabulary. One of these cases, presented by Halvorson (2019, Example 5.4.4) is the subject of the next section.

For the purpose of comparing universalism and nihilism, it is insubstantial whether we understand nihilism as presented as a mereological theory (i.e. in a language with a parthood predicate that satisfies the axioms of core mereology) or as the non-mereological part of this theory. As outlined in section 3.2, this can be justified as follows.

Assume that there is a sense in which the nihilist theory N interprets the universalist theory U. Assume further that the languages of N and U have disjoint signatures, but each theory is an extension of core mereology M and a (potentially empty) base theory. The translation supporting the interpretation cannot map mereological theorems of U to sentences containing mereological terms that are provable in N, as N contains the nihilist's rejection of composition. If U is interpretable in N, it is therefore interpretable in its non-mereological subtheory.

3.3 Equivalence and mereology

Halvorson (2019, pp. 147ff) claims that theories of universalists and nihilists are "weakly intertranslatable", which is his way of stating that they are m-dimensionally bi-interpretable (see also Halvorson, n.d.).

What Halvorson actually proves is not that the theories U and N (or the empty theory) presented above are bi-interpretable. Instead, he proves that in the particular case in which there are exactly two atoms (i.e. objects without proper parts), there are bi-interpretable nihilist and universalist theories describing the case.

Take the claim that there are exactly two objects without proper parts. A nihilist theory describing the case asserts that there are two objects. A universalist, on the other hand, asserts that there are three objects: the two atoms and its fusion.

The idea behind Halvorson's proof is that the nihilist can understand the universalist's claims about the fusion of the atoms as a claim about the pair of the atoms. The universalist can interpret the nihilist's theory as the restriction of their own theory to its non-mereological subtheory, which only quantifies over atoms.

I will now present the outline of the proof for the equivalence of these theories, and then discuss what this result tells us about the relation between universalism and nihilism.

Let T_U be the universalist's theory with signature $\Sigma = \langle P \rangle$, where

$$\begin{aligned} T_U = U \cup \{\exists x \exists y (x \neq y \wedge \forall z (Pzx \rightarrow z = x) \\ \wedge \forall z (Pzy \rightarrow z = y) \\ \wedge \forall u (\forall z (Pzu \rightarrow z = u) \rightarrow u = x \vee u = y))\} \end{aligned}$$

The theory T_U is therefore the universalist's theory described in section 3.2.2 extended by a sentence that that says that there are exactly two atoms. It contains the instances of the axiom schema for unrestricted comprehension in the language $\mathcal{L}$, which means that it entails the existence of a fusion of both atoms.[5]

Let T_N be the nihilist's theory with the empty signature $\Sigma = \langle \rangle$[6], where

$$T_N = \{\exists x \exists y (x \neq y \wedge \forall z (z = x \vee z = y))\}$$

The nihilist theory T_N only states that there are exactly two objects. The theories are bi-interpretable (or "weakly intertranslatable"), which one can readily show by constructing the respective interpretations (cf. Halvorson, 2019, Example 5.4.4).

The non-mereological subtheory of T_U interprets T_N. This means that there is an identity preserving 1-dimensional interpretation $\tau_1 : T_N \rightarrow T_U$. The translation τ_1 is given by the domain formula $\delta_{\tau_1}(x) := \neg \exists y (Pyx \wedge x \neq y)$, i.e. it maps objects of T_N to objects that are atoms according to T_U (Halvorson, 2019, pp. 148f). Accordingly, one can see that T_U proves every sentence of T_N under the interpretation τ_1, $T_U \vdash \tau_1(\phi)$ for every sentence ϕ in T_N

The interpretation of the universalist theory T_U in T_N is less direct, as T_N describes fewer objects than T_U. An interpretation $\tau_2 : T_U \rightarrow T_N$ needs to be 2-dimensional, as there are insufficiently many objects according to T_N to allow for a direct interpretation of the identity predicate. One needs therefore to translate using the domain formula $\delta_{\tau_2}(x_1, x_2) :\equiv x_1 = x_1 \wedge x_2 = x_2$ that maps variables of T_U to pairs of variables of T_N.

In order to interpret identity of T_U in T_N, one needs to define an equivalence relation for pairs of variables in T_N. This equivalence relation on pairs can be given by $E_{\tau_2}(x_1, x_2, y_1, y_2)$ defined as follows:

$$E_{\tau_2}(x_1, x_2, y_1, y_2) :\equiv (x_1 = y_1 \wedge x_2 = y_2) \vee (x_1 = y_2 \wedge x_2 = y_1)$$

Two pairs are equivalent if they are permutations of each other (Halvorson, 2019,

[5] And the uniqueness of this fusion, given the principles in M, see Varzi and Cotnoir, 2021, p. 27.

[6] As I argued in section 3.2.3, nihilist theories need to interpret universalism in their non-mereological subtheory, which in this case can be formulated without non-logical predicates.

p. 149). The interpretation then translates identity in T_U as E_{τ_2} in T_N,

$$\tau_2(x = y) :\equiv E_{\tau_2}(x_1, x_2, y_1, y_2),$$

it maps variables of T_U to pairs of variables of T_N. A translation of P can be given with

$$\begin{aligned}\tau_2(Pxy) :\equiv & \tau_2(P)(x_1, x_2, y_1, y_2)\\ & \equiv (x_1 = x_2 \wedge y_1 \neq y_2 \wedge (x_1 = y_1 \vee x_1 = y_2)) \vee (x_1 = x_2 \wedge y_1 = y_2 \wedge x_1 = y_1)\end{aligned}$$

The translation τ_2 and "the parthood relation to the relation that holds between a diagonal pair and non-diagonal pair that matches in one place" (Halvorson, 2019, p. 155). Here, parthood is defined as the relation that holds between a pair and itself (for identity) or between a diagonal pair and a non-diagonal pair of variables which match at one place. (Halvorson (2019, p. 149) differs from my presentation in understanding parthood as proper parthood; he therefore does not require that $\tau_2(P)$ is reflexive.)

Halvorson then gives a proof that τ_2 interprets T_U in T_N (2019, 149f). As parthood here is not understood as proper parthood, we cannot rely on the proof here. However, it is clear that $\tau_2(P)(x_1, x_2, y_1, y_2)$ fails to hold in T_N precisely for the case of distinct diagonal pairs, $T_N \vdash \neg\tau_2(P)(x, x, y, y) \leftrightarrow x \neq y)$.[7] Halvorson then sketches a proof that the pair of interpretations τ_1 and τ_2 demonstrates that T_U and T_N are bi-interpretable (cf. Halvorson, 2019, p. 155).

3.3.1 Interpreting the result

In this section, I briefly outline the limitations of this result and present the objection formulated in Warren (2015) against using this kind of bi-interpretability result in the context of metaphysics.

Halvorson Provides a very restricted equivalence result; it is limited to nihilist and universalist extensions of a theory that asserts the existence of exactly two atoms. But the approach Halvorson limits his discussion to a pair of theories that assert the existence of two atoms. But his approach can be extended to any pairs of nihilist and universalist theories extends theories describing finitely many atoms. This requires an n-dimensional interpretation $\tau_n : \mathcal{L}_U \to \mathcal{L}_N$ that interprets T_U in the corresponding T_N, where n is the number of atoms described by the base theory. The interpretation τ_n is given by the domain formula $\delta_n(x_1, \ldots, x_n)$ with the corresponding equivalence

[7] Assume otherwise, i.e. assume $T_N \vdash \tau_2(P)(x, x, y, y) \wedge x \neq y$. If we reduce the trivial identities $x = x$ and $y = y$ in $\tau_2(P)(x, x, y, y)$, this leaves $T_N \vdash (y \neq y \vee x = y) \wedge x \neq y$, so T_N would prove a contradiction.

relation E_{τ_n} over permutations of $x_1, \ldots, x_n$ that interprets identity. This allows the nihilist to n-interpret talk about the universalist's objects, i.e. atoms and composita, as talk about n-tuples of atoms. For each of the τ_n, $\tau_n P(x, y)$ can be defined as the relation that holds between n-tuples $\langle x_1, \ldots, x_n \rangle$ and $\langle y_1, \ldots, y_n \rangle$ just in case each of the x_i is one of the y_i.

But Halvorson here shifts the target, at least if we take his rhetoric seriously, as he announces his proof as covering the disagreement between universalists and nihilists (cf. Halvorson, 2019, pp. 147ff, 155). In fairness, Halvorson only explicitly talks about the bi-interpretability of universalist and nihilist theories that describe finitely many atoms cf. 2019, pp. 147ff. This means that he did not demonstrate that our general theories U and N are bi-interpretable. The interpretation of N in U is not the issue. Universalism allows for the definition of the predicate "is an atom", so that the interpretation τ_1 with its domain formula $\delta_{\tau_1}(x) := \neg\exists y(Pyx \land x \neq y)$ determines the mereology-free subtheory for every extension of U.

The problem for Halvorson's is that each interpretation $\tau_m : \mathcal{L}_U \to \mathcal{L}_N$ is an m-dimensional interpretation. Its domain formula is used to map variables of U to m-tuples of variables of N. Any m-dimensional interpretation τ_m, with its domain formula δ_m therefore only witnesses that a nihilist theory describing m many objects interprets its universalist extension. But this means that this τ_m does not witness that the nihilist theory that entails the existence of $m + 1$ many objects interprets its universalist extension. There is therefore no general interpretation that interprets finite universalist theories in their mereology-free subtheories; moreover, this means that there is no interpretation $\tau : \mathcal{L}_U \to \mathcal{L}_N$ that interprets U in N. A second problem arises due to the interpretation of identity of $\mathcal{L}_U$ as equivalence of m-tuples under permutation E_{τ_m}, which is not identity preserving in the sense of section 2.2.2: the sum of two objects a, b is identical to the sum b, a; the pairs $\langle a, b \rangle$ and $\langle b, a \rangle$ are equivalent under E_{τ_2}, but not identical. In the context of the nihilist's language with an empty signature, this interpretation allows for substitution in every context. The problem becomes clear if the nihilist's language is extended by additional non-logical predicates, as it is not generally the case that the pairs $\langle a, b \rangle$ and $\langle b, a \rangle$ satisfy the same formulas. The equivalence notion E_{τ_m} allows only for substitution in the 'simulated' mereological context, whereas identity guarantees unrestricted substitution.

But there is some sense in which universalism and nihilism are equivalent (cf. Putnam, 1987b). Warren (2015, pp. 248ff) takes the approach of interpreting universalism, formulated as a first-order theory, in a nihilist language that allows for plural quantification. Warren holds that universalism and nihilism are not equivalent in a sense that is relevant for ontology. The problem here is twofold: first, there cannot be a direct interpretation of $\mathcal{L}_U$ in $\mathcal{L}_N$– nihilism asserts that there are fewer objects and therefore does not in general define a sufficiently large domain, which means that identity in $\mathcal{L}_U$ cannot be interpreted by identity in $\mathcal{L}_N$. Second, identity of $\mathcal{L}_U$ cannot be uniformly interpreted as equivalence over a domain formula, as presented above.

However, the interpretation in a plural nihilist language preserves the "logical structure" of theories, which Warren understands as the "structure of proofs" (2015, p. 249). Because this includes the structure of quantification, this does not only mean that the theories U and N_p are mutually interpretable, as Warren holds (cf. 2015, p. 252). It also means that Warren's interpretations witness their bi-interpretability, as identity of $\mathcal{L}_U$ can be generally interpreted as identity of pluralities in $\mathcal{L}_{N_p}$. Therefore, we can understand Warren's result as synonymy result. *Pace* Warren, I think that this sense is therefore also important for metaphysical purposes. The ontological commitment to unrestricted composition and its pluralist correlate, a comprehension scheme for pluralities, both have the same expressive role.

