Theoretical Equivalence as Representational Equivalence

MSc Thesis *(Afstudeerscriptie)*

written by

Minzhe Li

under the supervision of **dr. Benno van den Berg** and **dr. Sebastian De Haro Ollé**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
June 25th, 2025

Members of the Thesis Committee:
dr. Benno van den Berg
dr. Sebastian De Haro Ollé
prof. dr. Sonja Smets
prof. dr. Albert Visser

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

This thesis explores the relationship between theoretical equivalence and representational equivalence. It mainly contains three parts: the first part summarizes current criteria of theoretical equivalence and proves some results about their comparative strengths; the second part develops a formal framework of representation and shows how formal criteria of theoretical equivalence are related to representational equivalence; the third part applies the framework to the hole argument in philosophy of physics, and gives a critical evaluation of the formalist response to the hole argument.

Acknowledgements

The thesis evolved from a winter project I completed in January 2024, which explored duality and theoretical equivalence under the supervision of Dr. Sebastian De Haro Ollé. He then encouraged me to further investigate this topic in my master's thesis. The basic idea of this thesis, i.e., to study the relationship between theoretical equivalence and representational equivalence, was formed during the summer of 2024, and an early proposal was written in September. Dr. Sebastian De Haro Ollé supervised the philosophy of physics part of the thesis, while Dr. Benno van den Berg supervised the logic part.

I would like to firstly express my sincere gratitude to my supervisors. Sebastian introduced me to the topic of theoretical equivalence and gave invaluable comments which pushed me to reflect further upon the weaknesses in my reasoning, particularly the distinction between metalanguage vs. object language and the relationship between pessimistic meta-induction and matter-of-fact naturalism. He also influenced the overall structure of the thesis, emphasizing the importance of a formal and rigorous presentation of anti-specific formalism of General Relativity. Benno, on the other hand, read fifty pages of rather primitive and chaotic notes I wrote during the early stages of my study. He spotted various mistakes I made throughout the proofs, and pushed me to think further about the justification of Morita equivalence, which eventually led to the result I proved in Appendix C. He also gave very helpful comments on the details of formalization such as the sorting of variables and the treatment of equality.

I would also like to give thanks to many other people who offered critical help during the process of writing this thesis. I would like to thank Professor Toby Meadows for correspondence concerning issues addressed in Section 2.5.1 and Professor Hajnal Andréka for correspondence concerning issues addressed in Appendix C. I would like to thank Professor Albert Visser for spotting many mistakes in the thesis and raising insightful advice for further research directions concerning interpretation. I would like to thank my classmate and roommate Ruiting Jiang, who read many of my early drafts and offered inspiring suggestions during our daily conversations, especially concerning the proof of Proposition 2.48. I also need to express my gratitude to my academic mentor, Professor Dick de Jongh. He listened to my early ideas about the thesis during the early summer of 2024, and gave very helpful information and advice for the timeline of finishing the thesis.

Finally, I need to thank my family for their unconditional support. This thesis would not have been possible without their encouragement and understanding during the challenging times of my research.

Contents

Chapter 1

Introduction

When are two scientific theories equivalent? This question is at the heart of recent debates in philosophy of science. A general criterion of theoretical equivalence will be able to tell us when two scientific theories are equivalent, and thus obviate the need to conduct further physical or philosophical investigations to choose from equivalent theories. It will also have significant consequences in the long-standing debate between the syntactic versus semantic conceptions of theories, i.e., the debate whether theories should be viewed syntactically, as sets of sentences, or semantically, as classes of models.[1] As argued in Halvorson (2012), if the criterion of theoretical equivalence is given in, say, syntactical rather than semantical terms, then it seems that we have a reason to prefer the syntactical view as it is able to characterize scientific theories up to equivalence and vice versa.

People have proposed various formal criteria as the standard of theoretical equivalence, including isomorphism, (first-order, single-sorted) definitional equivalence (Glymour, 1970, 1977, 1980), Quine equivalence (Quine, 1975), Morita equivalence (Barrett and Halvorson, 2016), categorical equivalence (Weatherall, 2016a, 2016b, 2017), definable categorical equivalence (Hudetz, 2019), etc. But, as many people complained,[2] scientific theories are not standalone formal structures. *They are used to say things about the world*, or, to *represent the world.* Thus, a purely formal characterization cannot fully capture the notion of theoretical equivalence. Consider the following example in classical mechanics (Taylor, 2005, p.173). The following equation in Newtonian mechanics describes the motion of a cart with mass m, subjected to a resistive force $f = -b\dot{x}$, and linked by a spring that exerts an elastic force $-kx$:

$$m\ddot{x} + b\dot{x} + kx = 0.$$

But precisely the same equation can also be used to represent a completely physical domain. Specifically, the movement of charges in a circuit with an inductor (inductance L), a capacitor (capacitance C), and a resistor (resistance R) can be described by the same equation if we interpret m as L, b as R, k as $1/C$, and x as the charge q.

$$L\ddot{x} + R\dot{x} + \frac{1}{C}x = 0.$$

Despite sharing the same formal structure, the two equations really say different things about the world.

Although people have recognized this problem in general, the proposed remedy has been to either consider *only* interpretative or representational equivalence (Coffey, 2014; Nguyen, 2017; Teitel, 2021), or stick to the formal criteria and add a requirement regarding interpretational or representational perspective, such as empirical equivalence, *in addition to* the formal requirement (Halvorson, 2019; Weatherall, 2021). The first option fixes the problem, but precisely as the criteria of interpretational or representational equivalence are informal, we cannot study them using tools of mathematics and

[1] See Halvorson (2012, 2013), Glymour (2013) for typical syntactical views of theories; See Suppes (2002), van Frassen (2008, 2014) for typical smentical views of theories. See Luz (2017) and Frigg (2022) for an overview of the debate.

[2] See Sklar (1982), Coffey (2014), Nguyen (2017), Teitel (2021), De Haro and Butterfield (2021).

formal logic, and we lack a rigorous method to apply such criteria in practice. However, while adding an extra interpretive requirement to the formal criteria may also solve the problem, it becomes unclear what the relationship between the formal and the representational criterion is, in particular, how the former relates to the latter. When we ask whether two theories are equivalent, we ultimately want to ask if they say the same thing about the world, but not whether they possess the same formal structure. But if we are ultimately asking a representational question, what guarantees that formal criteria will be of any use? Or, more specifically, how do they contribute to the representational equivalence that concerns us?

Thus motivated, this thesis approaches the question of theoretical equivalence from the perspective of representation. In Chapter 2, I summarize the current criteria of theoretical equivalence that have been proposed in the literature and prove some results about their comparative strengths that were not known before. Then in Chapter 3, I develop a formal framework of representation based on the idea of *representation as definition*. We analyze the structure of representation, define several relevant notions of equivalence with respect to representation, and then show how some of the previous formal criteria of theoretical equivalence correspond to certain species of representational equivalence. Chapter 4 discusses the implications of this framework in the debate about theoretical equivalence (in a relatively sketchy manner due to the space limit). Chapters 5, 6, and 7 then apply the framework of representation to a specific debate in the philosophy of physics, namely, the hole argument. More specifically, Chapter 5 introduces a formalism called *restricted set-theoretical language*, which allows us to schematically assign a canonical language to bare set-theoretical structures. This makes the formal framework of "representation as definition" developed in Chapter 3 applicable to physical models, which are typically bare set-theoretical structures. A specific restricted set-theoretical language $\mathcal{L}_M$ is also designed for Lorentzian manifolds. This gives a formalism of GR whose isomorphism criterion is precisely isometry, and is used in discussions of the hole argument in Chapters 6 and Chapter 7. Chapter 6 summarizes the hole argument and evaluates the formalist response. It is argued that while many defects of the formalist response can be remedied if we base the response on the correct notion of representational equivalence that is relevant there, there is still a critical assumption of the formalist response that remains unjustified. Chapter 7 then suggests a possible solution to the hole argument inspired by the previous evaluation of the formalist response, which avoids the commitment to the critical assumption.

The thesis also contains three appendices. Appendix A addresses the technical issue of non-disjoint languages in theoretical equivalence. Appendix B discusses the alternative definition of Morita extension given in Meadows (2024). Appendix C gives a proof of Beth's definability theorem for many-sorted logic, as an attempt to improve the previous result by Andreka, Madarasz, and Nemeti (2008, Theorem 2.5.1).

Chapter 2

Theoretical Equivalence

2.1 Introduction

This chapter summarizes current criteria of theoretical equivalence that have been proposed in the literature, compares their differences, and proves certain results with respect to their relative strengths which were not known before.

We classify all these criteria into three groups according to the formats in which these criteria are defined: the common-extension criteria, the coordinated-interpretation criteria, and the categorical criteria.

We use $\cong$ for isomorphism of models, $\equiv$ for logical equivalence. For a theory T, we also use $\mathcal{L}_T$ to denote its language, and Σ_T to denote its signature. $\overline{x}$ denotes a sequence of variables, and $\overline{x}^{\overline{\sigma}}$ denotes a sequence of sorted variables where $\overline{\sigma}$ is the corresponding sequence of sorts.

We use $\mathcal{M}|_\Sigma$ to denote the reduct of $\mathcal{M}$ to the language $\mathcal{L}_\Sigma$, and $T|_\Sigma$ to denote the set of all Σ-sentences entailed[1] by T.

Sometimes, instead of giving definitions, we give schemas for definitions. We will use bold font to emphasize that certain terms are meant to be placeholders for several different technical terms.

2.2 The Common-Extension Criteria

Common-extension criteria are given through the concept of "definitional extension". The basic idea is that a definitional extension of a model or a theory only adds new symbols that are defined out of the old symbols, and hence really adds nothing over and above the structure of the original model or theory. In other words, the extra structures in the extended theories or models are "free lunches", as they can already be defined from the original ones. Thus, if two theories/models can be made the same (modulo logical equivalence/isomorphism) by doing definitional extensions, then according to criteria in this group, we should really say that they are equivalent. As a schema,

Schema 2.1. Two theories/models are (common-extensionally) equivalent if they have a common **definitional extension** (modulo logical equivalence/isomorphism).

Different species of common-extensional criteria are obtained by plugging in different technical notions for the placeholder "definitional extension", i.e., by specifying what kind of structures are taken to be "free lunches".

2.2.1 Standard Definitional Equivalence

We first present the most standard notion of definitional extension below, where we are only allowed to define constants, functional and relational symbols, but not sort symbols. We refer to this as the

[1] Here we use the semantic notion of entailment as it is almost always available: ϕ is entailed by T if for every model $\mathcal{M}$ that satisfies T, $\mathcal{M}$ also satisfies ϕ.

standard definitional extension.