I think that plural nihilism, by extending the expressive power of the language, incurs similar metaphysical commitments as the mereological universalist. While unrestricted composition entails an ontological commitment to arbitrary fusions of atoms, a comprehension principle for plural logic entails the commitment to arbitrary collections of objects. I think that we should take this seriously as a metaphysical cost, if we interpret and compare the metaphysical assumptions required by a theory. We might understand both fusions and collections as objects merely as lightweight expressive devices, or as substantive metaphysical commitments, but Warren's and similar results I discuss in the following chapters put pressure on the idea that base objects of a theory are privileged with respect to how we understand its metaphysics (see also Linnebo and Rayo, 2012) The underlying meta-metaphysical assumption that I will continue to develop is that for realists in particular both the theoretical and metaphysical content of a theory should be independent of the particular languages used to express it (cf. Visser, 2015, p. 2)

4 Geometry

4.1 Overview

This chapter discusses theoretical equivalence in a multi-sorted context. In section 2.2.3, I have introduced both Morita equivalence and multi-dimensional bi-interpretation as equivalence concept that account for languages with sort-terms and corresponding quantifiers and variables. My aim is to illustrate how considerations of theory equivalence for multi-sorted languages provide arguments in a meta-metaphysical context. Additionally, I argue that the particular example, Barrett and Halvorson's argument against conceptual relativity, does not sufficiently clarify their realist conclusion, but provides the basis for a language-independent understanding of ontological commitments.

Barrett and Halvorson (2017a) apply Morita equivalence to particular theories in geometry. Their central concern is to argue against a form of conceptual relativity. Theoretical equivalence, according to Barrett and Halvorson, shows that the metaphysical commitments of a theory do not depend on the sort of its base objects and quantifiers.

I will first give an outline of the argument before filling in some of its central details.

For (certain) theories in geometry, there are Morita equivalent theories both in languages that allow for quantification over only points or only lines. Morita equivalence is the correct concept of theoretical equivalence in the context of these theories. The theories are therefore theoretically equivalent. As the theories are equivalent, in a privileged sense, they have the same ontological commitments. But the models of the theories differ – they have models with different cardinalities and different sorts of base objects.

Morita equivalence resolves that tension, as theories can be extended by new sort terms, which allows that new objects can be defined as "logical constructions" (2017, p. 1056) in a way that preserves both the theoretical and metaphysical content of the theory.

The number of base objects is not invariant between equivalent theories, even if there are equivalent theories can be conservatively extended to theories with corresponding sorts. The number of base objects is not a part of the language-independent commitments of theories. Instead, Barrett and Halvorson (2017a) claim that the definable sort structure is invariant between theoretically equivalent theories. I return in section 4.3 to the question of how to understand the metaphysics of logical constructions and what metaphysical commitments are shared between Morita equivalent theories.

In this chapter, I first motivate the idea that some theories of geometry can be equivalently formulated in different languages, but are not definitionally equivalent. I then present in some detail Barrett and Halvorson's proof for the Morita equivalence of the theories in question. On that basis, I discuss how Morita equivalence impacts the way we should understand the ontological commitments of a theory. My focus here is on the definition of new sorts, which Barrett and Halvorson call "logical constructions" (2017, p. 1056). I argue that Barrett and Halvorson's way of understanding Morita equivalence presents a shift in the notion of ontological commitment. This shift ask which sorts are definable by a theory and thereby deflates ontological questions.

4.2 Points, lines, or both

Consider Figure 4.1 (Barrett and Halvorson, 2017a, p. 1045):

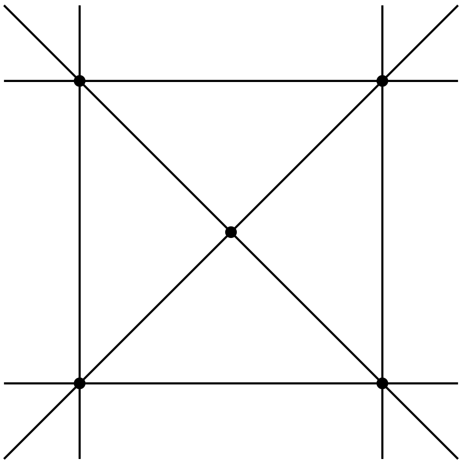

Figure 4.1: A diagram

The diagram can be described in different ways. Prima facie, it is a figure that consists of six lines with five points at which the lines intersect. But there are alternative ways of describing the figure. Intuitively, one can define points as intersections of distinct lines, and lines of pairs of distinct points.

Barrett and Halvorson (2017a, p. 1045) take this observation to argue that there are three different theories that describes the figure. Two of these theories are formulated in sorted languages that have a single sort term, either a sort term for points, or a sort term for lines. The third theory has both sort terms for points and for lines. These three theories are not logically equivalent, as they are formulated in different languages. Furthermore, they are not definitionally equivalent, as they have models of different cardinalities (cf. Barrett and Halvorson, 2017a, p. 1045). The first theory describes the figure as consisting of five points, the second as consisting of six line, and the third as consisting of eleven objects.

But there is a sense in which these three theories are equivalent descriptions of the figure. The idea that we can understand points in terms of lines, and lines in terms of points, can be made precise. This requires a concept of theoretical equivalence that allows for the comparison of theories with different sort terms. In particular, the concept needs to demonstrate how the identity predicates for the sorts of one theory can be defined in terms of an equivalence relation of the other.

Morita equivalence (cf. section 2.2.3) satisfies these requirements. Like m-dimensional bi-interpretation, it is therefore a good candidate for a concept of theoretical equivalence if one assumes that theories can be equivalent even if they have models of different cardinalities. Barrett and Halvorson (2017a) prove that specific theories in geometry formulated in terms of only points there are Morita equivalent theories that are formulated in terms of only lines. Both of these one-sorted theories further are Morita equivalent with a theory that has both sort terms for points and lines.

Barrett and Halvorson (2017a) argue as follows. Starting from a theory T that is expressed with sort symbols σ_p and σ_l, one can prove (under assumptions they assert are "natural" Barrett and Halvorson, 2017a, p. 1050, I will return to the limitations of the result) that there are theories T_p with σ_p and T_l with σ_l that are both Morita equivalent to T. Accordingly, T_p and T_l are also Morita equivalent.[1]

In the following, I present the proofs Barrett and Halvorson (2017a, pp. 1049ff) use to show that T_p, T_l, and T are Morita equivalent, following closely their presentation, but explicitly writing out definitions of new sort terms. My main philosophical interest in Morita equivalence is the question how the conservative definition of new sort terms

[1] I am mainly concerned with Barrett and Halvorson's (meta)-metaphysical conclusions. I will return to the limitations arising from the mathematical assumptions in my discussion in the second part of this chapter.

compares to conservative definition of predicates. I therefore present each step of extending a theory to a Morita extension in similar detail to Barrett and Halvorson (2017a). For each step, I make explicit how the theory is extended using an explicit definition of one of the forms I presented in section 2.2.3. Barrett and Halvorson's proofs draw heavily on the proofs of these propositions of Proposition (*Satz*) 4.59 and Proposition 4.89 in Schwabhäuser et al. (1983). Schwabhäuser et al. show how number variables (Proposition 4.59) and line variables (Proposition 4.89) can be eliminated (stepwise) from theories with both sorts of variables, as long as the theories satisfy specific conditions. The proof that the original theory T and its corresponding theory T_p (or T_l) in the reduced language are Morita equivalent consists in proving that one can return to a theory T_p^+ equivalent to t by stepwise Morita extension of T_p. Morita equivalence therefore describes a precise sense in which the theories in Schwabhäuser et al. (1983) are equivalent.

All theories in the following are formulated with some subset of the following vocabulary.

- The sort symbols σ_l and σ_p. I will follow the convention of Barrett and Halvorson (2017a) and use letters from the beginning of the alphabet to denote variables of sort σ_p and letters from the end to denote variables of sort σ_l.
- The predicate symbol Rax of arity $\sigma_p \times \sigma_l$, which says that point a lies on line x.
- The predicate symbol $Sabc$ of arity $\sigma_p \times \sigma_p \times \sigma_p$, which says that points a, b, and c are *colinear*.
- The predicate symbol Pxy of arity $\sigma_l \times \sigma_l$, which says that lines x and y intersect.
- The predicate symbol $Oxyz$ of arity $\sigma_l \times \sigma_l \times \sigma_l$, which says that lines x, y, and z intersect at a single point or are *copunctual*.

For better readability I will use parentheses for predicates if the terms are complex, in the form $S(a, b, c)$.

Elimination of line variables

Proposition 1
Let T be theory with signature $\Sigma = \{\sigma_p, \sigma_l, R, S\}$, and suppose that T entails the sentences:

1. $a \neq b \rightarrow \exists_{=1} x(Rax \wedge Rbx)$

2. $\forall x \exists a \exists b(Rax \wedge Rbx \wedge a \neq b)$

3. $Sabc \leftrightarrow \exists x(Rax \wedge Rbx \wedge Rcx)$

Then for every Σ-formula ϕ without free variables of sort σ_l, there is a Σ-formula ϕ^*, whose free variables are included in those of ϕ, that contains no variables of sort ϕ_l, and is such that $T \vDash \forall \bar{a}(\phi(\bar{a}) \leftrightarrow \phi^*(\bar{a}))$ (Barrett and Halvorson, 2017a, p. 1050).[2]

Theory T entails that two distinct points uniquely determine a line (1) and that every line has at least two distinct points that lie on it (2); every line therefore by characterised by two distinct points. Furthermore, T explicitly defines colinearity.[3] Note that if one reads off a structure from Figure 4.1, its theory satisfies sentences 1-3 (cf. Barrett and Halvorson, 2017a, p. 1054).

Given these assumptions, one can now show how terms used to formulate T can be defined in terms of T_p. Barrett and Halvorson (2017a, p. 1051) continue by proving their first theorem, which states that T_p and T are Morita equivalent.

Theorem 1

Let T be a theory that satisfies the assumptions of Proposition 1. Then there is a theory T_p in the signature $\Sigma_0 = \Sigma - \{\sigma_l, R\}$ that is Morita equivalent to T.

The Σ_0-theory T_p is given as $T_p = \{\phi^* \mid T \vDash \phi\}$; Proposition 1 entails that there are the sentences ϕ^*.

The proof of Theorem 1 proceeds by demonstrating that there are stepwise Morita extensions $T_p^1, T_p^2, T_p^3, T_p^4$ of T_p and T^+ of T such that T_p^4 and T^+ are logically equivalent.

The theory T_p^1 with signature $\Sigma_1 = \Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2\}$ is the Morita extension of T_p by the definition of a coproduct sort $\sigma_p \times \sigma_p$ and function symbols π_1, π_2 with arity $\sigma_p \times \sigma_p \longrightarrow \sigma_p$,

$$T_p^1 = T_p \cup \{\forall_{\sigma_p} a \forall_{\sigma_p} b \exists_{(\sigma_p \times \sigma_p)=1} x(\pi_1(x) = a \wedge \pi_2(x) = b)\}$$

The theory T_p^1 thereby includes a sort for pairs of points and its associated projection functions.

The theory T_p^2 with signature $\Sigma_2 = \Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i\}$ is the Morita extension of T_p^1 with a definition of a subsort σ_s and a function symbol i with arity $\sigma_s \longrightarrow \sigma_p \times \sigma_p$,

$$\begin{aligned} T_p^2 = T_p^1 \cup \{&\forall_{\sigma_p \times \sigma_p} x(\pi_1(x) \neq \pi_2(x) \leftrightarrow \exists_{\sigma_s} z(i(z) = x)) \\ &\qquad\qquad \wedge \forall_{\sigma_s} z_1 \forall_{\sigma_s} z_2(i(z_1) = i(z_2) \longrightarrow z_1 = z_2)\} \end{aligned}$$

[2] See also Schwabhäuser et al., 1983, Proposition 4.59.

[3] While Σ contains S as primitive predicate, one needs to prove later that colinear points define a line.

The theory T_p^2 thereby defines a sort σ_s for pairs of distinct points as subsort of $\sigma_p \times \sigma_p$. A pairs of distinct points characterise a unique line, but the same line can be described by multiple pairs of distinct points. A line then can be defined by an equivalence class of pairs of distinct colinear points.

The theory T_p^3 with signature $\Sigma_3 = \Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \sigma_l, \epsilon\}$ is the Morita extension of T_p^2 with a definition of a quotient sort σ_l and a function symbol ϵ with arity $\sigma_s \to \sigma_l$,

$$\begin{aligned} T_p^3 = T_p^2 \cup \{&\forall_{\sigma_s} x \forall_{\sigma_s} y[\epsilon(x) = \epsilon(y) \leftrightarrow S(\pi_1 \circ i(x), \pi_1 \circ i(y), \pi_2 \circ i(y)) \\ &\wedge S(\pi_2 \circ i(x), \pi_1 \circ i(y), \pi_2 \circ i(y))] \\ &\wedge \forall_{\sigma_l} z \exists_{\sigma_s} x(\epsilon(x) = z)\} \end{aligned}$$

The theory T_p^3 defines the sort σ_l of lines as quotient sort over colinear pairs of distinct points of sort σ_s. One can verify that T_p^2 satisfies the admissibility conditions for this definition by noting that pairs of distinct points $\langle a, b\rangle$ and $\langle c, d\rangle$ determine the same line if and only if both points a, c, d and b, c, d are each colinear.

The theory T_p^4 with signature $\Sigma_4 = \Sigma_0 \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \sigma_l, \epsilon, R\}$ is the Morita extension of T_p^3 with a definition of a predicate Rax, with arity $\sigma_p \times \sigma_l$,

$$T_p^4 = T_p^3 \cup \{\forall_{\sigma_p} a \forall_{\sigma_l} x(Raz \leftrightarrow \exists_{\sigma_p \times \sigma_p} x \exists \sigma_s y(\pi_1(x) = a \wedge i(y) = x \wedge \epsilon(y) = z)\}.$$

The predicate Raz expresses that point a is on line z. It is defined based on the observation that if a lies on z, there is another point b such that the pair $\langle a, b\rangle$ determines z (Barrett and Halvorson, 2017a, p. 1052).

With theory T_p^4, the sort terms and predicates of Σ are defined again from the reduced theory T_p. Doing so required the definition of terms that are not in Σ, which means that T_p^4 cannot be logically equivalent to T. Proving that T_p^4 is Morita equivalent to T therefore requires the Morita extension of T to a theory T^+ in a shared signature Σ^+.

The theory T^+ with $\Sigma^+ = \Sigma_4 = \Sigma \cup \{\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \sigma_l, \epsilon, R\}$ is the Morita extension of T with explicit definitions for the symbols $\sigma_p \times \sigma_p, \pi_1, \pi_2, \sigma_s, i, \epsilon$.[4] Interesting here is the definition of the quotient function ϵ, which maps a pair of distinct points to the line that is its quotient object, which can be given as follows

$$\forall_{\sigma_s} x \forall_{\sigma_l} y(\epsilon(a) = y \leftrightarrow R(\pi_1 \circ i(x), y) \wedge R(\pi_2 \circ i(x), y))$$

As required, a pair of distinct point determines a line if and only if both points are on the line.

[4] Note that R and S are already in Σ.

To verify that T_p^4 and T^+ are logically equivalent, one needs to show that $T_p^4 \vDash \phi$ for every sentence ϕ such that $T \vDash \phi$. One can verify that T_p^4 entails the conditions 1-3 of Proposition 1, and that therefore $T_p^4 \vDash \phi \leftrightarrow \phi^*$ for every Σ-sentence ϕ. Because $T_p^4 \vDash \phi^*$, for every $\phi \in Cn(T)$, $T_p^4 \vDash \phi$. Because T_p^4 and T^+ are logically equivalent, T_p and T are Morita equivalent. As Morita equivalence is strictly stronger than bi-interpretability, T and T_p are bi-interpretable.

Elimination of point variables

The elimination of point variables can be shown in an analogous way. Like Barrett and Halvorson (2017a) and Schwabhäuser et al. (1983), I will here give only their assumptions and the broad outline of the proof, as the case is analogous to the elimination of line variables.

Proposition 2 Let T be a theory formulated in $\Sigma = \{\sigma_p, \sigma_l, R, P, O\}$, and assume that T entails the following sentences:

1. $x \neq y \longrightarrow \exists_{\leq 1} a(Rax \wedge Ray)$
2. $\forall a \exists x \exists y(x \neq y \wedge Rax \wedge Ray)$
3. $Oxyz \leftrightarrow \exists a(Rax \wedge Ray \wedge Raz)$
4. $Pxy \leftrightarrow (x \neq y \wedge Oxyy)$[5]
5. $Pxy \leftrightarrow (x \neq y \wedge \exists a(Rax \wedge Ray)$

Then for every Σ-formula ϕ without free variables of sort σ_p, there is a Σ-formula ϕ^*, whose free variables are included in those of ϕ, that contains no variables of sort ϕ_p, and such that $T \vDash \forall \bar{x}(\phi(\bar{x}) \leftrightarrow \phi^*(\bar{x}))$ (Barrett and Halvorson, 2017a, p. 1053). Proposition 2 guarantees that for each formula ϕ in Σ, there is a logically equivalent formula ϕ^* that does not contain any variables ranging over points.

Proposition 2 requires of T that distinct lines have at most one point in common (1), that every point lies on two distinct lines (2), and that compunctuality of lines can be defined (3). Additionally, two lines intersect if and only if they are compunctual (4) or, equivalently given the definition of compunctuality (3), that there is a point that lies on both lines (5).

[5] Barrett and Halvorson (2017a) prints condition 4 (using my notation) with the typographical error ($Sxyy$ is a sort mismatch) 4 $Pxy \leftrightarrow (x \neq y \wedge Sxyy)$. Schwabhäuser et al. (1983) have the correct condition in their proof of Proposition 4.89.

Note again that the conditions 1-5 are true in a model constructed on Figure 4.1 (cf. Barrett and Halvorson, 2017a, p. 1054).

Theorem 2
Let T be a theory that satisfies the assumptions of Proposition 2. Then there is a theory T_l in the signature $\Sigma_0 = \Sigma - \{\sigma_p, R\}$ that is Morita equivalent to T.

Like in the case of the elimination of line variables, the proof of Theorem 2 proceeds by showing that there are chains of Morita extensions $T_l^1, \dots, T_l^n$ of T_l and $T', \dots, T^+$ of T such that T_l^n and T^+ are logically equivalent.

The proof consists in the stepwise definition of sorts, predicates, and corresponding functions to recover σ_p and R from T_l. In turn, this consists in the definition of the product sort $\sigma_l \times \sigma_l$ of pairs of lines, the subsort σ_i of $\sigma_l \times \sigma_l$ that are the intersecting lines (using conditions 1 and one of 4 and 5). The sort σ_p of points is then defined as quotient sort of σ_i that share an intersection,

$$\begin{aligned} \forall_{\sigma_i} a \forall_{\sigma_i} b[\epsilon(a) = \epsilon(b) \leftrightarrow\ & O(\pi_1 \circ i(a), \pi_1 \circ i(b), \pi_2 \circ i(b)) \\ & \wedge O(\pi_2 \circ i(a), \pi_1 \circ i(b), \pi_2 \circ i(b))] \\ & \wedge \forall_{\sigma_p} c \exists_{\sigma_i} a(\epsilon(a) = c) \end{aligned}$$

The predicate Rax then can be defined with the idea that a is the point of intersection of x with another line y, $Rax \leftrightarrow \exists y(Pxy \wedge a = \epsilon(x \times y))$.[6] Like for Theorem 1, T needs to be extended to a theory T^+ to include the additional sort and function symbols defined in the intermediary steps.

The proof of Theorem 2 shows that T_l and T are Morita equivalent, and are therefore also bi-interpretable.

For a theory that satisfy the assumptions of both Proposition 1 and 2, Theorems 1 and 2 combined prove that the theories T, T_p, and T_l are equivalent under the concepts of Morita equivalence and bi-interpretability.

What does this equivalence result show about geometry? Barrett and Halvorson (2017a, p. 1054) mention three theories that satisfies the conditions for both Propositions 1 and 2. As mentioned above, the first theory are theories that describes Figure 4.1. The purpose of this theory is mainly illustrative that a general idea can be made precise using the tool of Morita equivalence. The other two theories Barrett and Halvorson mention are more general geometrical theories.

These theories are projective and affine geometry in the signature σ_p, σ_l, R, which both satisfy the hypotheses of Theorem 1 and 2. Accordingly, there are formulations of both theories in the reduced languages with only one base sort (cf. Barrett and

[6] This definition of Rax uses a simplified notation for better readability.

Halvorson, 2017a, p. 1054f).

Barrett and Halvorson further claim that the result concerning affine geometry can be extended to theories of two dimensional Euclidean geometry and two dimensional Minkowski geometry by adding orthogonality (cf. 2017, p. 1055).[7]

We can understand the theory of Figure 4.1 as useful toy model that allows us to discuss meta-metaphysical questions arising in the context of multi-sorted languages. That the result applies to further theories means that the discussion can provide a better understanding of the relation between ontological commitments of a theory and its signature, including its base objects.

4.3 Defining new objects?

What does that result tell us about the metaphysical commitments of theories T, T_p, and T_l that are Morita equivalent?

Morita equivalence sheds lights on two primary questions. The first question concerns the role of language for the ontological commitments of our theories. The second, related question, concerns whether the definition of new sort terms is acceptable, and how it impacts fundamentality claims.

I think that the upshot is twofold: Barrett and Halvorson's understanding of Morita equivalence shows how debates concerning ontological commitments are deflated. Morita equivalence (and other equivalence notions) shows how the ontology of a theory can be language independent. This view holds that a realist interpretation of the geometrical theories in question gives a permissive answer about which sorts of objects exist. The second upshot is that for equivalent theories, their mathematical content, divorced of metaphysical assumptions, does not determine which sorts of objects are fundamental. The argument here precedes by symmetry considerations. If the construction of new sort terms is acceptable, equivalence results show that theories do not single out a particular objects as fundamental or not-constructed.

Barrett and Halvorson's primary (meta-)metaphysical use of Morita equivalence is to argue that conceptual relativism is false (2017, sect. 6).

They present the conceptual realist with the following dilemma. On the one hand, the relativist may hold that the T_p and T_l are not equivalent. Then these theories say different things about the world; their diverging ontological commitments are not

[7] See also Schwabhäuser et al., 1983, Proposition 4.93 concerning eliminability of point variables for absolute geometry.

simply a matter of the particular language used. On the other hand, the relativist may hold that T_p and T_l are theoretically equivalent. In that case, the Morita equivalence shows that T_p and T_l have the same ontological commitments, even if their base quantifiers range over different sorts of objects (cf. Barrett and Halvorson, 2017a, p. 1060f). Both horns of the dilemma reject conceptual relativism; the ontological commitments of a theory does not depend on the particular choice of language.

The first horn of the dilemma requires that there is a reason to hold that T_p and T_l are not equivalent. The fact that T_p and T_l are Morita equivalent means that there cannot be a theoretical reason to hold that they are inequivalent, as Morita equivalence preserves the theorems. Like Barrett and Halvorson (2016b, Theorem 4.4) show, Morita extensions are conservative extensions in the multi-sorted context. Accordingly, the Morita equivalence of T_p and T_l means that they have the same mathematical content, as they can be conservatively extended to theories entail the same theorems.[8]

To hold that the theories T_p and T_l are not equivalent is therefore to claim that there are extra-theoretical reasons to distinguish them. One particular line of reasoning here is to claim that while Morita extension is conservative with respect to the mathematical content of the theories, it fails to preserve metaphysically relevant features such as ontological commitments or fundamentality. I will address the challenge that Morita equivalence does not preserve ontological commitments first. Following that, I will outline how Morita equivalence puts pressure on the notion of fundamentality.