Definition 2.2. Let T (respectively, $\mathcal{M}$) be a first-order theory (model) of language $\mathcal{L}$. An *(explicit) definition* Φ_α of a symbol α by T ($\mathcal{M}$) is given as follows:

- if α is a relational symbol R, then $\Phi_\alpha = \forall \overline{x}(R(\overline{x}) \leftrightarrow \phi_R(\overline{x}))$, where $\overline{x}$ is a sequence of variables with the same length as the arity of R, and $\phi_R(\overline{x})$ is a formula in $\mathcal{L}$;
- if α is a constant symbol c, then $\Phi_\alpha = \forall x(x = c \leftrightarrow \phi_c(x))$, where $\phi_c(x)$ is a formula in $\mathcal{L}$, and,
  - T satisfies the admissibility condition, $T \vdash \exists! x(\phi_c(x))$;
  - ($\mathcal{M}$ satisfies the admissibility condition, $\mathcal{M} \models \exists! x(\phi_c(x))$);
- if α is a function symbol f, then $\Phi_\alpha = \forall \overline{x} \forall y(f(\overline{x}) = y \leftrightarrow \phi_f(\overline{x}, y))$, where $\overline{x}$ is a sequence of variables with the same length as the arity of f, $\phi_f(\overline{x}, t)$ is a formula in the language of T, and,
  - T satisfies the admissibility condition, $T \vdash \forall \overline{x} \exists! y(\phi_f(\overline{x}, y))$;
  - ($\mathcal{M}$ satisfies the admissibility condition, $\mathcal{M} \models \forall \overline{x} \exists! y(\phi_f(\overline{x}, y))$);

Then we define the standard definitional extension of a theory/model.

Definition 2.3. Let T be a theory in signature Σ and T^+ be a theory in signature Σ^+ such that $\Sigma \subseteq \Sigma^+$. We say that T^+ is a *standard definitional extension* of T if there is a set of definitions for each symbol in $\Sigma^+ - \Sigma$ by T, say, $\Phi = \{\Phi_\alpha | \alpha \in \Sigma^+ - \Sigma\}$, such that T^+ is logically equivalent to $T \cup \Phi$.

Using a convention, we will use Φ for a set of definitions, Φ_s for the specific definition of the symbol s in Φ, and ϕ_s the formula equated with s in the definition Φ_s.

Definition 2.4. Let $\mathcal{M}$ be a model in signature Σ and $\mathcal{M}^+$ a model in signature Σ^+ such that $\Sigma \subseteq \Sigma^+$. We say that $\mathcal{M}^+$ is a *standard definitional extension* of $\mathcal{M}$ if there is a set of definitions Φ for each symbol in $\Sigma^+ - \Sigma$ by $\mathcal{M}$ such that:

- $\mathcal{M}^+|_\Sigma \cong \mathcal{M}$, and,
- $\mathcal{M}^+ \models \Phi_\alpha$ for each $\alpha \in \Sigma^+ - \Sigma$.

We have the following theorem (Corollary 4.6.13 in Halvorson (2019)),

Theorem 2.5. *Let T be a theory in signature Σ^+ and T a theory in signature Σ such that $\Sigma \subseteq \Sigma^+$. If T^+ is a definitional extension of T, then T^+ is conservative over T in the sense that, for any Σ-formula ϕ, $T^+ \vdash \phi$ if and only if $T \vdash \phi$.*

The similar theorem holds for models as well.

Theorem 2.6. *If $\mathcal{M}^+$ is a definitional extension of $\mathcal{M}$, then $\mathcal{M}^+$ is conservative over $\mathcal{M}$ in the sense that, for any Σ-formula ϕ, $\mathcal{M}^+, \overline{a} \models \phi$ if and only if $\mathcal{M}, \overline{a} \models \phi$.*

This is simply a consequence of the coincidence lemma.

Definition 2.7. We say that two theories/models with disjoint signatures are *definitionally equivalent* if they have a common definitional extension.

Note that for a theory T^+ to be a definitional extension of T, we only require that T^+ be logically equivalent to $T \cup \Phi$, but not identical. Equally, we only require $\mathcal{M}^+|_\Sigma$ to be isomorphic to $\mathcal{M}$, but not identical. If we require $T^+ = T \cup \Phi$ and $\mathcal{M}^+|_\Sigma = \mathcal{M}$ instead, then the definition of definitional equivalence needs to be changed to saying that two theories/models have a common definitional extension *up to logical equivalence/isomorphism*, so that we guarantee that logically equivalent theories or isomorphic models are always considered definitionally equivalent.

This is the standard presentation of definitional equivalence. We define two further notions for future use.

Definition 2.8. Let T be a theory in signature Σ. A *potential definition* of T is a tuple $\langle a, \phi \rangle$ such that one of the following three holds:

- $a = n$, and ϕ is a Σ-formula with n distinct free variables;
- $a = 0 \to 1$, and ϕ is a Σ-formula with one free variable and $T \models \exists! x(\phi(x))$;
- $a = n \to 1$, and ϕ is a Σ-formula with n free variables and $T \models \forall \overline{x} \exists! y(\phi(\overline{x}, y))$.

Intuitively, ϕ is a formula that can potentially be used for definition and a the arity of the symbol that ϕ may define.

Definition 2.9. Let T^+ be a definitional extension of T, namely $T^+ = T \cup \Phi$ for some set of definitions Φ. Let s be a symbol in the signature of $T^+ - T$. We say that s *is defined by a potential definition* $\langle a, \phi \rangle$ *of* T, if Φ_s defines s as a symbol with arity a by the formula ϕ.

We note two features of this criterion.

1. The definitional equivalence is only defined for theories with disjoint signatures.
2. The definitional extension is limited to definitions of non-logical symbols.

Item 1 is addressed in Appendix A, and item 2 is left for future work. For the following discussion, we will always assume that it suffices to give a criterion of equivalence for theories with disjoint signatures and we will restrict ourselves to cases where the logic is fixed.

2.2.2 Morita Equivalence

Definitional equivalence is mainly designed for first-order single-sorted logic, where we understand "definitional extension" as first-order definability. In the context of many-sorted logic, we can still have the standard definitional extension and consequently the standard definitional equivalence, as long as we further require that definitions of symbols preserve arities. However, many people think that in such contexts it is natural to give a more general notion of definitional extension where we are allowed to define not only relation symbols, function symbols, and constants, but also sort symbols . Thus, we have the notion of *Morita extension*, which generalizes the standard notion of definitional extension. The notion was first proposed and studied in Andréka, Madarász and Németi (2008) under the name "generalized definitional extension"; Barrett and Halvorson (2016) then introduced it to the philosophical discussion of theoretical equivalence under the name "Morita extension".

We present the version of Morita equivalence in Barrett and Halvorson (2016) as follows:

Definition 2.10. Let T (respectively, $\mathcal{M}$) be a many-sorted first-order theory (model) in signature Σ. Let σ be a new sort symbol. We say that Φ_σ is a *definition of σ in T* (respectively $\mathcal{M}$) if one of the following holds:

- Product:
$$\Phi_\sigma^{prod} := \forall x^{\sigma_0}\forall y^{\sigma_1}\exists! z^\sigma(\pi_0(z) = x^{\sigma_0} \wedge \pi_1(z) = y^{\sigma_1});$$
where π_0 and π_1 are functions with arity $\sigma \to \sigma_0$ and $\sigma \to \sigma_1$ and σ_0 and σ_1 are sort symbols in Σ;

- Coproduct
$$\Phi_\sigma^{cop} := \forall z^\sigma(\exists x^{\sigma_0}(z^\sigma = p_0(x^{\sigma_0})) \vee \exists x^{\sigma_2}(z^\sigma = p(x^{\sigma_1}))) \wedge \forall x^{\sigma_0}\forall y^{\sigma_1}(p_0(x^{\sigma_0}) \neq p_1(y^{\sigma_1}))$$
where p_0 and p_1 are functions with arity $\sigma_0 \to \sigma$ and $\sigma_1 \to \sigma$ and σ_0 and σ_1 are sort symbols in Σ;

- Subsort
$$\Phi_\sigma^{sub} := \forall x^{\sigma_0}(\phi_\sigma(x^{\sigma_0}) \leftrightarrow \exists y^\sigma(\pi(y^\sigma) = x^{\sigma_0})) \wedge \forall x^{\sigma_0}\forall y^{\sigma_0}(\pi(x^{\sigma_0}) = \pi(y^{\sigma_0}) \to x^{\sigma_0} = y^{\sigma_0})$$
where σ is a new sort symbol, π is a function with arity $\sigma \to \sigma_0$, σ_0 a sort symbol in Σ, and T proves that $\exists x^{\sigma_0}\phi_\sigma(x^{\sigma_0})$ (respectively, $\phi_\sigma(x^{\sigma_0})$ is non-empty in $\mathcal{M}$);

- Quotient
$$\Phi_\sigma^{quo} := \forall x^{\sigma_0}\forall y^{\sigma_0}(\pi(x^{\sigma_0}) = \pi(y^{\sigma_0}) \leftrightarrow \phi_\sigma(x^{\sigma_0}, y^{\sigma_0})) \wedge \forall z^\sigma\exists x^{\sigma_0}(z^\sigma = \pi(x^{\sigma_0}))$$
where ϕ_σ is a Σ-formula with at most two free variable of sort σ_0, and T proves that ϕ_σ gives an equivalence relation (respectively, ϕ_σ defines an equivalence relation).

Definition 2.11. Let T (respectively, $\mathcal{M}$) be a many-sorted first-order theory (model) in signature Σ and T^+ ($\mathcal{M}^+$) be a many-sorted first-order theory in signature Σ^+ such that $\Sigma \subseteq \Sigma^+$. We say that T^+ ($\mathcal{M}^+$) is a *Morita extension* of T ($\mathcal{M}$) if there is a set of definitions Φ for each symbol in $\Sigma^+ - \Sigma$ by T ($\mathcal{M}$) such that:

- for any new function symbol f appearing in the definition of a new sort symbol σ, $\Phi_f = \Phi_\sigma$, and
- $T^+ \equiv T \cup \Phi$ (respectively, $\mathcal{M}^+|_\Sigma \cong \mathcal{M}$ and $\mathcal{M}^+ \models \Phi$).

Definition 2.12. We say that T is a *Morita descendant* of T' if there is a finite sequence of Morita extensions $T, ..., T_n$ such that $T_n \equiv T'$.

Similarly, we say that $\mathcal{M}$ is a *Morita descendant* of $\mathcal{M}'$ if there is a finite sequence of Morita extensions $\mathcal{M}, ..., \mathcal{M}_n$ such that $\mathcal{M}_n \cong \mathcal{M}'$.

Definition 2.13. We say that two theories T and T' (or two models $\mathcal{M}$ and $\mathcal{M}'$) are Morita equivalent if they have a common Morita descendant.

Again, for later use, we define the notion of a potential Morita definition.

Definition 2.14. Let T be a theory in signature Σ. A *potential Morita definition of T* is a tuple $\langle a, \phi \rangle$ such that one of the following holds:

- $a = \langle \sigma_0, ..., \sigma_n \rangle$, and ϕ is a Σ-formula with n free variables of sorts $\sigma_0, ..., \sigma_n$;
- $a = \sigma$, and ϕ is a Σ-formula with one free variable of sort σ and $T \models \exists! x^{\sigma}(\phi(x^{\sigma}))$;
- $a = \langle \sigma_0, ..., \sigma_n \rangle \to \sigma$, and ϕ is a Σ-formula with n free variables of sorts $\sigma_0, ..., \sigma_n$ and one free variable of sort σ, and $T \models \forall \overline{x} \exists! y^{\sigma}(\phi(\overline{x}, y))$.
- $a = \sigma_i \times \sigma_j$, $\phi = \top$.[2]
- $a = \sigma_i + \sigma_j$, $\phi = \top$.
- $a = \sigma_i$, ϕ is a Σ-formula with one free variable of sort σ_i, and $T \models \exists x^{\sigma_i}(\phi(x^{\sigma_i}))$.
- $a = \sigma_i$, ϕ is a Σ-formula with two free variables of sort σ_i, and T entails that ϕ defines an equivalence relation.