Ontological considerations are at the core of the second horn of the dilemma. Barrett and Halvorson argue that Morita equivalent theories have the same ontological commitments; conceptual relativism therefore fails, as the ontological commitments of a theory do not depend on its language (2017, p. 1060, cf. sect. 5). Arguing that T_p and T_l are not equivalent for metaphysical considerations therefore requires that Morita equivalence does not preserve ontological commitments.

The second horn of the dilemma accepts that Morita equivalence is the correct equivalence concept for theories like T_p and T_l. I think that this is correct, as Morita equivalence presents a clear understanding of how equivalent theories can be formulated in richer and sparser language, using different base sorts. Whether this is (meta-)metaphysically adequate is a different issue. Barrett and Halvorson argue that Morita extension is not only mathematically conservative, but also ontologically conservative (2017, sect. 5).

[8] See also the proof that Morita equivalence preserves expressive power, Theorem 4.6. (Barrett and Halvorson, 2016b).

Their central argument is to show that the quantifiers ranging over objects of a new sort defined by Morita extension can be explicitly defined as complex quantifier expression of the original language (cf. 2017, pp. 1056ff). They interpret Morita extension as ontologically conservative: for every way of defining a new sort terms and its variables, the corresponding quantifier can be understood as complex quantifier expression constructed from quantifiers that were already in the language (cf. Barrett and Halvorson, 2016b, Theorem 4.4).

Barrett and Halvorson hold that this definition of objects of new sorts as "logical constructions" does not add to the ontological commitments of a theory. They suggest that "[a]dding [...] new sort symbols and associated quantifiers to the theory T, therefore, does not increase one's ontological commitments. Rather, it is just a way of making more explicit the ontological commitments of the original theory T." (Barrett and Halvorson, 2017a, p. 1056, cf. p. 1059)

Note, however, that the ontological commitments of a theory are not simply restricted to its base sorts, as the Morita equivalence of T_p and T_l shows. As they are Morita equivalent, these theories have logically equivalent Morita extensions T_p^+ and T_l^+. These extensions have the same ontological commitments, qua logical equivalence, and are metaphysically conservative over the original theories. Accordingly, T_p cannot have more ontological commitments than T_l, and vice versa.

Ontological commitments of a theory therefore do not distinguish objects of a base sort (and their quantifiers) from those that can be constructed using explicit definitions. Consequently, objects of the sorts σ_p and σ_l are ontological commitments of both T_p and T_l, regardless of whether they are base objects or as definable as "logical constructions". The ontological commitments of a theory are therefore objects and sorts that are *definable* in that theory and for which the (extended) theory entails existence claims. This 'definability structure' is invariant between Morita equivalent theories, rather than the objects of base sort in each language.

Barrett and Halvorson (2017a) are not explicit about this particular understanding of ontological commitments, and are not explicit about what exactly the ontological commitments of the theories are – only that they are made explicit in the construction of new sorts. This is sufficient in the context of their argument against conceptual relativism. Barrett and Halvorson needed to show that Morita extension is ontologically conservative and that Morita equivalence therefore preserves ontological commitments. Their permissive view on ontological commitments supports that this view. *Pace* conceptual relativism, ontological commitments of theories are independent of the particular language used.

This leaves the concern that Morita equivalence does not preserve fundamentality. Due to their respective signatures, Morita equivalent theories $T, T_p, T_l \ldots$ have different primitive terms.

Like definitional equivalence, the primitive terms of a theory, both sort terms and predicates, are not invariant under Morita equivalence. For the purpose of these equivalence concepts, it does not matter whether a particular term is primitive or explicitly defined.

Recall that the proof of the Morita equivalence of T_p and T_l requires the stepwise construction of logically equivalent theories T_p^+ and T_l^+, with a shared signature Σ^+. Morita equivalence therefore only requires that the same sorts of objects are definable from both of the reduced signatures, not that a distinction between primitive terms and defined terms is preserved.

Fundamentality considerations now present a problem for the metaphysical adequacy of Morita equivalence. Theories, e.g. T_p and T_l with different primitive (sort) terms are Morita equivalent. While both T_p and T_l are committed to both points and lines, as their extensions in Σ^+ show, one can ask whether points or lines, or both, are the fundamental objects of (projective) geometry.

I think that this response is difficult to sustain. I have more to say in chapter 5, but I think that Morita equivalence and similar equivalence concept put pressure on the idea the question which of the definable objects and properties of a theory are fundamental.

The general argument is as follows. Morita equivalence provides an explication how theories of projective geometry, for example, can be expressed using different signatures. Both T_p and T_l are sufficient to define the 'full' theory T. The geometrical theorems therefore do not single out any of these theories as fundamental. The alternative of taking T and its unified signature (or further extensions) as fundamental can be rejected on the same basis.

As the proofs of Proposition 1 and 2 show (see in particular the proofs in Schwabhäuser et al., 1983, Propositions 4.59 and 4.89), T can again be reduced by eliminating sort terms, while preserving all of its theorems. These proofs, and the corresponding proofs of Theorems 1 and 2 in Barrett and Halvorson (2017a), explain how the basic sorts of both T_p and T_l can be eliminated from T without losing any theorems. Neither sort of object is indispensable for the 'full' theory T. Singling out any theory and signature as fundamental is therefore not supported by the theorems of T, T_p, and T_l.

I think that realists should embrace that metaphysical assumptions of a theory are language independent. As long as the languages are sufficiently expressive, the choice of language is insubstantial. Definitional extension and reduction of a language by both quantifiers of new sorts and predicates does not change the metaphysics of a theory; extending one's language in that way primarily introduces expressive devices for stating the same theoretical content (Linnebo and Rayo, 2012).

Accordingly, the permissive understanding of the ontological commitments of the theory is the correct realist approach: theories are committed to the entirety of objects and properties they define. The question which objects and properties are fundamental for a theory, on the other hand, is not substantive.

In the next chapter, I provide a more extensive discussion on the role of equivalence in interpreting theories. The central insight is that formal equivalence that preserves theoremhood between equivalent theories, i.e. that is at least as strong as mutual faithful interpretability, describes a limitation on the metaphysical interpretation of theories. Realists need to take into account that the theoretical and metaphysical content of a theory is independent of any particular language used to express the theory. This does not mean, however, that interpretation is underdetermined in an anti-realist sense. Realism, however, cannot require that theories have a fundamental formulation.

5 Equivalence and Metaphysics

5.1 Overview

In the first chapter, I motivated approaching disputes about the substance of metaphysical disagreements by considerations of theory equivalence. In the second chapter, I introduced a range of formal concepts of theory equivalence of varying strength. These concepts rely on the (related) ideas that equivalent theories must share their theorems under an appropriate translation, or that they can be conservatively extended into a theory in a unified language.

Both approaches explicate how a theory can be stated in different languages, and that in this sense, theories (or their content) are independent of the particular language used.

In chapters three and four, I presented commonly studied cases of formally equivalent theories. These cases are paradigmatic examples in which philosophers draw (meta-)metaphysical conclusions from the fact that some relation of formal equivalence holds between theories.

In the mereological case (chapter 3), the formal theories studied are meant as explications of metaphysical claims about what kinds of objects exist. The fact that universalist and nihilist descriptions of a world can be bi-interpreted entails, according to Halvorson (2019), that the number of base type is not an invariant of theories. How many objects and, to an extent, what kind of objects exists is therefore something that metaphysics cannot answer. We have seen that this is a very limited result that does not show what Halvorson claims that it shows. However, the case provided reasons to think that metaphysical commitments cannot simply be reduced to the quantification over fundamental objects.

The geometrical case (chapter 4) is introduced with a slightly different focus. Barrett and Halvorson (2017a) understand the case as a case study in the discussion whether metaphysical assumptions of a theory, such as its ontological commitments, are relative to the concepts or languages used to present the theory. As geometrical theories can be expressed using different basic concepts and terms, their metaphysical assumptions are not relative to a particular language – the theory in question can be identified across languages that do not share primitive (non-logical) terms.

The geometrical case in particular raises the question whether there is a metaphysical significance of the "base objects" corresponding to the base sorts of a language. Against the idea of conceptual relativity, Barrett and Halvorson argue that metaphysical commitments of a theory are independent of the linguistic resources used to state the theory. Instead, equivalent formulations of a theory share their commitments even if they quantify over different sorts of base objects. These commitments may be brought out through the availability of "logical objects", i.e. by what sorts of objects can be constructed on the basis of any of the equivalent formulation of the theory.

Whatever metaphysical implications the geometrical theories in question have, Barrett and Halvorson (2017a) hold, is therefore independent of any particular geometrical language; conceptual relativism cannot be true.

Both cases give reason to think that the metaphysics of a theory does not simply correspond to the linguistic primitives or base objects of a formulation of a theory: there are equivalent theories that do not share their primitives and have models with domains of different cardinality.

The aim of this last chapter is to understand how equivalence concepts relate to the project of interpreting theories. Interpretation is here understood as the broadly semantic project of describing how linguistic objects, or representational objects in general, relate to the world. Part of the overall interpretative task is to establish which metaphysical assumptions are made by the interpreted theory, i.e. how the world in general needs to be in order for the theory to be true. This question is raised pointedly by Barrett (2017, p. 94), who asserts that equivalence between theories, theories "'say[ing] the same thing'", provides a way of accessing how one should interpret the theories.

Coffey (2014) argues that the formalism of a theory underdetermines its interpretation and purely formal relations between formalisms therefore cannot account for central features of interpretation.

Instead, Coffey proposes that theory equivalence should be explicated as sameness of interpretation in an approach called "*interpretative equivalence*" (or "*interpretational equivalence*", see Weatherall, 2019, pp. 12ff). Considerations of formal equivalence at best may inform and constrain the interpretation of formalisms (Coffey, 2014, pp. 836f).

I argue in this chapter that Coffey fails to show the inadequacy of formal concepts of equivalence: interpretation cannot both have the features Coffey relies on in his argumentation and give raise to a non-trivial equivalence concept. Instead, formal equivalence results fundamentally limit the scope of what interpretation can establish. For precisely this reason, formal inter-theoretic features provide an appropriate perspective for our understanding of theories.

The rest of this chapter will proceed as follows.

First, I briefly outline how Coffey understands the relation between theory equivalence and interpretation. Second, I will present Coffey's argument against formal concepts of theory equivalence – they cannot capture an important asymmetry in the interpretation of equivalent theories. I will then argue that Coffey's argument fails; it relies on assumptions about interpretation that Coffey cannot rely on for methodological reasons. While Coffey's argument fails, I think that the way it fails provides important insight on the interdependence between formal theory equivalence and the interpretation of theories: formal concepts of theory equivalence put pressure on the idea that metaphysical interpretation can give a unique characterisation of the metaphysical assumptions of a theory.

The final part of this chapter concludes with a brief summary of the thesis.

5.2 Coffey's argument against formal equivalence concepts

This section presents Coffey's argument against formal concepts of theory equivalence. Coffey holds that concepts of theory equivalence that only track relations between the formalisms of a theory cannot account for central features of how (equivalent) theories are interpreted. On their own, formalisms of theories are abstract (mathematical) objects and do not have any representational content. Interpretation is then the semantic task of describing the relation between a formalism and whatever it is meant to capture. Interpretation describes the word-world (or, more generally, representational device-world) relations that connect a theory with its target system.[1]

Concepts of theory equivalence are supposed to capture whether theories are the same with regards to their theoretical content, i.e. whether they say the same thing about the world.

[1] Dewar (2023) dubs this "external interpretation". This equally applies if the formalism in questions are the sentences of the syntactic view, or other mathematical objects like the models of the semantic view. In both cases, the formalism itself does not determine a unique relation to the world – interpretation describes the relation of this apparatus to the world.

Formal approaches to theory equivalence are motivated by the idea that equivalent formalisms, e.g. equivalent uninterpreted theories, afford the same interpretations: if there are way in which the formalisms of theories are mutually reconstructable, these formalism can be used to express the same content. If differences between formalisms are insignificant for what the uninterpreted theories can say about the world, so the formal perspective, the formalisms can be interpreted by exactly the same systems.

Coffey disagrees with the assertion that formal concepts of equivalence capture whether theories have the same content: formalisms of theories underdetermine their interpretation and afford a wide range of interpretations to an extent that additional scientific and philosophical assumptions are always required. Without these assumptions, the formalism of a theory does not have any content, as there is no way of understanding how a formalism relates to whatever it is supposed to capture (Coffey, 2014, pp. 823f). The content of a theory, in this view, cannot simply be the possible, potentially convoluted, ways a formalism might be interpreted. Giving a particular interpretation is needed to clear up the formalisms, e.g. eliminate "surplus structure", or otherwise describe how to understand the scientific significance of different elements and features of the representational apparatus (Knox, 2014).

Because their formalisms are insufficient guides to interpretation, formal approaches to theory equivalence cannot generally establish whether two theories "say the same thing", as, according to Coffey, there is no sense in which theories say anything without prior interpretation (Coffey, 2014, p. 824). A concept of theory equivalence therefore needs to be established on the basis of interpretation as primary notion. For each case, interpreting the particular theories in question determines whether they are equivalent given the interpretative standards and assumptions applicable to each theory in the context of the particular case.

The central philosophical tension between the two approaches can be summarised as follows: Coffey holds that formalisms do not in themselves have any theoretical content, i.e. do not say anything about the world, but first have to be interpreted. Theoretical equivalence therefore can only be a relation between interpreted theories, not between their uninterpreted formalisms. Equivalence of theories has to be sameness of their interpretation.

Proponents of formal concepts of theory equivalence contend that the formalisms do not determine an interpretation, but afford a range of interpretations. Methodologically, comparisons of the correct or intended interpretations might not be available. However, one might be able to determine whether the assertions of a theory can be recovered in another theory. If this is the case, as Coffey acknowledge, we can be sure that the theories, or their formalisms, afford the same interpretations (cf. Coffey, 2014, pp. 836f).

Coffey, however, holds that formal concepts of theory equivalence cannot account for a central feature of interpretation: philosophers often hold that some formulations are privileged and reflect the metaphysical assumptions of a theory in their formalism. Privileged formulations can be directly interpreted, which means to understand (some of) their terms as corresponding to concrete objects described by the theory (Coffey, 2014, p. 831).

Other, equivalent, theories might be derivative reformulations of these metaphysically privileged theories, and need to be indirectly interpreted (cf. Coffey, 2014, pp. 827f, 831). Because formal concepts of theory equivalence track facts about mutual interpretability of theories, or similar symmetrical features, they do not account for the fact that equivalent theories may be treated differently in how they are interpreted.

5.2.1 The "asymmetry" of theoretical equivalence

The symmetrical relation of theory equivalence, according to Coffey, still has to account for an "asymmetry of reformulation" (Coffey, 2014, p. 827). According to Coffey, this asymmetry obtains in some cases in which the interpretative practice of philosophers and scientists understands one theory as more fundamental than another despite their theoretical equivalence (cf. Coffey, 2014, pp. 827, 831).

For Coffey, this means that two theories T and S can be theoretically equivalent even if S is a reformulation of T, but not vice versa (Coffey, 2014, p. 827).

Formal concepts of theoretical equivalence cannot support these judgements as they track symmetrical relations between theories.

If two theories T and S are formally equivalent under an equivalence relation E, there is a particular range $A_1, \dots A_n$ of formal features shared by them. The features $A_1, \dots, A_n$ exhaust the theoretically relevant features under E. Features out of that range cannot be a reason for treating one of them as more fundamental than the other.

Take now a formal feature, A_{n+1} outside of that initial range, i.e. one that is not preserved by our formal equivalence relation E. If that feature makes T more fundamental than S, it is a reason to treat T and S as inequivalent. But then the range of features $A_1, \dots, A_n, A_{n+1}$ determines a new equivalence relation E^*. Neither E nor E^* explain the "asymmetry of reformulation". Under E the theories are equivalent but no relevant feature explains why T is more fundamental. For E^*, a relevant feature explains the difference in fundamentality, but also entails that T and S are not equivalent (cf. Coffey, 2014, p. 834).

Definitional equivalence, for example, does not preserve the choice of primitive predicates; definitionally equivalent theories can have different signatures. Adopting definitional equivalence as concept of theory equivalence means that the choice of signature is not relevant for theory equivalence.

As a consequence, considering the choice of signature as a reason to hold that one theory is more fundamental than another would simply give reason to assert that the theories are inequivalent and therefore reject definitional equivalence as adequate concept of theory equivalence in favour of a different concept of theory equivalence. Formal concepts of theory equivalence can therefore not justify that a theory is a reformulation of a more fundamental theory (cf. Coffey, 2014, pp. 827, 834).

Coffey takes this argument as a reason to reject formal approaches to theory equivalence.

Instead, Coffey's preferred approach to theoretical equivalence, interpretative equivalence, is supposed to explain how equivalent theories can exhibit the asymmetry of reformulation (cf. Coffey, 2014, pp. 834ff). I will here reconstruct Coffey's argument and his examples, with a focus on the role of metaphysical considerations that support the argument.

The central idea in Coffey's argument is that theories can be interpreted using different methods. A mentioned above, interpretation is to describes how the formalism of a theory connects to the system it is meant to capture.

For the sake of presentation, I will here say that a model M interprets a theory T, or that M is the interpretation of T. This is meant as a way to express the word-world relations between theory and is not necessarily in terms of model theoretic interpretation. I will follow Coffey in that the task of interpreting a theory involves a description of the metaphysical commitments of that theory (cf. Coffey, 2014, p. 385). The assertion that a model M (correctly) interprets T therefore should be understood as also asserting that M represents the metaphysics of T.

Two theories T, S are interpretatively equivalent if there is a model that is the correct or intended interpretation of both T and S – in other words, if both T and S are made true by the same system. Note that according to Coffey, theories can have different interpretations. It can be the case that a model M interprets a theory T in one context, given one set of assumptions for interpretation, but that a metaphysically distinct model N interprets T in a second context. Judgements about the equivalence for a pair of theories therefore depend on how they each are interpreted, and thereby on the context of interpretation (cf. Coffey, 2014, p. 385).

That theories can be interpreted in different ways is central for Coffey's interpretative equivalence. The central assumption here is that interpretation can understand the formalisms of theories in different ways. In particular, interpretation may give different metaphysical weight to terms of different theories (cf. Coffey, 2014, p. 835).

This allows Coffey to distinguish fundamental theories from equivalent, but derived, reformulations of these theories (cf. Coffey, 2014, pp. 835ff). The central assumption here is that only fundamental theories are understood "to contain explicit representations" (Coffey, 2014, p. 836) of the metaphysics of the theory. Coffey does not assume that for every theory, there is a metaphysically perspicuous fundamental formulation (cf. 2014, p. 836), but his argument against formal concepts of theory equivalence relies on the assumption that there are cases in which the formalism of a fundamental theory "explicitly represents the ontology of the underlying theory" (Coffey, 2014, p. 831).

I will therefore assume that some theories can be *directly* interpreted. In this case, I will say that a model M is a "d-interpretation" of T, or that a model M "d-interprets" a theory T. For the purpose of my discussion, it is sufficient to understand d-interpretation as disquotational, but it extends to more sophisticated ways of interpreting a fundamental theory (see for example Knox, 2014, p. 867f).

Other theories' formalisms are not metaphysically perspicuous. These theories have to be interpreted in a different way and the formalism is only indirectly connected to the metaphysics of the theory (cf. Coffey, 2014, pp. 823ff, 838). I will say that a model M is an "i-interpretation" for a theory T, or that M "i-interprets" T. The same model M can d-interpret a theory T but i-interpret another theory S. In his argument against formal Coffey discusses one kind of indirect interpretation, the case in which a theory is understood as a "derivative" reformulation of another theory (2014, pp. 823ff). In this case, a reformulation is i-interpreted by first recovering a fundamental theory, which is then d-interpreted (cf. pp. 836f).

A *derived theory* S, relative to a more fundamental theory T, is interpretatively equivalent to that more fundamental theory. However, their shared interpretation K is a d-interpretation of the more fundamental T, but only i-interprets S. The theories S and T then both should be understood as having the metaphysics of K, the metaphysics found by directly interpreting the fundamental formulation (Coffey, 2014, p. 831).

Coffey's examples then have the following structure.

Theories T and S are formally equivalent, in the sense that one can reconstruct the formalism of S from the formalism of T, and vice versa. However, one of the theories, S, is derived from the fundamental formulation T. The theory T should be directly interpreted, because its formalism "explicitly represents" its metaphysics (Coffey, 2014, p. 831). The d-interpretation M of T reflects its metaphysics. Coffey understands S as merely useful reformulation of T that not itself expressing its metaphysical assumptions. While the formalism of S suggests an ontology in virtue of its base objects, this should not be taken seriously: S should not be d-interpreted precisely because its face value metaphysics is incorrect. Instead, S can and should be i-interpreted, by first recovering the formalism of T. Both theories share their correct metaphysical interpretation, M, and are therefore interpretatively equivalent.

The interpretative approach therefore both explains why T and S are equivalent – they both are metaphysically interpreted by M – but also accounts for the asymmetry of reformulation. As fundamental theory, only T, but not S, can be d-interpreted by M. The fact that these theories are formally equivalent is only instrumental for explaining their interpretative equivalence (Coffey, 2014, pp. 834ff).

Coffey's presents his argument against formal concepts of theory equivalence on the examples of the Newtonian and Lagrangian formulations of classical dynamics, and potential and field formulations of Maxwell's equations of classical electrodynamics (Coffey, 2014, pp. 827ff).

According to Coffey, both of these pairs are formally equivalent (cf. 2014, pp. 828, 830f). Physicists can transfer their results between both formalisms without loss. However, Coffey argues that despite the formal equivalence, physicists and philosophers single out one formulation as fundamental and determine the other theory to be a reformulation (cf. 2014, pp. 828, 831). Only the Newtonian formulation, with an "associated ontology of material bodies possessing mass, interacting via the mediation of different types of forces" (p. 829) is "explicit" about the ontology of classical dynamics (Coffey, 2014, p. 831). The Lagrangian formulation instead represent systems as points in configuration space, and has physical content only due to its relation to the Newtonian formulation (Coffey, 2014, pp. 831ff). Similarly, the field formulation of Maxwell's equations assumes the right kind of objects, fields, and therefore can be seen as fundamental. The field formulation can be recovered from the potential formulation; interpretation therefore can reject that the base objects of the field formulation, potentials, are metaphysical commitments of the potential formulation. Note that in both cases, the possibility of i-interpretation depends on the formal equivalence of the theories.

But as Coffey suggests, we "should expect the elements of these formalisms to be interpreted rather differently" 2014, p. 836. For Coffey, and others he quotes for that purpose (cf. p.831f), the Newtonian ontology simply *is* the ontology of classical dynamics. Accordingly, the Lagrangian formulation is a useful reformulation of a more fundamental theory. Considerations about whether a formalism reflects the correct metaphysics therefore give reason to determine a theory as fundamental or as derivative reformulation.