Definition 2.15. Let T^+ be a Morita extension of T, namely $T^+ = T \cup \Phi$ for some set of definitions Φ. Let s be a symbol in the signature of $T^+ - T$. We say that *s is defined by a potential Morita definition $\langle a, \phi \rangle$ of T*, if

- Φ_s defines s as a constant, relational or functional symbol with arity a by the formula ϕ;
- Φ_s defines s as a new product of σ_0 and σ_1, where $a = \sigma_0 \times \sigma_1$;
- Φ_s defines s as a new coproduct of σ_0 and σ_1, where $a = \sigma_0 + \sigma_1$;
- Φ_s defines s as a new subsort of $a = \sigma_0$ by the domain formula ϕ;
- Φ_s defines s as a new quotient of $a = \sigma_0$ with ϕ being the equivalence relation.

It may be worried that there is some arbitrariness in the notion of Morita extension as we have chosen particularly to allow definitions of sort symbols by product, coproduct, subsort, and quotient. Theorem 2.5.1 by Andréka, Madarász and Németi (2008) shows that Morita extension coincides with implicit definability in many-sorted logic, and hence is a natural criterion to consider.

Still, it can be weakened by rejecting any of these four operations as acceptable methods to define new sorts. But notice that in the many-sorted context, at the very least we should be allowed to define sort symbols by primitive sort symbols: substituting a sort symbol with another would be a case of notational variant and we would certainly want to count notational variants as cases of theoretical equivalence.

Thus, we give the following notion of primitive Morita extension.

Definition 2.16. Let T (respectively, $\mathcal{M}$) be a many-sorted first-order theory (model) in signature Σ. Let σ be a new sort symbol. We say that Φ_σ is a *primitive definition of σ by T (respectively $\mathcal{M}$)* if there is a sort symbol σ_0 in Σ such that $\Phi_\sigma := \forall x^{\sigma_0}(\exists y^{\sigma}(\pi(y^{\sigma}) = x^{\sigma_0}) \wedge \forall x^{\sigma_0} \forall y^{\sigma_0}(\pi(x^{\sigma_0}) = \pi(y^{\sigma_0}) \to x^{\sigma_0} = y^{\sigma_0}))$ where π is a function with arity $\sigma \to \sigma_0$.

[2] For the case of product and coproduct, no associated formulas are needed. We include the argument ϕ here only for the sake of uniformity.

The notion of primitive Morita extension is then defined as an extension that only allows primitive definitions. Consequently, the notion of primitive Morita equivalence can also be defined. In the following, we use the superscript to specify which operations we allow in a certain variant of Morita extension. For instance, we may use the notation "Morita^{p,c,s,q} extension" for the standard Morita extension, and "Morita$^{\emptyset}$ extension" for the primitive Morita extension.

2.3 The Coordinated-Interpretation Criteria

The fundamental concept for the group of coordinated-interpretation criteria is "interpretation" or "translation". The basic idea is that two theories are equivalent if they can mutually interpret or translate what the other is talking about, and sometimes a further requirement (which we may call "the requirement of coordination") is that they can "check" that the other's interpretation or translation is "correct". Different criteria are then given based on different understandings of "interpretation" or "translation", and different requirement of coordination. For convenience, we will mainly focus on the case of theories. Similar coordinated-interpretation criteria can also be given for classes of models as in Hodges (1993, Section 5.4)

Normally, people do not distinguish between interpretation and translation. However, there is a good sense in which these two notions could be distinguished. In the following, we use *interpretation* for cases where a theory is included in some definitional extension of another theory, and *translation* for cases where we have a mapping of formulas that preserves theoremhood of theories. Both translation and interpretation can be seen as a way by which one theory/model "understands" another, and in many cases they coincide: whenever theory T_1 is included in a definitional extension of another theory T_2, we have a natural translation manual from T_1 to T_2, and vice versa.

We first present the popular notion of "relative interpretation", as given in Meadows (2024). Since the definition is given in the form of translation, we use the name "relative translation" instead.[3]

Definition 2.17. Let T_1 and T_2 be two theories in signature Σ_1 and Σ_2. We say that t is a *relative translation* from $\mathcal{L}_{\Sigma_1}$ to $\mathcal{L}_{\Sigma_2}$ in T_2 if

- for each sort symbol $\sigma \in \Sigma_1$, $t(\sigma)$ is a sort symbol in Σ_2, and $\delta_{t,\sigma}$ is a Σ_2-formula with at most one free variable $x^{t(\sigma)}$;

- for each relational symbol R of arity $\langle \sigma_0, ..., \sigma_n \rangle$, $t(R)$ is a Σ_2-formula with at most free variables $x^{t(\sigma_0)}, ..., x^{t(\sigma_n)}$;

- for each function symbol f of arity $\langle \sigma_0, ..., \sigma_n \rangle \to \sigma$, $t(f)$ is a Σ_2-formula with at most free variables $t(x^{\sigma_0}), ..., t(x^{\sigma_n}, y^{\sigma})$, and that T_2 satisfies the admissibility condition: $T_2 \vdash \forall x^{t(\sigma_0)} ... \forall x^{t(\sigma_n)} (\delta_{t,\sigma_0}(x^{t(\sigma_0)}) \wedge ... \wedge \delta_{t,\sigma_n}(x^{t(\sigma_n)}) \to \exists ! y^{t(\sigma)} (\delta_{t,\sigma}(y^{t(\sigma)}) \wedge t(f)(x^{t(\sigma_0)}, ..., x^{t(\sigma_n)}, y^{t(\sigma)})))$;

- for each constant symbol c of sort σ, $t(c)$ is a Σ_2-formula with at most one free variable x^{σ}, and that T_2 satisfies the admissibility condition: $T_2 \vdash \exists ! x^{t(\sigma)} (\delta_{t,\sigma}(x^{t(\sigma)}) \wedge t(c)(x^{t(\sigma)}))$.

[3] Note that the above definition of relative interpretation as given in Meadows (2024) assumes formulas are in functional normal forms, i.e., we only have formulas of the form $f(\overline{x}) = y$, and not iteration of functional terms such as $f(\overline{g(x)}) = y$. This is harmless as each formula can be transformed into functional normal forms modulo logical equivalence. There is also a direct way by which we can define a translation, as given by Halvorson (2019, Definition 4.5.4). The essential idea is that we induce a translation of Σ_1-terms $t(\overline{x})$ to $Ft(\overline{x}, y)$, which intuitively says that y is the object denoted by $t(\overline{x})$.

Now, t naturally induces a function from Σ_1-formulas to Σ_2-formulas defined as below, which we also denote by t:

- for $\phi := R(x^{\sigma_0}, ..., x^{\sigma_n})$, $t(\phi) := \bigwedge_i \delta_{\sigma,t}(x^{t(\sigma_i)}) \to t(R)(x^{t(\sigma_0)}, ..., x^{t(\sigma_n)})$;
- for $\phi := f(x^{\sigma_0}, ..., x^{\sigma_n}) = y^{\sigma}$, $t(\phi) := (\bigwedge_i \delta_{\sigma,t}(x^{t(\sigma_i)}) \wedge \delta_{\sigma,t}(y^{\sigma})) \to t(f)(x^{t(\sigma_0)}, ..., x^{t(\sigma_n)}, y^{t(\sigma)})$;
- for $\phi := x^{\sigma} = c$, $t(\phi) := \delta_{\sigma,t}(x^{t(\sigma)}) \to t(c)(x^{t(\sigma)})$;
- for $\phi := \neg\psi(x^{\sigma_0}, ..., x^{\sigma_n})$, $t(\phi) := \bigwedge_i \delta_{\sigma,t}(x^{t(\sigma_i)}) \to \neg t(\psi)$;
- for $\phi := \phi_1 \wedge \phi_2$, $t(\phi) := t(\phi_1) \wedge t(\phi_2)$;
- for $\phi := \forall x^{\sigma}\psi$, $t(\phi) := \forall x^{t(\sigma)}(t(\psi)(x^{t(\sigma)}))$.

We say t supports a relative translation of T_1 in T_2 if for any Σ_1-formula ϕ, if $T_1 \vdash \phi$, then $T_2 \vdash t(\phi)$.

Let $Mod(T)$ be the category whose objects are models of T and whose morphisms are elementary embeddings between models. Note that a relative translation t induces a functor $t^{\dagger}$ from $Mod(T_2)$ to $Mod(T_1)$, where a symbol s in Σ_1 is interpreted in $t^{\dagger}(\mathcal{M})$ as $t(s)$ is interpreted in $\mathcal{M}$.

The prefix "relative" is meant to suggest that we allow translation relativized to a subdomain. In the case of relative translation, we have a corresponding notion of relative *interpretation*.

Definition 2.18. Let T_1 and T_2 be two first-order theories in signature Σ_1 and Σ_2. We say T_1 *relatively interprets* T_2 if there is a Morita^s extension T_1^+ of T_1 such that $T_2 \subseteq T_1^+|_{\Sigma_2}$. And we say that the tuple $\langle T_1, T_1^+, T_2, \rangle$ is a *relative interpretation* of T_1 in T_2.

It is not hard to see that relative translation corresponds to relative interpretation.

Theorem 2.19. *Let T_1 be a Σ_1-theory and T_2 be a Σ_2-theory. There is a relative translation of T_1 in T_2 if and only if there is a relative interpretation of T_1 in T_2.*

Proof. Suppose that there is a relative translation t of T_1 in T_2. Then we can define a Morita^s extension T_2^+ of T_2 as follows: for each symbol $\alpha \in \Sigma_1$, we define α as the subsort of $t(\alpha)$ with the domain formula $\delta_{t,\alpha}$; for each relational symbol R in Σ_1, we use $t(R)$ as ϕ_R in the definition R, and similarly for functional symbols and constants. Then we prove by induction that for any Σ_1-formula ϕ, $T_2^+ \models \forall\overline{x}^{\overline{\sigma}}\phi$ iff $T_2^+ \models \forall\overline{x}^{t(\overline{\sigma})}t(\phi)$. Thus, for any ϕ, if $T_1 \models \phi$, then $T_2 \models t(\phi)$, and hence $T_2^+ \models \phi$. Therefore, $T_2 \subseteq T_2^+|_{\Sigma_1}$.

Conversely, suppose there is a relative interpretation $\langle T_2, T_2^+, T_1, \rangle$ of T_1 in T_2. Then we can define a relative translation t of T_1 in T_2 as follows: for each symbol $\alpha \in \Sigma_1$, suppose it is defined in T_2^+ as the subsort of $\sigma' \in \Sigma_2$ with the domain formula ϕ_{α}, then we define $t(\alpha)$ as σ', and $\delta_{t,\alpha}$ as ϕ_{α}; similarly, if $R \in \Sigma_1$ is defined by ϕ_R, then we define $t(R)$ as ϕ_R, and similarly for functional symbols and constants. Again we prove by induction that for any Σ_1-formula ϕ, $T_2^+ \models \forall\overline{x}^{\overline{\sigma}}\phi$ iff $T_2^+ \models \forall\overline{x}^{t(\overline{\sigma})}t(\phi)$. Now for any Σ_1-sentence ϕ, if $T_1 \models \phi$, then $T_2^+ \models \phi$, and so $T_2 \models t(\phi)$. Therefore, t is a relative translation of T_1 in T_2. □

While interpretation and translation often coincide, there is an advantage in adopting the perspective of interpretation: we have a general schema from which we can generate different notions of interpretation in a coherent manner and gives us a unified picture as to what resources we are allowed to use in a specific case of interpretation.

Schema 2.20. We say T_1 **interprets** T_2 if there is a **definitional extension** T_1^+ of T_1 such that $T_2 \subseteq T_1^+|_{\Sigma_{T_2}}$.

Thus, for instance, if we adopt Morita^s extension as the accepted notion of definitional extension in the above schema, we obtain relative interpretation. To take another example, we could adopt the standard definitional extension as the accepted notion of definitional extension, and we name the resulting notion of interpretation as *strict* interpretation.

Definition 2.21. Let T_1 and T_2 be two first-order theories in signatures Σ_1 and Σ_2. We say that T_1 *strictly interprets* T_2 if there is a standard definitional extension T_1^+ of T_1 such that $T_1^+|_{\Sigma_2} \equiv T_2$.

Strict interpretation also corresponds to a notion of translation. As one may expect, it is a notion translation that does not allow relativization or renaming of sorts. In fact, it is the notion of translation adopted by Barrett and Halvorson (2016), and we refer to it here as *strict translation.*

Definition 2.22. Let T_1 and T_2 be two theories in signatures Σ_1 and Σ_2. We say that t is a *strict translation* from $\mathcal{L}_{\Sigma_1}$ to $\mathcal{L}_{\Sigma_2}$ in T_2 if

- t is identity on sort symbols;
- for each relational symbol R of arity $\langle \sigma_0, ..., \sigma_n \rangle$, $t(R)$ is a Σ_2-formula with at most free variables $x^{\sigma_0}, ..., x^{\sigma_n}$;
- for each function symbol f of arity $\langle \sigma_0, ..., \sigma_n \rangle \to \sigma$, $t(f)$ is a Σ_2-formula with at most free variables $x^{\sigma_0}, ..., x^{\sigma_n}, y^{\sigma}$, and that T_2 satisfies the admissibility condition: $T_2 \vdash \forall x^{\sigma_0} ... \forall x^{\sigma_n} \exists ! y^{\sigma}(t(f)(x^{\sigma_0}, ..., x^{\sigma_n}, y^{\sigma}))$;
- for each constant symbol c of sort σ, $t(c)$ is a Σ_2-formula with at most one free variable x^{σ}, and that T_2 satisfies the admissibility condition: $T_2 \vdash \exists ! x^{\sigma}(t(c)(x^{\sigma}))$.

Now, t naturally induces a function from Σ_1-formulas to Σ_2-formulas defined as below, which we also denote by t:

- for $\phi := R(x^{\sigma_0}, ..., x^{\sigma_n})$, $t(\phi) := t(R)(x^{t(\sigma_0)}, ..., x^{t(\sigma_n)})$;
- for $\phi := f(x^{\sigma_0}, ..., x^{\sigma_n}) = y^{\sigma}$, $t(\phi) := t(f)(x^{t(\sigma_0)}, ..., x^{t(\sigma_n)}, y^{t(\sigma)})$;
- for $\phi := x^{\sigma} = c$, $t(\phi) := t(c)(x^{t(\sigma)})$;
- for $\phi := \neg\psi$, $t(\phi) := \neg t(\psi)$;
- for $\phi := \phi_1 \wedge \phi_2$, $t(\phi) := t(\phi_1) \wedge t(\phi_2)$;
- for $\phi := \forall x^{\sigma}\psi$, $t(\phi) := \forall x^{t(\sigma)}(t(\psi)(x^{t(\sigma)}))$.

Theorem 2.23. *Let T_1 be a Σ_1-theory and T_2 be a Σ_2-theory. There is a strict translation of T_1 in T_2 if and only if there is a strict interpretation of T_1 in T_2.*

Proof. The proof is exactly analogous to the proof of the previous theorem. □

We can also define the notion of Morita translation, which corresponds to Morita interpretation.[4] For the sake of convenience, we consider here only relational languages.[5]

Definition 2.24. Let Σ_1 and Σ_2 be the many-sorted languages for T_1 and T_2. A *Morita translation* of $\mathcal{L}_{\Sigma_1}$ in $\mathcal{L}_{\Sigma_2}$ is a function $f = f_0 \cup f_1 \cup f_2 \cup f_3$ defined as follows:

- f_0 (the map for sorts): for each sort σ in Σ_1, $f_0(\sigma) = \{\langle 0, \overline{\sigma_0}\rangle, ..., \langle n, \overline{\sigma_n}\rangle\}$ where $\overline{\sigma_0}, ..., \overline{\sigma_n}$ are sequences of sorts in Σ_2;

- f_1 (the map for variables): for each variable x^σ in Σ_1, $f_1(x^\sigma)$ is a set of sequences of Σ_2-variables $\{\overline{x}^{\overline{\sigma}}_{\langle i,\overline{\sigma}\rangle} | \langle i, \overline{\sigma}\rangle \in f_0(\sigma)\}$;
  - We require that $\overline{y}^{\overline{\beta}}_{\langle i,\overline{\beta}\rangle}, \overline{z}^{\overline{\gamma}}_{\langle j,\overline{\gamma}\rangle} \in f_1(x^\sigma)$ are disjoint if $i \neq j$.
  - We require that variables appeared in $f_1(x^\sigma)$ and $f_1(y^\gamma)$ are disjoint if x^σ and y^γ are distinct variables;
  - For convenience, we call $k \in \Pi_{x^\sigma \in \Sigma_1} f_1(x^\sigma)$ an assignment of f. We write $A(f) := \Pi_{x^\sigma \in \Sigma_1} f_1(x^\sigma)$. And we write $\tilde{k}(x^\sigma) := \langle i, \overline{\sigma}\rangle$ for $k(x^\sigma) = \overline{x}^{\overline{\sigma}}_{\langle i,\overline{\sigma}\rangle}$.

- f_2 (domain formulas): for each variable x^σ in Σ_1, $f_2(x^\sigma)$ is a function from $A(f)$ to Σ_2-formulas, such that for each assignment k, $f_2(x^\sigma)(k)$ is a Σ_2-formula with free variables $k(x^\sigma)$.
  - We require that if $\tilde{k}(x^\sigma) = \tilde{k}'(y^\sigma)$, then $f_2(x^\sigma)(k) = f_2(y^\sigma)(k')[k'(y^\sigma) \mapsto k(x^\sigma)]$;

- f_3 (the translation of atomic formulas): for each atomic Σ_1-formula $R(\overline{x}^{\overline{\sigma}})$ (including equality), $f_3(R(\overline{x}^{\overline{\sigma}}))$ is a function from $A(f)$ to Σ_2-formulas, such that for each assignment k, $f_3(R(\overline{x}^{\overline{\sigma}}))(k)$ is a Σ_2-formula with free variables $k(x^{\sigma_1}), ..., k(x^{\sigma_n})$.
  - We require that if $\tilde{k}(x^{\sigma_i}) = \tilde{k}'(y^{\sigma_i})$ for all σ_i, then $f_3(R(\overline{x}^{\overline{\sigma}}))(k) = f_3(R(\overline{y}^{\overline{\sigma}}))(k')[k'(\overline{y}^{\overline{\sigma}}) \mapsto k(\overline{x}^{\overline{\sigma}})]$

Note that f together with an assignment k extends to a map from Σ_1-formulas to Σ_2-formulas as follows:

- for $\phi := R(x^{\sigma_1}, ..., x^{\sigma_n})$, $f(\phi)(k) := \bigwedge_{1 \leq i \leq n} f_2(x^{\sigma_i})(k) \to f_3(R)(k)$;

- for $\phi := \neg\psi(x^{\sigma_1}, ..., x^{\sigma_n})$, $f(\phi)(k) := \bigwedge_{1 \leq i \leq n} f_2(x^{\sigma_i})(k) \to \neg f(\psi)(k)$;

- for $\phi := \psi_1 \wedge \psi_2$, $f(\phi)(k) := f(\psi_1)(k) \wedge f(\psi_2)(k)$;

- for $\phi := \forall x^\sigma \psi$, $f(\phi)(k) := \bigwedge_{k' = k[k(x^\sigma) \mapsto k'(x^\sigma)]} \forall k'(x^\sigma)(f(\psi)(k'))$;

We say that f is a *Morita translation* of T_1 in T_2 if for any Σ_1-formula ϕ, if $T_1 \vdash \phi$ then $T_2 \vdash f(\phi)(k)$ for all assignment k.

Lemma 2.25. *If f is a Morita translation of T_1 in T_2, then for any Σ_1-formula ϕ, if ϕ has free variables $x^{\sigma_1}, ..., x^{\sigma_n}$, then $f(\phi)(k)$ has free variables $k(x^{\sigma_1}), ..., k(x^{\sigma_n})$.*

[4] Definition 2.24 essentially corresponds to what Visser (2021) calls "relative, multi-dimensional, piecewise and non-identity-absolute interpretation".

[5] Particularly, it is much more elegant to treat the functionality of f as simply a theorem of T_1, which T_2 must preserve, than to treat it as an ad hoc admissibility condition that the translation must satisfy.

Proof. Easy induction. □

Lemma 2.26. *If f is a Morita translation of T_1 in T_2, and s is a Morita translation of T_2 in T_3, then there is a Morita translation of T_1 in T_3.*

Proof. We define a new translation g of T_1 in T_3 as follows.

Let $\overline{\sigma} = \langle \sigma_1, ..., \sigma_n \rangle$ be a sequence of sorts in Σ_1. Suppose $l \in \Pi_{\sigma_i \in \Sigma_2} s(\sigma_i)$. By $l(\overline{\sigma})$ we mean the sequence $\langle \langle \pi_0(l(\sigma_1)), ..., \pi_0(l(\sigma_n)) \rangle, \langle \pi_1(l(\sigma_1)), ..., \pi_1(l(\sigma_n)) \rangle \rangle$.