Two aspects are worth emphasizing at this point, and I will return to them in the next section.

First, while Coffey holds that reformulation is an asymmetric matter, reformulation is underwritten by a formal equivalence of the theories. Indirect interpretation can only understand a theory as derived from a more fundamental one because the formalisms are mutually reconstructible.

Second, the designation of theories as derived or fundamental explicitly relies on metaphysical considerations. A fundamental theory simply has the right base objects, i.e. is correct about the ontology of the target system, a derived theory does not. Coffey does not tell us how to arrive at the judgement that the Newtonian formulation of classical mechanics is metaphysically privileged but assume that this is the case and defers to the judgements of philosophers and practising physicists (cf. Coffey, 2014, pp. 830ff).

Coffey's position is unstable. His argument against formal concepts of theory equivalence relies on metaphysical demands on interpretation that are incoherent – interpretation cannot provide a unique shared interpretation for a pair of formally equivalent theories.

Not only does that mean that his argument against formal equivalence concepts fails, but also that interpretation cannot be the basis for a concept of theory equivalence in the way presented by Coffey.

5.3 Why Coffey's argument fails

Coffey does not explicate the concept of interpretation that is the basis for interpretative equivalence and instead defers to philosophers and scientists who individually interpret each theory in question (cf. Coffey, 2014, pp. 830f). However, Coffey uses the assumption that the correct method of interpreting a theory respects the correct metaphysics of the target system. This assumption, however, undermines interpretation as a basis for theory equivalence.

In this section, I will argue that Coffey faces a dilemma. Summarized, the dilemma consists of the following two options for how a method of interpretation is chosen.

As a first option, one could grant Coffey metaphysical interpretation as a primitive notion, for which the intended interpretation is already known. In this case, the metaphysical interpretation of theories that are mere reformulations are not independently established, but need to be mediated by the relations between the formalisms. However, having prior access to the correct metaphysics of the theory is methodologically undermining.

Alternatively, Coffey may assume that there is a standard way (in a context) for determining the metaphysical interpretation of a theory. I will here discuss *prima-facie* interpretation as broadly disquotational method of interpretation, but the argument extends to every standardised way of interpreting theories. In the cases Coffey describes this leaves interpretation underdetermined.

In either case, facts about the formal relations between the theories, and not facts about interpretation, explain why the theories are equivalent. Coffey's cases therefore do not provide an argument against formal concepts of theory equivalence.

With this summary concluded, I will now start explaining the argument in detail.

Metaphysical interpretation

For the first horn of the dilemma, I will grant a notion of metaphysical interpretation, potentially primitive, that correctly determines M as the metaphysical interpretation of T. Coffey describes this case in his argument against formal concepts of theory equivalence, as presented in the previous section.

The structure of the case is pictured by 5.1. Theory T provides a metaphysically privileged representation of the physics (or other subject specific content) in question.[2] It can therefore be directly interpreted as corresponding to the metaphysical assumptions of M.

$$\begin{tikzcd}
\mathrm{S} \arrow[r, "\tau"] \arrow[dr, dashed] & \mathrm{T} \arrow[d] \\
& \mathrm{M}
\end{tikzcd}$$

Figure 5.1: Interpretation provides correct metaphysics

The theory S, on the other hand, is a formally equivalent reformulation of T. For

[2] The theory is privileged in Coffey's sense, i.e. the "formulation [. . .] explicitly represents the ontology of the underlying theory" (Coffey, 2014, p. 831).

the sake of of clarity, I will assume that T and S are bi-interpretable, but the argument generalizes to other concepts of formal equivalence that preserve theoremhood.[3] As a reformulation of T, S has to be i-interpreted by M. Because the theories T and S share their (intended) interpretation M, they are equivalent according to Coffey.

For Coffey, this means that the theories share the metaphysical assumptions of the fundamental theory T. The surface ontology of S is not relevant for its interpretation; attempting to directly interpret S fails because it presents the wrong metaphysical picture. To interpret S is to understand it as a reformulation and therefore to retrieve its fundamental theory.

The derived theory S is therefore best understood as a piece of mathematics that is useful for scientific practice, but does not have independent metaphysical significance.

Under its i-interpretation (dashed arrow) the theory S has the same metaphysical commitments, M, as T under its d-interpretation. Instead of interpreting S directly, a translation τ has to be considered to recover T from S. Reflecting Coffey's description of the case, the translation τ is adequate for non-interpretative purpose and allows for the reconstruction of T.

The picture then exhibits the asymmetry required by Coffey: both T and S share their intended interpretation, but S is a reformulation of T. Only T "explicitly" reflects the metaphysical assumptions of the theory (Coffey, 2014, p. 831, cf. p. 836).

We can now state a clear problem for Coffey. Why should only one of the theories be interpreted directly? We might be led to believe that the fact that S is a reformulation of T, and not an independent theory, is merely a historical accident. Coffey rejects this understanding of "reformulations-as-provenance" precisely on the basis of metaphysical considerations (2014, fn. 24). The useful reformulation has not interpretation in a "real" structure and cannot be understood in terms of the metaphysics of our world (cf. Coffey, 2014, fn. 19).

But these metaphysical considerations make interpretation inadequate as basis for judgements about the equivalence of theories. If the question is whether theories share their interpretation, we cannot assume the metaphysical picture in which the theories are to be interpreted. The theories, and their interpretations, tell us what is real *according to the theory*. Presupposing a particular metaphysical background here does not work on a methodological level. The standard for theory equivalence is that theories have a shared intended interpretation. We therefore presuppose a particular metaphysics to distinguish between fundamental theory and reformulation.

[3] We may understand T as the Newtonian formulation of classical dynamics, S as the Lagrangian formulation.

We can equally imagine a different community of scientists that formulate a theory of classical dynamics first in terms of the Lagrangian picture. Because the theories S and T are bi-interpretable, their scientific practice can be the same as the practices of our actual historical scientists – wherever our scientists used a statement in the Newtonian language, the Lagrangian scientists could have used the corresponding translation. This includes every statement used in the "external interpretation" (Dewar, 2023) that is required to connect the theory to the phenomena it is used to describe. The empirical basis for the theory and scientific practice would be the same, but every statement involving theoretical terms would have been replaced with its translation.

Because the Lagrangian formulation was the first one developed, it would have been natural for realists to assume that the base-objects of the Lagrangian formulation exist. After all, the formulation is empirically supported and the best available theory that accounts for the dynamics of classical systems. A later developed Newtonian formulation would then appear as a reformulation of the Lagrangian formulation, useful in some contexts, but not reflecting the "real" metaphysics of classical dynamics.

With this parallel hypothetical in mind, we can see that interpretative equivalence cannot rely on assumptions about the "correct" metaphysics.

In his argument, Coffey already assumes that classical dynamics has a particular metaphysics, the realist interpretation of the Newtonian formulation. According to Coffey, any successful interpretation has to match this metaphysics; interpretations that go against this metaphysics are not available. This poses a methodological problem. Assuming the correct metaphysics for determining is circular under the realist picture Coffey takes as theoretical background. Theories are the best, or only, guide towards the metaphysics of the world. What the metaphysical assumptions of the theory are is therefore a central question for the interpretation of the theory. But Coffey then cannot assume that successfully interpreting a theory requires that the interpretation matches the metaphysics of the target system.

Therefore, Coffey cannot rely on the central assumption for the argument that there is an "asymmetry of reformulation" that is not explained by formal concepts of theory equivalence (Coffey, 2014, p. 828).

Coffey recognises that a problem arises for the interpretative approach if no interpretation can be singled out as fundamental. In these cases, Coffey suggests that formal equivalence of theories entails that their interpretation has to be shared, whatever the correct interpretation turns out to be. In this way, interpretation is still the underlying notion that explains theoretical equivalence (Coffey, 2014, pp. 836ff).

My objection to Coffey's argument means that this is the general case: for methodological reasons, interpretation cannot assume a particular metaphysics of the target system. But this means that the theoretical work is done by the formal equivalence relation and not by interpretation.

Therefore Coffey's argument against formal concepts of theory equivalence fails if we grant a notion of interpretation that produces the correct metaphysics.

Standardised direct interpretation

However, Coffey assumes that the theories in question are formally equivalent. As I have argued, we cannot simply assume a particular metaphysical picture a successful interpretation needs to support.

In terms of the hypothetical given above, one cannot simply assume that the Lagrangian formulation misses the metaphysics of classical systems and is therefore simply a piece of useful mathematics that cannot be directly interpreted. The situation for interpreting theories is not simply asymmetrical, as pictured in 5.1, but should be rather visualised in the symmetrical 5.2.

This brings us to the second horn of the dilemma, which assumes a standard method for interpreting a theory. We will see that this leaves interpretation underdetermined.

A standard method for interpretation provides an interpretation for any theory; .

Per the assumption of formal equivalence of T and S, the formalism of T can be recovered from S, and *vice versa*. Accordingly, there is not only a mapping $\tau : S \longrightarrow T$ that recovers T from S, but also a mapping $\sigma : T \longrightarrow S$ that recovers S from T. Analogous to the interpretation M, a model N d-interprets S, but i-interprets T, via the mapping σ. To follow Coffey's argumentation, we will assume that that M and N do not have the same metaphysics.

$$\begin{tikzcd}
\mathrm{S} \arrow[r, shift left, "\tau"] \arrow[d] \arrow[dr, dashed] & \mathrm{T} \arrow[l, shift left, "\sigma"] \arrow[d] \arrow[dl, dashed] \\
\mathrm{N} & \mathrm{M}
\end{tikzcd}$$

Figure 5.2: Prima-facie interpretations

This second mapping allows that there is a way to interpret both T and S as committed to the alternative metaphysical assumptions captured by N.

Assuming that there are direct interpretations of both S and T available, the symmetrical fact that their formalisms are mutually recoverable means that their interpretation is systematically underdetermined. Each of T and S is d-interpreted by a model, M and N, respectively. In addition to their direct interpretations, each theory can be indirectly interpreted. But both models have indirect interpretations, as supported by their formal equivalence.

If we take both theories at face value (or interpret both indirectly), S and T are inequivalent under interpretative equivalence. Interpreting only one theory indirectly, by recovering the other theory, means that they are equivalent.

Now the defender of interpretative equivalence is put into an awkward position: they have to explain both why we can interpret one of the theories indirectly, i.e. why its indirect interpretation should be preferred over a face-value understanding.

As I argued above, Coffey cannot simply break the symmetry by insisting that either M or N exhibits the correct metaphysics. In order to explain theoretical equivalence, determining the correct method of interpretation cannot depend on assumptions about the correct metaphysics. We cannot rely on metaphysical considerations to specify a unique interpretation for these theories. Similarly, any *theoretical* reason cannot break this asymmetry: because the theories share their theorems under the translation that witnesses their equivalence, any theoretical consideration in favour of one theory equally supports a formally equivalent theory cf. Coffey, 2014, 839f.

How to correctly interpret the theories is therefore underdetermined by formal equivalence, including claims about what the metaphysics of the theory are. We would need to rely on reasons that are external to the theories to decide which interpretation is correct. Interpretative equivalence then cannot get off the ground, as for formally equivalent theories, no unique interpretation can be provided. *Pace* Coffey, this is not due to the fact that in some cases it is difficult to identify the correct interpretation, but due to the fact that interpretation is systematically underdetermined.

But this means again that interpretation does not explain the equivalence of the theories S and T. Instead, their formal equivalence explains how the theories afford the same interpretations: if a model M d-interprets T, M interprets any formally equivalent theory S, even if the interpretation might be derived in a potentially convoluted way.

On both approaches to interpretation, granting metaphysically correct interpretation and standardising direct interpretation, Coffey's argument against formal equivalence therefore fails.

The argument fails because it assumes that equivalent theories share a unique and specific correct interpretation, at least in a particular context of interpretation. For Coffey, strong realism for which metaphysical commitments can be read off from a privileged formulation is the background assumption which would be challenged by underdetermination (Coffey, 2014, sects. 2, 7).

In the next section, I will present an alternative to this position. Formal equivalence, I will argue, show that interpretation itself is limited. The realist hope that a privileged formulation gives access to the metaphysics of a theory is misguided. However, this does not by itself entail anti-realism. Instead, the adequate reaction that not only the theoretical (e.g. scientific or mathematical) content of theories is language-independent (Visser, 2015, p. 2), but that formal equivalence provides a better understanding of what it means that metaphysics is language independent.[4] Considering the limitation formal equivalence imposes on interpretation provides a better understanding of what we may be realist about.

5.4 Equivalence as limit to interpretation

Interpretation does not provide an adequate basis for theory equivalence in the way Coffey imagines. Interpreting formalisms of theories is an important philosophical and scientific task; it requires additional philosophical and scientific tools to understand what a theory says about the world. However, considerations of formal equivalence demonstrate both limitations on interpretation of theories and limitations to the Quinean approach to (meta-)metaphysics.

Instead of assuming that a particular formulation of a theory will be perspicuous with respect to its metaphysical assumption, and other formulations are to be treated as derivative, we should allow for the possibility that all (equivalent) formulations capture both the scientific and metaphysical content of the theory.

This means to reconsider Coffey's assumption that the direct interpretations of T and S in M and N are metaphysically distinct, i.e. to investigate whether a relation corresponds to the arrows labelled "?" in 5.3.

If results about the formal equivalence of theories are to be taken seriously, the interpretation of a theory is underdetermined by its theoretical content. If there are ways to mutually "recover" the respective formalisms, the theories simply capture the same facts about the target system.[5]

[4] If one accepts this distinction.

[5] I take it as plausible that equivalent theories preserve theoremhood, i.e. that we should at least require that equivalent theories are mutually faithfully interpretable.

$$\begin{tikzcd}
S \arrow[r, shift left, "\tau"] \arrow[d] \arrow[dr, dashed] & T \arrow[l, shift left, "\sigma"] \arrow[d] \arrow[dl, dashed] \\
N \arrow[r, dashed, shift left, "?"] & M \arrow[l, dashed, shift left]
\end{tikzcd}$$

Figure 5.3: Inner models

What does that entail for the prospects of determining a unique set of metaphysical assumptions?

The mereological case discussed in chapter 3 provides a useful basis for sharpening how to respond to that question.

Assume that universalism and nihilism are theoretically equivalent, either in the way Halvorson describes, i.e. by every pair of descriptions of a world with finitely many atoms being bi-interpretable, or in that universalism in general is mutually interpretable with nihilism in a language of plural logic (Halvorson, 2019, pp. 147ff; Warren, 2015, app.). Now, the usual model theoretic semantics entail that these theories have different models – in the case of finite worlds the models differ with respect to the cardinality of the domain.If we take model theoretic semantics as a way of spelling out the metaphysical assumptions of a theory, universalism and nihilism have different metaphysical interpretations.[6]

Under Coffey's approach to interpretation, the following issue arises. We can acknowledge that there is a way in which the formalism of each of the theories can be recovered by the formalism of the other – there are mappings between the languages of the theories that witness that mereological nihilism and universalism are bi-interpretable. As described earlier, both theories also have direct interpretations in their own classes of models.Each of these interpretations, the classes of nihilistic and universalist models, are now available as metaphysical interpretations for both theories. As in Coffey's cases, indirect interpretations are available, supported by the translations that witness the mutual interpretability of the theories.

How are these interpretations related, i.e., how should we think about the relation labelled "?" in 5.3? Understanding direct interpretation in model theoretic terms, the first step of the answer is already given: a translation mapping that witnesses

[6] As mentioned above, model theoretic semantics is one way in which one can describe which metaphysical commitments are carried by a theory by describing the required word-world relations. There are other and potentially more selective ways of describing this relation which amount to different approaches to interpretation.

the relative interpretability of a theory S in a theory T induces inner models for each model of S in models of T (Friedman and Visser, 2014, p. 3).

In the mereological case, the equivalence relation between theories, understood as sets of sentences, determines particular relations between the interpretations of these theories. This point generalises, as long as interpretation is understood in model theoretic terms. Each of the formal relations discussed in chapter 2 provides a way in which models of one theory can be recovered in the models of an equivalent theory.

This has implications for how we understand interpretation (here: the model classes) for both theories. There are three options available. First, one can assert that one of the model classes is the correct metaphysical interpretation for both theories. One would here have to introduce independent reasons to prefer one of the interpretations over the other. The response corresponds to Coffey's solution to single out, if possible, a privileged formulation in cases of equivalent theories in physics.

This solution runs into the problem mentioned above: there cannot be any theoretical reason to prefer one interpretation over the other.

Any reason provided to single one interpretation as correct therefore need to be extra-theoretical, e.g. in the form of aesthetic or pragmatic considerations.

But this response is inadequate for the argumentative purpose at hand. Adducing extra-theoretical considerations amounts to a demand for a stricter notion of theory equivalence for the purpose of metaphysics under which the theories in question are not equivalent. This kind of response therefore would shift the question back from the limits equivalence results impose on interpretation to the question whether the theories are equivalent in the first place.

The second option is to assert that the metaphysical commitments are what is shared between both model classes. Metaphysical questions that are not uniformly decided by both theories would not be part of their metaphysical content. Metaphysical interpretation then is given by the union of both model classes. This would be to simply accept that the theories have the same metaphysical commitments – whatever the metaphysical commitments are is less determined than the face value reading of the theories suggest.

In the mereological case, this would amount to the assertion that both theories are committed to the existence of the atoms, but do not have any commitment to the existence of composite objects. Instead of deciding whether there are composite objects, both theories have the linguistic resources to talk about arbitrary collections of these shared objects, be it as composita, pluralities, or sequences of objects. If we may, we can understand the fact that the theories provide these linguistic resources as the theories being committed to corresponding "logical constructions" (Barrett and Halvorson, 2017a, p. 1056).

Geometry illustrates that this is a plausible suggestion on how equivalent descriptions express the same ontological commitments even if they do not share their base objects (see chapter 4).

Take, for example, a pyramid that is equivalently described with theories in terms of only its nodes, its edges, or the planes of its faces. Theories with these base objects do not have the same surface-level ontological commitments, as they quantify over different sorts of objects. But they describe the same situation: all describe the same pyramid. The geometrical object, the pyramid, is a language-independent ontological commitment of theories. Like Figure 4.1, the pyramid can be constructed using the base sorts of different theories.

One might worry that this example refers to a new sort of object, the pyramid, and thereby suggests a privileged theory in a language that quantifies over pyramids. But this amounts to the first response and faces the same challenges for justifying that the theory is privileged.

In other cases, we have no independent description of the situation in the first place.

Coffey's examples are formally equivalent theories that share their theorems under adequate translation but do not share their base objects. Given the understanding developed in this response, their formal equivalence then means that these theories share their ontological commitments, even if these commitments are not explicitly expressed in the formalism of either theory.[7]

The third alternative holds that both model classes are distinct, but correct interpretations of both theories. The interpretations are separate ways of presenting the ontological commitments, where the inner model constructions show how the presentations are related. This alternative is motivated by the idea that we have to use some theory to express our theories and their metaphysics (Button, 2013, Ch. 19). The equivalence relations between the theories, and the corresponding relations between

[7] Kaveh (2023) provides a similar understanding, according to which the ontology of a theory consists of the objects shared by all models of its equivalence class. In contrast to the discussion above, he not hold that theories can have ontological commitments independent of their base objects. Because each formulation of classical dynamics quantifies over different base objects, they share no objects. From there, Kaveh (2023) concludes classical dynamics has no ontological commitments. Only its predictive power, but not any ontology, should be taken as essential for classical dynamics. Positing different ontologies for different formulations of classical mechanics therefore does not mean that they are genuine alternatives, as they differ only with respect to an inessential part of the theory. On this basis, Kaveh rejects the understanding of (physical) theories as descriptive; his view is beyond the scope of this thesis.

their models, then allows us to still capture how the descriptive content of a theory is independent of the particular language used (cf. Visser, 2015, p. 2).

Under stricter equivalence notions (iso-congruence, bi-interpretability, and synonymy), one can go a step further and hold that the theories do not have different models, as for every model of one theory one can find an isomorphic model of the other theory (cf. Friedman and Visser, 2014, p. 4).

The first response is not promising, for the reasons discussed throughout this chapter. Formal equivalence puts pressure on realists that require a privileged or fundamental formulation of a theory.

The second and third responses illustrate the limitations equivalence put on interpretation. However, they also illustrate why the specific worry that interpretation is underdetermined does not mean that the resulting position is anti-realist: interpretation still describes the metaphysical commitments of a theory. Theoretical terms are interpreted as reflecting language independent objects or structure, as realists demand, and are not simply treated as purely instrumental. What changes, relative to realist pictures that require a privileged formulation, is that the relation between representational apparatus and metaphysical commitments is less direct. But this simply reflects the understanding that choice of language or representational apparatus does not reflect any theoretical content of the theory.

While the resulting views are still realist, there is still a question on how to understand the limitations formal equivalence imposes on interpretation.

I think that the limitation has both a doxastic and, closely related, a metaphysical aspect.

The doxastic implication is that the theories do not make a distinction between the interpretations. While there might be a difference between the models of the theories, this difference cannot be expressed with the conceptual tools provided by the theories. On the basis of the theories, one cannot explain the difference in interpretation, and why we should distinguish the corresponding metaphysical pictures.

Any potential difference between the interpretations is beyond what we can formulate using the conceptual resources available in the theories. So if there is a difference between the metaphysical pictures of these interpretations, one would need to explain them using different or additional theories.

Interpretation of a theory is then not simply empirically underdetermined with respect to some metaphysical issues, as various strands of scientific anti-realism would suggest. Instead, their formal equivalence means that these theories do not answer the putative metaphysical question. As a consequence, the limited interpretation can remain realist, as it can provide answers to all metaphysical questions that are meaningful from the perspective of the theories. If we stop here, formal equivalence describes which questions a realist might have cannot be answered by particular theories.

Alternatively, one may understand the limitations as a positive (meta-)metaphysical stance. In this sense, formal equivalence relations between theories provide a better understanding of the metaphysical commitments of theories. Formal equivalence allows us to describe metaphysical commitments of theories that are not explicitly represented by a privileged formulation.

At that point, one might disagree how to label of these commitments. For example, we might hesitate to call mereological nihilism's commitment to arbitrary collections of objects, if formulated in plural logic with comprehension and witnessed by its bi-interpretability with universalism, an "ontological" commitment, instead of a metaphysical commitment more generally. These terminological questions aside, bi-interpretability shows here that nihilism, at least as discussed in chapter 3 has a metaphysical assumption that corresponds to universalism's commitment to arbitrary sums of objects.

Like in the doxastic understanding of how interpretation is limited, we do not have a case of simple underdetermination. Formal equivalence of theories allows the realist to better understand the language-independent metaphysical assumptions of theories.

5.5 Lessons

Finally, let me summarise where we stand after the discussion in this thesis. Metaphysics is under the suspicion that some debates are purely verbal. Concepts of theoretical equivalence can be leveraged to explicate why one would think that particular metaphysical disputes are merely verbal – theories that are equivalent under an applicable concept of theory equivalence have the same theoretical commitments.

This raises the question which notion of theory equivalence should be used in the context of metaphysics.