Let σ be a sort in Σ_1. For each $k \in \Pi_{\sigma_i \in \Sigma_1} f(\sigma_i)$, and $l \in \Pi_{\sigma_i \in \Sigma_2} s(\sigma_i)$, we define the sequence $\sigma_{k,l} = \langle \pi_0(k(\sigma)), \pi_0(l(\pi_1(k(\sigma))), \pi_1(l(\pi_1(k(\sigma))))) \rangle$. We enumerate all $\sigma_{k,l}$ by function $\alpha : \{\sigma_{k,l}\} \mapsto \mathbb{N}$ according to the lexicographic order of $\langle \pi_0(\sigma_{k,l}), \pi_1(\sigma_{k,l}) \rangle$. Thus, we define $g_0(\sigma) = \{\langle \alpha(\sigma_{k,l}), \pi_2(\sigma_{k,l}) \rangle | k \in \Pi_{\sigma_i \in \Sigma_1} f(\sigma_i), l \in \Pi_{\sigma_i \in \Sigma_2} s(\sigma_i)\}$.

Let $g_1(x^\sigma) = \{l(k(x^\sigma)) | k \in A(f), l \in A(s)\}$.

Let h be a g-assignment, let $g_2(x^\sigma)(h) := s(f_2(x^\sigma)(k))(l)$ for any f-assignment k and s-assignment l such that $l(k(x^\sigma)) = h(x^\sigma)$. It is easy to check that g_2 is well-defined: for any k, k', l, l' if $l(k(x^\sigma)) = l'(k'(x^\sigma))$, then $s(f_2(x^\sigma)(k))(l) = s(f_2(x^\sigma)(k'))(l')$. And by by Lemma 2.25, $s(f_2(x^\sigma)(k))(l)$ is a Σ_3-formula with free variables $l(k(x^\sigma))$, which equals $h(x^\sigma)$.

Let $g_3(R(x^{\sigma_1}, ..., x^{\sigma_n}))(h) := s(f_3(R(x^{\sigma_1}, ..., x^{\sigma_n}))(k))(l)$ for any f-assignment k and s-assignment l such that $l(k(x^{\sigma_i})) = h(x^{\sigma_i})$, for $1 \leq i \leq n$. Again, it is easy to check that g_3 is well-defined. Similarly, according to Lemma 2.25, $g_3(R(x^{\sigma_1}, ..., x^{\sigma_n}))(h)$ is a Σ_3-formula with free variables $h(x^{\sigma_1}), ..., h(x^{\sigma_n})$.

It is straightforward to check by induction that g is a translation of T_1 in T_3.

□

Theorem 2.27. *Let T_1 be a Σ_1-theory and T_2 be a Σ_2-theory. There is a Morita translation of T_1 in T_2 if and only if there is a Morita interpretation of T_1 in T_2.*

Proof. Suppose that there is a Morita translation f of T_1 in T_2. Let $\sigma \in \Sigma_1$ be arbitrary and $f(\sigma) = \{\langle 0, \overline{\sigma_0} \rangle, ..., \langle n, \overline{\sigma_n} \rangle\}$. Then we can define σ in T_2 as follows:[6]

- for each $\overline{\sigma_i} = \langle \sigma_i^1, ..., \sigma_i^m \rangle$, we define $\sigma_i^\dagger$ as the product of $\sigma_i^1, ..., \sigma_i^m$;
- for each $\sigma_i^\dagger$, we define σ_i^* as the quotient sort of $\sigma_i^\dagger$ by the equivalence relation that corresponds to $\bigwedge_{1 \leq j \leq m} f(=^{\sigma_i^j})(k)$, where $k(\sigma) = \langle i, \overline{\sigma_i} \rangle$;
- for each σ_i^*, we define σ_i^+ as the subsort of σ_i^* with the domain formula that corresponds to $\bigwedge_{1 \leq j \leq m} f_2(x^{\sigma_i^j})(k)$, where $k(\sigma) = \sigma_1$;
- finally, we define σ as the coproduct of $\sigma_1^+, ..., \sigma_n^+$.

Once all sorts in Σ_1 are defined, relational symbols in Σ_1 can then be defined as disjunctions of $f_3(R)(k)$ for all different assignments k. Thus, we have a Morita extension T_2^+ of T_2. It is routine to prove by induction that for any Σ_1-formula ϕ, $T_2^+ \vdash \phi$ iff $T_2^+ \vdash f(\phi)(k)$ for any assignment k. Thus, $T_1 \subseteq T_2^+|_{\Sigma_1}$, and therefore T_1 is Morita interpreted in T_2.

Now assume that T_1 is Morita interpreted in T_2, and let $\langle T_2, T_2^+, T_1, \rangle$ the Morita interpretation. By Lemma B.3, there is a sequence of pure Morita successors $S_0, ..., S_n$ such that $S_0 = T_2$ and T_2^+ is a standard definitional extension of S_n. We prove by induction on the length of n, that there is a Morita translation f of S_n in S_0. Suppose a new sort symbol σ of S_n is defined in S_{n-1} as:

[6] This requires that we can do arbitrarily finite products and coproducts in a Morita extension.

- a product of σ_0 and σ_1; then we define the translation f as follows:
  - $f_0(\sigma) = \{\langle 0, \langle \sigma_0, \sigma_1 \rangle \rangle\}$,
  - f_1 can be chosen arbitrarily,
  - $f_2(x^\sigma)(k) := \top$,
  - $f_3(x^\sigma = y^\sigma)(k) := k(x^\sigma) = k(y^\sigma)$,
  - $f_3(\pi_0(x^\sigma, y^{\sigma_0}))(k) := [k(x^\sigma)]^1 = y^{\sigma_0}$, where $[k(x^\sigma)]^1$ is the first element of $k(x^\sigma)$,
  - $f_3(\pi_1(x^\sigma, y^{\sigma_1}))(k) := [k(x^\sigma)]^2 = y^{\sigma_1}$, where $[k(x^\sigma)]^2$ is the second element of $k(x^\sigma)$;
- a coproduct of σ_0 and σ_1; then we define the translation f as follows:
  - $f_0(\sigma) = \{\langle 0, \sigma_0 \rangle, \langle 1, \sigma_1 \rangle\}$,
  - f_1 can be chosen arbitrarily,
  - $f_2(x^\sigma)(k) := \top$,
  - $f_3(x^\sigma = y^\sigma)(k)$ is defined as $k(x^\sigma) = k(y^\sigma)$ if $\tilde{k}(x^\sigma) = \tilde{k}(y^\sigma)$, $\bot$ otherwise,
  - $f_3(p_0(x^{\sigma_0}, y^\sigma))(k)$ is defined as $k(y^\sigma) = x^{\sigma_0}$ if $\tilde{k}(y^\sigma) = \langle 0, \sigma_0 \rangle$, $\bot$ otherwise,
  - $f_3(p_1(x^{\sigma_1}, y^\sigma))(k)$ is defined as $k(y^\sigma) = x^{\sigma_1}$ if $\tilde{k}(y^\sigma) = \langle 1, \sigma_1 \rangle$, $\bot$ otherwise;
- a subsort of σ_1 by domain formula ϕ; then we define the translation f as follows:
  - $f_0(\sigma) = \{\langle 0, \sigma_0 \rangle\}$,
  - f_1 can be chosen arbitrarily,
  - $f_2(x^\sigma)(k) := \phi(k(x^\sigma))$,
  - $f_3(x^\sigma = y^\sigma)(k) := k(x^\sigma) = k(y^\sigma)$;
- a quotient of σ_1 by equivalence relation ϕ; then we define the translation f as follows:
  - $f_0(\sigma) = \{\langle 0, \sigma_0 \rangle\}$,
  - f_1 can be chosen arbitrarily,
  - $f_2(x^\sigma)(k) := \top$,
  - $f_3(x^\sigma = y^\sigma)(k) := \phi(k(x^\sigma) = k(y^\sigma))$.

Thus, we have a Morita translation f of S_{n+1} in S_n. By inductive hypothesis, we have a Morita translation s of S_n in S_0, by Lemma 2.26, we have a translation $s \circ f$ of S_{n+1} in S_0.

Since T_2^+ is a definitional extension of S_n, there is a canonical translation l of T_2^+ in S_n. Then $s \circ (f \circ l)$ is a Morita translation of T_2^+ in $S_0 = T_2$. Then its restriction to Σ_1 will be a Morita translation of T_1 in T_2. □

With different notions of definitional extension, we can reproduce different notions of interpretation in the literature, which further correspond to different notions of translation, just as we did above.

Definitional Extension	Interpretation		Translation	
	Name	Other Names	Name	Other Names
Standard Definitional Extension	Strict Interpretation	N/A	Strict Translation	"Translation" in Barrett and Halvorson (2016) and Halvorson (2019)
Morita^s Extension	Relative Interpretation	N/A	Relative Translation	"Relative interpretation" in Visser (2021)
Morita^p Extension	Multi-dimensional Interpretation	N/A	Multi-dimensional Translation	"Multi-dimensional Interpretation" in Visser (2021)
Morita^c Extension	Piecewise Interpretation	N/A	Piecewise Translation	"Piecewise Interpretation" in Visser (2021)
Morita^q Extension	Quotientive Interpretation	N/A	Quotientive Translation	"Non-identity-absolute Interpretation" in Visser (2021)
Morita^{s,q} Extension	Relative, Quotientive Interpretation	"Interpretation" in Button and Walsh (2018)	Relative, Quotientive Translation	N/A
Morita^{p,s,q} Extension	Multi-dimensional, Relative and Quotientive Interpretation	N/A	Multi-dimensional, Relative and Quotientive Translation	"Generalized translation" in Halvorson (2019)
Morita^{p,c,s,q} Extension	Morita Interpretation	"Morita Interpretation" in Meadows (2024)	Morita Translation	N/A

Now we have a series of notions of interpretation. Given a specific notion of interpretation, we can give different coordinated-interpretation criteria based on weaker or stronger requirements of coordination (see, e.g., Visser (2004, 2009), Meadows (2024)).

Schema 2.28. We say that two theories T_1 and T_2 are

- **interpretably equivalent** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1 in T_2 such that for any model $\mathcal{M}_1$ of T_1, $s^\dagger(t^\dagger(\mathcal{M}_1)) = \mathcal{M}_1$, and for any model $\mathcal{M}_2$ of T_2, $t^\dagger(s^\dagger(\mathcal{M}_2)) = \mathcal{M}_2$.

- **bi-interpretable** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1

in T_2 such that for any model $\mathcal{M}_1$ of T_1, $s^\dagger(t^\dagger(\mathcal{M}_1))$ is provably isomorphic to $\mathcal{M}_1$ and for any model $\mathcal{M}_2$ of T_2, $t^\dagger(s^\dagger(\mathcal{M}_2))$ is **provably isomorphic** to $\mathcal{M}_2$.

- **iso-congruent** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1 in T_2 such that for any model $\mathcal{M}_1$ of T_1, $s^\dagger(t^\dagger(\mathcal{M}_1))$ is isomorphic to $\mathcal{M}_1$, and for any model $\mathcal{M}_2$ of T_2, $t^\dagger(s^\dagger(\mathcal{M}_2))$ is isomorphic to $\mathcal{M}_2$.

- **elementary-congruent** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1 in T_2 such that for any model $\mathcal{M}_1$ of T_1, $s^\dagger(t^\dagger(\mathcal{M}_1))$ is elementarily equivalent to $\mathcal{M}_1$, and for any model $\mathcal{M}_2$ of T_2, $t^\dagger(s^\dagger(\mathcal{M}_2))$ is elementarily equivalent to $\mathcal{M}_2$.

- **mutually interpretable** if each **interprets** the other.

The advantage of using interpretation in our criteria of theoretical equivalence is that different notions of interpretations are clearly classified according to which structures are considered as free lunches, i.e., what notion of definitional extension we adopt. The drawback is that sometimes finding the corresponding translation at the sentential level could be difficult. In particular, in standard many-sorted logic where variables are sorted, we do not have a formula that intuitively speaks about the coproduct of two sorts. This means that while we can extend our theories or models by coproduct, we may not be able to give a straightforward translation of it. Visser (2009) develops the method of "piecewise" translation to deal with coproduct. Alternatively, if we have a sort with at least two distinguished elements, then one may use this sort as an auxiliary tool, translate coproduct via product and quotient: use $(a, b, 1)$ modulo b to represent $a \in \sigma_1$, and use $(a, b, 2)$ modulo a to represent b in σ_2 (Halvorson, 2019). Visser (2021) sketches a similar way of doing it.

2.4 The Categorical Criteria

The Categorical Criteria, as the name suggests, look at the categories assotiated with different theories. The most common category that people focus on is the semantical category of a theory, whose objects are models of the theory and arrows elementary embeddings between models.

Meadows (2024) discusses three notions of equivalence in category theory that could be applied to discuss the equivalence between semantical categories of theories: categorical isomorphism, categorical equivalence, and objective categorical equivalence. The first two are familiar; the third one is new.

Definition 2.29. We say that two categories $\mathbb{C}$ and $\mathbb{D}$ are *objectively equivalent* if there is a functor F from $\mathbb{C}$ to $\mathbb{D}$ and a functor G from $\mathbb{D}$ to $\mathbb{C}$ such that for any object $A \in \mathbb{C}$, $F \circ G(A) \cong A$, and for any object $B \in \mathbb{D}$, $G \circ F(B) \cong B$.

So we have the following definition.

Definition 2.30. We say that two theories T_1 and T_2 are

- *categorically isomorphic* if their semantical categories are isomorphic.

- *categorically equivalent* if their semantical categories are equivalent.

- *objectively categorically equivalent* if their semantical categories are objectively equivalent.

We have the following results.

Theorem 2.31. *Two theories are categorically equivalent iff they are categorically isomorphic.*

This was proved in Meadows (2024, Corollary 2.9). For future reference, we present the proof below. The proof assumes that there is a least inaccessible cardinal Ω and objects in $Mod(T)$ are models of T within the universe V_Ω.

Lemma 2.32. *Let $Mod(T_1)$ and $Mod(T_2)$ be the categories of models of T_1 and T_2 respectively. Let t be a functor from $Mod(T_1)$ to $Mod(T_2)$ and s be a functor from $Mod(T_2)$ to $Mod(T_1)$ such that t and s witness the categorical equivalence of T_1 and T_2. Then there is an isomorphism f between $Mod(T_1)$ and $Mod(T_2)$ that respects t and s modulo isomorphism of the models (i.e. $f(\mathcal{M}) \cong t(\mathcal{M})$ for each model $\mathcal{M} \in Mod(T_1)$ and $f^{-1}(\mathcal{N}) \cong s(\mathcal{N})$ for each model $\mathcal{N} \in Mod(T_2)$).*

Theorem 2.31 then follows as a corollary.

Proof. Note that t establishes a bijection f between isomorphism classes of models of T_1 and those of models of T_2. Let $(\cdot)^*$ be a function that chooses a representative for each isomorphism class of $Mod(T_1)$, and let $\pi_{\cdot}$ be a function which chooses an isomorphism from a model to its representative and that $\pi_{[\mathcal{M}]^*} = id_{[\mathcal{M}]^*}$ for each model $\mathcal{M}$ of T_1. $(\cdot)^\dagger$ and $\rho_{\cdot}$ are defined analogously for $Mod(T_2)$. For each isomorphism class $[\mathcal{M}]$ of T_1, note that $|[\mathcal{M}]| = |f([\mathcal{M}])| = \Omega$, and so we can choose a bijection $H_{[\mathcal{M}]}$ between $[\mathcal{M}]$ and $f([\mathcal{M}])$, which maps $[\mathcal{M}]^*$ to $f([\mathcal{M}])^\dagger$. Let H be a functor defined as follows:

- for each object $\mathcal{M}$ in $Mod(T_1)$, $H(\mathcal{M}) = H_{[\mathcal{M}]}(\mathcal{M})$;
- for each arrow h from $\mathcal{M}$ to $\mathcal{N}$ in $Mod(T_1)$, $H(h) = (\rho_{H(\mathcal{N})})^{-1} \circ t(\pi_{\mathcal{N}} \circ h \circ (\pi_{\mathcal{M}})^{-1}) \circ \rho_{H(\mathcal{M})}$.

It is easy to check that H is an isomorphism between $Mod(T_1)$ and $Mod(T_2)$. And by construction, H respects t and s modulo isomorphism of models. □

Theorem 2.33. *If two theories are Morita equivalent then they are categorically equivalent.*

This was proved in Barrett and Halvorson (2016, Theorem 5.6).

Instead of looking at the semantic category of a theory, we can also look at the syntactical category of a theory. A syntactical category is a category whose objects are formulas in contexts and arrows are provable functions between them.

Definition 2.34. Given a first-order many-sorted theory T with signature Σ, its *syntactical category* $\mathbb{C}_T$ is a category defined as follows:

- The objects are equivalence classes of Σ-*formulas in contexts* which we denote as $\overline{x}.\phi$, where $\overline{x}$ a list of variables and ϕ is a Σ-formula with at most free variables in x. Let p be the function that sends variables to sorts. Two Σ-formulas in contexts $\overline{x}.\phi$ and $\overline{y}.\psi$ are equivalent if $p(\overline{x}) = p(\overline{y})$ and $\phi = \psi[\overline{x}/\overline{y}]$.
- The arrows are also are equivalent classes of Σ-*formulas in contexts* $\{\overline{z}.\theta\} : \{\overline{x}.\phi\} \to \{\overline{y}.\psi\}$ such that
  - $\theta \vdash_{\overline{x},\overline{y}} \phi \wedge \psi$;
  - $\theta \wedge \theta[\overline{y}/\overline{z}] \vdash \overline{y} = \overline{z}$;
  - $\phi \vdash_{\overline{x}} \exists y \theta$.

There is also a corresponding notion of Morita equivalence defined in Johnstone (2003), which we label as J-Morita equivalence, to distinguish it from the notion of Morita equivalence we defined before.

Definition 2.35. Two cartesian theories T_1 and T_2 are *J-Morita equivalent*, if their syntactical categories are equivalent.

Definition 2.36. Two regular theories T_1 and T_2 are *J-Morita equivalent*, if the effectivization of their syntactical categories are equivalent.

Definition 2.37. Two coherent theories T_1 and T_2 are *J-Morita equivalent*, if the effectivization of the positivization of their syntactical categories are equivalent.

Definition 2.38. Two geometrical theories T_1 and T_2 are *J-Morita equivalent*, if the effectivization of the infinite positivization of their syntactical categories are equivalent.

Tsementzis (2017) proves the following results.

Theorem 2.39. *For coherent theories T_1 and T_2, they are J-Morita equivalent iff they are standard Morita equivalent.*

Theorem 2.40. *For geometrical theories T_1 and T_2, they are J-Morita equivalent iff they are Morita equivalent with subsort, quotient, product and infinite coproduct.*

Theorem 2.41. *For regular theories T_1 and T_2, they are J-Morita equivalent iff they are Morita equivalent with subsort, quotient, product.*

Theorem 2.42. *For cartesian theories T_1 and T_2, they are J-Morita equivalent iff they are Morita equivalent with subsort, and product.*

And it is not entirely clear (justified) why for different types of theories we have different notions of J-Morita equivalence. There seems to be some arbitrariness just as in the case of standard Morita equivalence.

2.5 Results about Comparative Strengths

Meadows (2024) paints a map of theoretical equivalence:

$$\begin{tikzcd}
\text{Relatively Interpretably Equiv} \arrow[r] \arrow[d] & \text{Morita Interpretably Equiv} \arrow[r] \arrow[d] & \text{Categorical Isomorphism} \arrow[d, leftrightarrow] \\
\text{Bi-interpretable} \arrow[r] \arrow[d] & \text{Morita Bi-interpretable} \arrow[r] \arrow[d] & \text{Categorical Equivalence} \arrow[d] \\
\text{Iso-congruence} \arrow[r] \arrow[d] & \text{Morita-congruence} \arrow[r] \arrow[d] & \text{Objective Equivalence} \arrow[d] \\
\text{Mutually Interpretable} \arrow[r] & \text{Morita Mutual} \arrow[r] & \text{Category Mutual}
\end{tikzcd}$$

To see how the notions in this map are defined, recall that we have a series of schemas of coordinated-interpretation criteria based on different requirements of coordination (Schema 2.28): **interpretably equivalent**, **bi-interpretable**, **iso-congruent**, **elementary-congruent**, **mutually interpretable**. If we plug in relative interpretation in the above schemas, we get the leftmost column. And if we plug in Morita interpretation, then we get the middle column. The rightmost column has already been defined in Definition 2.30. It can be obtained by dropping the requirements related to interpretation in Schema 2.28 and focusing solely on the requirements for coordinations — that is, functors no longer need to be generated from interpretations.

There are two deficiencies in the above map. First, the common-extension criteria are not included. Second, it is not known whether certain arrows are invertible. In particular, we have the following questions (Meadows, 2024):

$$\begin{tikzcd}
\text{Relatively Interpretably Equiv} & \text{Morita Interpretably Equiv} & \text{Categorical Iso} \\
\text{Bi-interpretable} \arrow[u, "?"] \arrow[ur, "?"] & \text{Morita Bi-interpretable} \arrow[u, "?"] & \text{Categorical Equiv} \\
\text{Iso-congruence} \arrow[u, "?"] \arrow[ur, "?"] \arrow[urr, "?"] & \text{Morita-congruence} \arrow[u, "?"] \arrow[ur, "?"] & \text{Objective Equiv}
\end{tikzcd}$$

The following two sections address each of these two deficiencies.

2.5.1 Situating the Common-Extension Criteria

The two most important common-extension criteria are definitional equivalence and Morita equivalence. Meadows (2024, Proposition 7.1) shows that definitional equivalence can be positioned in the map: it is the same as strict interpretably equivalence.

Proposition 2.43. *Let T_1 and T_2 be two first-order theories in signatures Σ_1 and Σ_2. Then T_1 and T_2 are definitional equivalent iff they are relatively interpretably equivalent.*

It is also claimed that by Corollary 5.14 in Meadows (2024), Morita interpretably equivalence is the same as Morita equivalence. But a valid proof is still lacking.[7]

In the following, we give the proof that Morita equivalence implies Morita interpretably equivalence. For the sake of rigor, we define Morita interpretation and Morita interpretably equivalence formally.

Definition 2.44. Let T_1 be a Σ_1-theory, and T_2 a Σ_2-theory. We say that T_1 *Morita interprets* T_2 if there is a Morita descendant T_1^+ of T_1, such that $T_1^+|_{\Sigma_2} \equiv T_2$. A Morita interpretation t is then a tuple (T_1, T_1^+, T_2).