This thesis focuses on the idea that metaphysical equivalence might be weaker than logical equivalence, with a focus on the formal equivalence concepts presented in chapter 2. The driving question is to investigate how these equivalence concepts impact our understanding of the metaphysical assumptions and commitments of theories. Choosing a concept of formal equivalence of theories substantially constrains what distinctions in a formalism can be taken to correspond to a distinction made by the theory. Under the concepts discussed, theories with different signatures can have the same metaphysics – equivalent theories that do not share their primitive terms of their base objects might still present the same metaphysical picture.

The case studies in chapters 3 and 4 illustrate how notions of formal equivalence can inform discussions about the metaphysics of a theory. Reasonably strong concepts of theory equivalence, concepts that find application in mathematical contexts, allow for the preservation of all theorems between equivalent theories even if the theories have different model classes. The focus on theory equivalence gives reason to hold that no particular formulation of a theory uniquely describes its metaphysical commitments.

Attempts to instead take interpretation as fundamental concepts and use it to explain theory equivalence are faced with the problem of finding a plausible independent concept of interpretation. This approach shows little promise, as interpretation is limited by formal equivalence relations. As soon as formal equivalence allows for a mutual reconstruction of the theories, and therefore preserves theoremhood between equivalent theories, a theory does not specify a unique metaphysical interpretation that would be required in a successful concept of interpretative equivalence.

Interpreting a theory, describing the relation between its representational apparatus and its target systems, is therefore constrained by formal notions of theory equivalence. These formal notions explicate in which sense both the theoretical content of a theory and its metaphysical commitments are independent of the particular language used to express the theory.

Bibliography

Barrett, Thomas William (2017), *Structure and Equivalence of Theories*, PhD thesis, Princeton University, Princeton, NJ.

– (2019), "Equivalent and Inequivalent Formulations of Classical Mechanics", *British Journal for the Philosophy of Science*, 70, 4, pp. 1167-1199.

Barrett, Thomas William and Hans Halvorson (2016a), "Glymour and Quine on Theoretical Equivalence", *Journal of Philosophical Logic*, 45, 5, pp. 467-483.

– (2016b), "Morita Equivalence", *Review of Symbolic Logic*, 9, 3, pp. 556-582.

– (2017a), "From Geometry to Conceptual Relativity", *Erkenntnis*, 82, 5, pp. 1043-1063.

– (2017b), "Quine's Conjecture on Many-Sorted Logic", *Synthese*, 194, 9, pp. 3563-3582.

– (2022), "Mutual Translatability, Equivalence, and the Structure of Theories", *Synthese*, 200, 3, pp. 1-36.

Belleri, Delia (2017), "Verbalism and Metalinguistic Negotiation in Ontological Disputes", *Philosophical Studies*, 174, 9, pp. 2211-2226.

Button, Tim (2013), *The Limits of Realism*, Oxford.

– (2022), "Symmetric Relations, Symmetric Theories, and Pythagrapheanism", *Philosophy and Phenomenological Research*.

Button, Tim and Sean P. Walsh (2018), *Philosophy and Model Theory*, ed. by Sean Walsh and Wilfrid Hodges, Oxford, UK.

Carnap, Rudolf (1950), "Empiricism, Semantics and Ontology", *Revue Internationale de Philosophie*, 4, 11, pp. 20-40.

– (1963), "Replies and Systematic Expositions", in *? Iteschilpp:Prc*, ed. by Paul Arthur Schilpp, pp. 859-1013.

Coffey, Kevin (2014), "Theoretical Equivalence as Interpretative Equivalence", *British Journal for the Philosophy of Science*, 65, 4, pp. 821-844.

De Sa, Dan López (2007), "The Many Relativisms and the Question of Disagreement", *International Journal of Philosophical Studies*, 15, 2, pp. 269-279.

– (2015), "Expressing Disagreement: A Presuppositional Indexical Contextualist Relativist Account", *Erkenntnis*, 80, 1, pp. 153-165.

Dewar, Neil (2023), "Interpretation and Equivalence; or, Equivalence and Interpretation", *Synthese*, 201, 4, pp. 1-24.

Enderton, Herbert B. (2001), *A Mathematical Introduction to Logic (Second Edition)*, ed. by Herbert B. Enderton, Second Edition, Academic Press, Boston.

Florio, Salvatore (2023), "On Type Distinctions and Expressivity", *Proceedings of the Aristotelian Society*, 123, 2, pp. 150-172.

French, Rohan (2019), "Notational Variance and its Variants", *Topoi*, 38, 2, pp. 321-331.

Friedman, Harvey M. and Albert Visser (2014), "When Bi-Interpretability Implies Synonymy", *Logic group preprint series*, 320, pp. 1-19.

Glymour, Clark (1970), "Theoretical Realism and Theoretical Equivalence", *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1970, pp. 275-288.

– (2013), "Theoretical Equivalence and the Semantic View of Theories", *Philosophy of Science*, 80, 2, pp. 286-297.

Halvorson, Hans (2013), "The Semantic View, If Plausible, is Syntactic", *Philosophy of Science*, 80, 3, pp. 475-478.

– (2019), *The Logic in Philosophy of Science*.

– (n.d.), "Quantifier Variance Without Collapse".

Haslanger, Sally (2000), "Gender and Race: (What) Are They? (What) Do We Want Them to Be?", *Noûs*, 34, 1, pp. 31-55.

Hicks, Michael Townsen and Jonathan Schaffer (2017), "Derivative Properties in Fundamental Laws", *British Journal for the Philosophy of Science*, 68, 2.

Hirsch, Eli (2002), "Quantifier Variance and Realism", *Philosophical Issues*, 12, 1, pp. 51-73.

– (2009), "Ontology and Alternative Languages", in *Metametaphysics: New Essays on the Foundations of Ontology*, ed. by David Chalmers, David Manley, and Ryan Wasserman, pp. 231-58.

Kaveh, Shahin (2023), "Physical Theories Are Prescriptions, Not Descriptions", *Erkenntnis*, 88, 5, pp. 1825-1853.

King, Jeffrey C. (2019), "Structured Propositions", in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Summer 2019.

King, Jeffrey C., Scott Soames, and Jeff Speaks (2014), *New Thinking About Propositions*, ed. by Scott Soames and Jeffrey Speaks, New York, NY, USA.

Knox, Eleanor (2014), "Newtonian Spacetime Structure in Light of the Equivalence Principle", *British Journal for the Philosophy of Science*, 65, 4, pp. 863-880.

Lefever, Koen and Gergely Székely (2019), "On Generalization of Definitional Equivalence to Non-Disjoint Languages", *Journal of Philosophical Logic*, 48, 4, pp. 709-729.

Linnebo, Øystein and Agustin Rayo (2012), "Hierarchies Ontological and Ideological", *Mind*, 121, 482, pp. 269-308.

McEldowney, Paul Anh (2020), "On Morita Equivalence and Interpretability", *Review of Symbolic Logic*, 13, 2, pp. 388-415.

McGrath, Matthew and Devin Frank (2020), "Propositions", in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Winter 2020.

McSweeney, Michaela Markham (2016), "An Epistemic Account Of Metaphysical Equivalence", *Philosophical Perspectives*, 30, 1, pp. 270-293.

– (2019), "Following Logical Realism Where It Leads", *Philosophical Studies*, 176, 1, pp. 117-139.

Meadows, Toby (2021), "Relative Interpretation Between Logics", *Erkenntnis*, pp. 1-18.

Miller, Kristie (2005), "What is Metaphysical Equivalence?", *Philosophical Papers*, 34, 1, pp. 45-74.

– (2014), "Defending Substantivism About Disputes in the Metaphysics of Composition", *Journal of Philosophy*, 111, 9-10, pp. 529-556.

– (2017), "A Hyperintensional Account of Metaphysical Equivalence", *Philosophical Quarterly*, 67, 269, pp. 772-793.

Mossakowski, Till, Răzvan Diaconescu, and Andrzej Tarlecki (2009), "What is a Logic Translation?", *Logica Universalis*, 3, 1, pp. 95-124.

Niebergall, Karl-Georg (2000a), "On the Logic of Reducibility: Axioms and Examples", *Erkenntnis*, 53, 1-2, pp. 27-61.

– (2000b), "On the Logic of Reducibility: Axioms and Examples", *Erkenntnis*, 53, 1-2, pp. 27-61.

Pelletier, Francis Jeffry and Alasdair Urquhart (2003), "Synonymous Logics", *Journal of Philosophical Logic*, 32, 3, pp. 259-285.

Potter, Michael (1998), "Classical Arithmetic is Part of Intuitionistic Arithmetic", *Grazer Philosophische Studien*, 55, 1, pp. 127-41.

Putnam, Hilary (1987a), *Representation and Reality*.

– (1987b), "Truth and Convention: On Davidson's Refutation of Conceptual Relativism", *Dialectica*, 41, 1-2, pp. 69-77.

Quine, W. V. (1956), "Unification of Universes in Set Theory", *Journal of Symbolic Logic*, 21, 3, pp. 267-279.

Sambrotta, Mirco (2019), "Scientific Models and Metalinguistic Negotiation", *Theoria. An International Journal for Theory, History and Foundations of Science*, 34, 2, p. 277.

Schwabhäuser, W., W. Szmielew, and A. Tarski (1983), *Metamathematische Methoden in der Geometrie*, Hochschultext, Springer, Berlin, Heidelberg, ISBN: 9783642694189.

Sider, Theodore (2011), *Writing the Book of the World*.

Stalnaker, Robert C. (1976), "Propositions", in *Issues in the Philosophy of Language: Proceedings of the 1972 Colloquium in Philosophy*, ed. by Alfred F. MacKay and Daniel D. Merrill, pp. 79-91.

Szczerba, L. W. (1977), "Interpretability of Elementary Theories", in, *Logic, Foundations of Mathematics, and Computability Theory: Part One of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975*, ed. by Robert E. Butts and Jaakko Hintikka, Dordrecht, pp. 129-145.

Tran-Hoang, Paul Anh (2021), "On the Virtue of Categoricity", *Notre Dame Journal of Formal Logic*, 62, 1, pp. 107-146.

Tsementzis, Dimitris (2017), "A Syntactic Characterization of Morita Equivalence", *Journal of Symbolic Logic*, 82, 4, pp. 1181-1198.

Unger, Peter (1980), "The Problem of the Many", *Midwest Studies in Philosophy*, 5, 1, pp. 411-468.

Van Inwagen, Peter (1987), "When Are Objects Parts?", *Philosophical Perspectives*, 1, pp. 21-47.

– (1990), *Material Beings*, Ithaca.

Varzi, Achille (2019), "Mereology", in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Spring 2019.

Varzi, Achille and A. J. Cotnoir (2021), *Mereology*, Oxford.

Vetter, Barbara (2015), *Potentiality: From Dispositions to Modality*, Oxford, England and New York, NY, USA.

Visser, Albert (2006), "Categories of theories and interpretations", in *Logic in Tehran*, ed. by Ali Enayat, Iraj Kalantari, and Mojtaba Moniri, Lecture Notes in Logic, Cambridge University Press, pp. 284-341.

– (2015), "Extension and Interpretability", English, Logic Group preprint series, 329.

Warren, Jared (2015), "Quantifier Variance and the Collapse Argument", *Philosophical Quarterly*, 65, 259, pp. 241-253.

Weatherall, James Owen (2019), "Part 2: Theoretical Equivalence in Physics", *Philosophy Compass*, 14, 5.

Wiggins, David (1968), "On Being in the Same Place at the Same Time", *Philosophical Review*, 77, 1, pp. 90-95.

Wigglesworth, John (2017), “Logical Anti-Exceptionalism and Theoretical Equivalence”, *Analysis*, 77, 4, pp. 759-767.

Williamson, Timothy (2013), *Modal Logic as Metaphysics*, Oxford, England.

Woods, Jack (2018), “Intertranslatability, Theoretical Equivalence, and Perversion”, *Thought: A Journal of Philosophy*, 7, 1, pp. 58-68.