Recall that two theories are interpretably equivalent (as a schema) if they are mutually interpretable and the functions t^* and s^* between their class of models induced by interpretations are

[7] This point is confirmed by the author in private correspondence, to which I am very thankful.

inverses of each other. In the case of Morita interpretation, interpretations do not decide uniquely a functor between the categories of models, thus we only require that there be functors *compatible with the interpretations* that are inverses of each other.

Definition 2.45. Let T_1 and T_2 be two theories with signatures Σ_1 and Σ_2. Suppose T_1 Morita interprets T_2 via the Morita interpretation $t = (T_1, T_1^+, T_2)$.

We say that a functor $t^\dagger$: $\mathrm{Mod}(T_1) \to \mathrm{Mod}(T_2)$ *is compatible with this Morita interpretation*, if for all models $\mathcal{M}$ of T_1, there is some model $\mathcal{M}'$ of T^+ such that:

- $\mathcal{M}'|_{\Sigma_1} = \mathcal{M}$ and,
- $\mathcal{M}'|_{\Sigma_2} = t^\dagger(\mathcal{M})$.

Definition 2.46. We say that two theories T_1 and T_2 are *Morita interpretably equivalent* if there is a Morita interpretation t of T_2 in T_1 and a Morita interpretation s of T_1 in T_2 such that there are functors $t^\dagger$ and $s^\dagger$ compatible with t and s respectively, such that $t^\dagger \circ s^\dagger = id$ and $s^\dagger \circ t^\dagger = id$.

Theorem 2.47. *If two theories are Morita equivalent then they are interpretably equivalent.*

Proof. Assume that T_1 and T_2 are Morita equivalent with signatures Σ_1 and Σ_2. Then there is a theory T_3 which is a common Morita extension of T_1 and T_2. Since $T_3|_{\Sigma_2} \equiv T_2$ and $T_3|_{\Sigma_1} \equiv T_1$, there is a Morita interpretation $t = (T_2, T_3, T_1)$ of T_1 in T_2 and a Morita interpretation $s = (T_1, T_3, T_2)$ of T_2 in T_1.

As in Barrett and Halvorson (2016, p. 572), we can define a Π_1 functor from $Mod(T_3)$ to $Mod(T_1)$ as follows:

- $\Pi_1(\mathcal{M}) = \mathcal{M}_{\Sigma_1}$
- $\Pi_1(h) = h_{\Sigma_1}$ for every arrow $h : \mathcal{M} \to \mathcal{N}$ in $Mod(T_3)$, where the family of maps h_{Σ_1} is defined to be $h_{\Sigma_1} = \{h_\sigma : M_\sigma \to N_\sigma$ such that $\sigma \in \Sigma_1\}$.

By Propositions 5.2, 5.3, and 5.5 in Barrett and Halvorson (2016), Π_1 witnesses the categorical equivalence between $Mod(T_1)$ and $Mod(T_3)$, and let Π^1 be its quasi-inverse.

Similarly, we can define a Π_2 functor from $Mod(T_3)$ to $Mod(T_2)$ that witnesses the categorical equivalence between $Mod(T_2)$ and $Mod(T_3)$, and let Π^2 be its quasi-inverse.

By transitivity of categorical equivalence, $\Pi_1 \circ \Pi^2$ and $\Pi_2 \circ \Pi^1$ witness the categorical equivalence between $Mod(T_1)$ and $Mod(T_2)$. By Lemma 2.28, there is an isomorphism f between $Mod(T_1)$ and $Mod(T_2)$ that respects the $\Pi_1 \circ \Pi^2$ and $\Pi_2 \circ \Pi^1$ modulo isomorphism of models.

We only need to show that f and f^{-1} are compatible with Morita interpretations s and t respectively. Let $\mathcal{M}_1$ be an arbitrary model of T_1. Let $\mathcal{M}_1^+$ be an arbitrary expansion of $\mathcal{M}_1$ in T_3. Since Π^1 is a categorical equivalence, $\Pi^1(\Pi_1(\mathcal{M}_1^+)) \cong \mathcal{M}_1$ and hence $\Pi^1(\mathcal{M}) \cong \mathcal{M}_1^+$. As functors preserve isomorphisms, $\Pi_2(\Pi^1(\mathcal{M})) \cong \Pi_2(\mathcal{M}_1^+) = (\mathcal{M}_1^+)|_{\Sigma_2}$. Substituting $\Pi_2(\Pi^1(\mathcal{M}))$ for $(\mathcal{M}_1^+)|_{\Sigma_2}$ in $\mathcal{M}_1^+$ for the interpretation of symbols in Σ_2, we have a new model $\mathcal{M}'$. Since $\Pi_2(\Pi^1(\mathcal{M})) \cong (\mathcal{M}_1^+)|_{\Sigma_2}$, $\mathcal{M}' \cong \mathcal{M}_1^+$ and hence is a model of T_3. And by construction, $\mathcal{M}'|_{\Sigma_1} = \mathcal{M}_1$ and $\mathcal{M}'|_{\Sigma_2} = \Pi^1(\mathcal{M})$. Thus, f is compatible with t. That f^{-1} is compatible with t is proved similarly. Thus, T_1 and T_2 are Morita interpretably equivalent. □

It is not yet known whether the converse also holds.

2.5.2 Completing the Map

Now we try to complete the map by answering the question marks.

We first show that there are iso-congruent theories which are not categorically equivalent, and thus answer all the question marks in the second row negatively.

Proposition 2.48. *There are iso-congruent theories which are not categorically equivalent.*

Proof. Let Σ_1 be a signature with a single sort, a predicate P, a binary relational symbol R, and infinitely many constants $c_0, c_1, ...$ Let T_1 be a Σ_1-theory which says:

- $c_i \neq c_j$ for $i \neq j$;
- $P(c_i)$ for each i;
- $\exists x \exists y (R(x, y) \wedge \neg P(x) \wedge \neg P(y) \wedge \forall z \forall w (R(z, w) \leftrightarrow ((z = x \wedge w = y) \vee (z = y \wedge w = x))))$.
- There are infinitely (arbitrarily finitely) many elements that are not P.

Let Σ_2 be a signature with a single sort, a predicate Q, and infinitely many constants $d_0, d_1, ...$ Let T_2 be a Σ_2-theory which says:

- $d_i \neq d_j$ for $i \neq j$;
- $Q(d_i)$ for each i;
- There are infinitely (arbitrarily finitely many) many elements that are not Q.

We first prove that they are not categorically equivalent. Consider the model $\mathcal{M}_1$ of T_1, which contains exactly two elements a, b which are P but are not named by any constant, and R holds between c, d. Note that the automorphism group of $\mathcal{M}_1$ contains two distinct normal subgroups of order 2, one is the subgroup generated by the permutation between a and b, and the other by the permutation between c and d. However, the models of T_2 contain at most one such normal subgroup of order 2 (when there are exactly two elements that are Q but are not named by any constant).[8] Since categorical equivalence preserves automorphism groups up to isomorphism, T_1 and T_2 are not categorically equivalent.

We then prove that they are iso-congruent. Let t be the translation from Σ_2-formulas to Σ_1-formulas which sends $Q(x)$ to $P(x)$ and d_i to c_i. It is easy to check that t supports an interpretation of T_2 in T_1. Let t^* be the induced function from models of T_1 to models of T_2.

Let s be the translation from Σ_1-formulas to Σ_2-formulas which sends $P(x)$ to $Q(x) \wedge x \neq c_0 \wedge x \neq c_1$, c_i to d_{i+2}, and $R(x, y)$ to $(x = d_0 \wedge y = d_1) \vee (x = d_1 \wedge y = d_0)$. We keep the domain constant in both t and s. It is easy to check that s supports an interpretation of T_1 in T_2. Let s^* be the induced function from models of T_2 to models of T_1.

Let $\mathcal{M}$ be an arbitrary model of T_1. Let a, b be the elements in $\mathcal{M}$ of which R holds. Let S be the infinite set of elements in $\mathcal{M}$ which are not P and does not contain a, b. We pick an arbitrary bijection g from S to $S \cup \{a, b\}$. Let k be the partial function that sends a to $(c_0)^{\mathcal{M}}$, b to $(c_1)^{\mathcal{M}}$, $(c_j)^{\mathcal{M}}$ to $(c_{j+2})^{\mathcal{M}}$ for $j \geq 2$, and remains identity on all points which are P but are not constants. It is easy to check that $f = g \cup k$ is an isomorphism from $\mathcal{M}$ to $s^*(t^*(\mathcal{M}))$.

[8] The part of the proof that the automorphism groups of $\mathcal{M}_1$ and $\mathcal{M}_2$ are not isomorphic was pointed out to me by my classmate and roommate Ruiting Jiang during our oral conversations. I shall express my sincere gratitude to him.

Let $\mathcal{N}$ be an arbitrary model of T_2. Let V be the infinite set of elements in $\mathcal{N}$ which are not Q. We pick an arbitrary bijection g from V to $V \cup \{(d_0)^{\mathcal{N}}, (d_1)^{\mathcal{N}}\}$. Let k be the partial function that sends $(d_i)^{\mathcal{N}}$ to $(d_{i+2})^{\mathcal{N}}$, and remains identity on all points which are Q but are not constants. It is easy to check that $f = g \cup k$ is an isomorphism from $\mathcal{N}$ to $t^*(s^*(\mathcal{N}))$.

□

Now we turn to arrows in the first rows, and answer the question whether bi-interpretability implies definitional equivalence. The original version of bi-interpretability is defined in Meadows (2024) based on relative interpretation (i.e., interpretation with subsort). We give an example to show that (relative) interpretability does not imply definitional equivalence.

The definition of bi-interpretability based on relative interpretation can be given as follows. Note that the formulation is based on Hodges (1993) and Visser (2021).

Definition 2.49. Let T_1 and T_2 be two theories with signatures Σ_1 and Σ_2 respectively. Let t and s be interpretations of T_1 in T_2. We say that t and s are *homotopic* if there is a Σ_2-formula $\chi(x, y)$ such that T_2 proves that $\chi(x, y)$ is a bijection and that for each $R \in \Sigma_1$

$$T_2 \vdash \forall x_0, \dots, x_n \forall y_0, \dots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow\ (t(R)(x_0, \dots, x_n) \ \leftrightarrow\ s(R)(y_0, \dots, y_n)) \Big)$$

Or equivalently, for every model $\mathcal{M}$ of T_2, $\chi^{\mathcal{M}}$ is a bijection, and that for any $\overline{a}$, $\overline{b}$ in $\mathcal{M}$ such that $\mathcal{M} \models \chi(\overline{a}, \overline{b})$, we have $\mathcal{M} \models s(R)(\overline{a})$ iff $\mathcal{M} \models t(R)(\overline{b})$.

And we say that $\chi(x, y)$ is a homotopy from s to t.

We use $\chi(\overline{x}, \overline{y})$ as an abbreviation for $\bigwedge_{i<n} \chi(x_i, y_i)$.

Definition 2.50. T_1 and T_2 are *(relatively) bi-interpretable* iff there is a relative interpretation t of T_1 in T_2 and a relative interpretation s of T_2 in T_1 such that $s \circ t$ is homotopic to the identity interpretation of T_1 in itself, and $t \circ s$ is homotopic to the identity interpretation of T_2 in itself.

Meadows (2024) instead uses the prefix "strictly" to emphasize that he uses the notion of relative interpretation and does not allow interpretation by product, coproduct, or quotient.

We show that relative bi-interpretability does not imply definitional equivalence.

Theorem 2.51. *There are two theories which are (relatively) bi-interpretable but not definitionally equivalent.*

Proof. Let Σ_1 be a signature with a single sort, two constants c_0 and c_1, and two binary relational symbols R and J. Let Σ_2 be a signature with a single sort, a constant d_0, and two binary relational symbols Q and K.

Let $\mathcal{M}_1$ be the following model of Σ_1:

$$\begin{tikzcd}
c_0 \arrow[r, "R"] \arrow[d, leftrightarrow, "J"'] & \cdot \arrow[r, "R"] \arrow[d, leftrightarrow, "J"'] & \cdot \arrow[r, "R"] \arrow[d, leftrightarrow, "J"'] & \cdot \arrow[r, "R"] \arrow[d, leftrightarrow, "J"'] & \cdots \\
c_1 \arrow[r, "R"] & \cdot \arrow[r, "R"] & \cdot \arrow[r, "R"] & \cdot \arrow[r, "R"] & \cdots
\end{tikzcd}$$

And Let $T_1 = Th(\mathcal{M}_1)$.

Let $\mathcal{M}_2$ be the following model of Σ_2:

$$\begin{tikzcd}
d_0 \arrow[r, "Q"] & \cdot \arrow[r, "Q"] \arrow[d, leftrightarrow, "K"'] & \cdot \arrow[r, "Q"] \arrow[d, leftrightarrow, "K"'] & \cdot \arrow[r, "Q"] \arrow[d, leftrightarrow, "K"'] & \cdots \\
 & \cdot \arrow[r, "Q"] & \cdot \arrow[r, "Q"] & \cdot \arrow[r, "Q"] & \cdots
\end{tikzcd}$$

And Let $T_2 = Th(\mathcal{M}_2)$.

We prove that T_1 and T_2 are (relatively) bi-interpretable. Let t be the relative translation from Σ_1 formulas to Σ_2 formulas with domain formula $x \neq c_0$ and which sends c_0 to $Q(d_0, x)$, c_1 to $\exists z(Q(d_0, z) \wedge K(z, x))$, $R(x, y)$ to $Q(x, y)$, and $J(x, y)$ to $K(x, y)$. It is easy to check that t supports an interpretation of T_1 in T_2. Let t^* be the induced function from models of T_1 to models of T_2.

Let s be the relative translation from Σ_2 formulas to Σ_1 formulas with domain formula $x \neq c_1$ and which sends d_0 to c_0, $Q(x, y)$ to $R(x, y)$, and $K(x, y)$ to $J(x, y)$. Again, it is easy to check that s supports an interpretation of T_2 in T_1. Let s^* be the induced function from models of T_2 to models of T_1.

Note that $R(x, y)$ defines an isomorphism from $\mathcal{M}_1$ to $t^*(s^*(\mathcal{M}_1))$, and $Q(x, y)$ defines an isomorphism from $\mathcal{M}_2$ to $s^*(t^*(\mathcal{M}_2))$. Therefore, T_1 and T_2 are (relatively) bi-interpretable.

We then prove that T_1 and T_2 are not definitionally equivalent. Suppose (towards a contradiction) that they are. Now there is a definitional extension T_3 of T_1 and T_4 of T_2 such that $T_3 \equiv T_4$. For convenience, we refer to the point related to x by J or K the J-partner or the K-partner of x. Let us call a point whose J-partner and K-partner are distinct a *half-hearted* point. Note that T_2 and consequently T_4 proves that $\neg\exists x(K(d_0, x))$, in particular, $T_4 \vdash \exists y((J(d_0, y) \wedge \neg K(x, y) \wedge \exists z(K(z, y))))$, i.e., that the J-partner of d_0 has d_0 as its J-partner, but not as its K-partner. Thus the J-partner of d_0, which we label as h_1, is a half-hearted point. But then T_2 and consequently T_4 also prove that the K-partner of h_1 has h_1 as its K-partner, but not as its J-partner, and hence the K-partner of h_1, which we label as h_2, is also a half-hearted point. The same reasoning applies to the J-partner of h_2 and so on. Thus, for arbitrary n, T_4 will prove that there are n half-hearted points $h_1, h_2, ...$ mutually linked by K and J alternatively.

By distance we mean the distance between two points in the Gaifman graph. Note that for any natural number n, T_4 proves that the number of points within the n-neighborhood of d_0 is below some finite number i. Then T_4 proves that for any natural number n, there will always be some half-hearted points linked by K and J as above, but are outside of the n-neighborhood of d_0. Reasoning within T_4, suppose x, y, z, w are four such points, and we have xKy, yJz and zKw (without loss of generality). Then we have yJz but not xJw. Let ϕ_j be the definition of J in T_2, then we have $\phi_J(y, z) \wedge \neg\phi_J(x, w)$. Thus, since definitional extension is conservative, for any n, T_2 proves that there are four points x, y, z, w outside the n-neighborhood of d_0 such that $\phi_J(y, z) \wedge \neg\phi_J(x, w)$. But notice that for any model $\mathcal{M}$ of T_2, and for any such four points x, y, z, w outside the n-neighborhood of d_0, $N_n^{\mathcal{M}}(y, z) \cong N_n^{\mathcal{M}}(x, w)$. So, the locality rank of the query defined by ϕ_J must be greater than n. By Theorem 4.12 and Theorem 4.13 in Libkin (2004), ϕ_J must have quantifier rank higher than $\log_3(\frac{2n+1}{3})$. Since this holds for arbitrary n, ϕ_J will have infinite quantifier rank, which is impossible.

Therefore, T_1 and T_2 are not definitionally equivalent.

□

But this would be true if we adopt a stronger notion of bi-interpretability, in particular, if we do not allow relativize quantification in interpretation, i.e., we do not allow non-trivial domain fomula. We can show that, in this case, (strict) bi-interpretability indeed implies definitional equivalence.

We prove the following lemma.

Lemma 2.52. *Let T_1 and T_2 be two theories with signatures Σ_1 and Σ_2 respectively. Let s, t be two (strict) interpretations of T_1 in T_2. Let χ be a homotopy from s to t. We prove that for each formula $\phi(\overline{x})$ we have:*

$$T_2 \vdash \forall \overline{x} \forall \overline{y} \Big(\chi(\overline{x}, \overline{y}) \to (s(\phi)(\overline{x}) \leftrightarrow t(\phi)(\overline{y})) \Big)$$

Proof. By routine induction on structures of ϕ. The base case is guaranteed by definition. The inductive cases for conjunctions and negations are trivial. For the inductive case of existential quantifiers, we notice that T_2 proves that χ is a bijection of the entire domain (as we do not allow relativization here), and therefore T_2 proves that if there exists an x such that $s(\phi)(\overline{x}, x)$, then there must exist a y with $\chi(x, y)$ and $t(\phi)(\overline{y}, y)$ follows by inductive hypothesis. □

Theorem 2.53. *If two theories are (strictly) bi-interpretable, then they are definitionally equivalent.*

Proof. Let T_1 and T_2 be two theories with signatures Σ_1 and Σ_2 respectively. Let t be the interpretation of T_1 in T_2 and s be the interpretation of T_2 in T_1. Let $\chi(x, y)$ be the formula that witnesses the isomorphism from $\mathcal{M}_1$ to $t^*(s^*(\mathcal{M}_1))$ for any model $\mathcal{M}_1$ of T_1, and let $\theta(x, y)$ be the formula that witnesses the isomorphism between $s^*(t^*(\mathcal{M}_2))$ and $\mathcal{M}_2$ for any model $\mathcal{M}_2$ of T_2. By Proposition 2.43, we only need to prove that T_1 and T_2 are relatively interpretably equivalent.

We describe a distinct relative interpretation t' of T_1 in T_2 as follows.

- for each relational symbol R in Σ_1, let $t'(R) := \forall \overline{y}(t(\chi)(\overline{x}, \overline{y}) \to t(R)(\overline{y}))$,
- for each functional symbol f in Σ_1, let $t'(f) := \forall \overline{y}(t(\chi)(\overline{x}, \overline{y}) \to t(f)(\overline{y}))$,
- for each constant symbol c in Σ_1, let $t'(c) := \forall y(t(\chi)(x, y) \to t(c)(y))$.

We prove by induction that (Lemma A) for any Σ_1-formula ϕ, $T_2 \vdash \forall \overline{x} \forall \overline{y}$ $(t(\chi)(\overline{x}, \overline{y}) \to (t'(\phi)(\overline{x}) \leftrightarrow t(\phi)(\overline{y})))$.

- $\phi := R(\overline{x})$. By the definition of t', we only need to show that $T_2 \vdash \forall \overline{x} \forall \overline{y}(t(\chi)(\overline{x}, \overline{y}) \to (\forall \overline{z}(t(\chi)(\overline{x}, \overline{z}) \to t(R)(\overline{z})) \leftrightarrow t(R)(\overline{y})))$ which is clear as T_1 proves that χ is a bijection over the entire domain, and hence T_2 proves that $t(\chi)$ is a bijection.
- The cases for functional symbols and constants are similar.
- The cases for conjunction and negation are trivial.
- $\phi := \forall x \psi(\overline{x}, x)$. By inductive hypothesis, $T_2 \vdash \forall \overline{x} \forall x \forall \overline{y} \forall y(t(\chi)(\overline{x}, x; \overline{y}, y) \to (t'(\psi)(\overline{x}, x) \leftrightarrow t(\psi)(\overline{y}, y)))$. Since T_1 proves that χ is a bijection, T_2 proves that $t(\chi)$ is a bijection as well, so for any y there exists an x such that $t(\chi)(x, y)$. Thus, $T_2 \vdash \forall \overline{x} \forall \overline{y}(t(\chi)(\overline{x}, \overline{y}) \to (\forall x t'(\psi)(\overline{x}, x) \to \forall y t(\psi)(\overline{y}, y)))$. Similarly, for any x there exists a y such that $t(\chi)(x, y)$, so $T_2 \vdash \forall \overline{x} \forall \overline{y}(t(\chi)(\overline{x}, \overline{y})$ $\to (\forall y t(\psi)(\overline{y}, y) \to \forall x t'(\psi)(\overline{x}, x)))$.

As a corollary, for any Σ_1-sentence χ, $T_2 \vdash t'(\chi) \leftrightarrow t(\chi)$. So for any χ, if $T_1 \vdash \chi$, then $T_2 \vdash t(\chi)$ since t is an interpretation, and then $T_2 \vdash t'(\chi)$. Thus, we have shown that t' is also an interpretation of T_1 in T_2.

Since T_1 interprets T_2 by s, Lemma A then gives us that (Lemma B) $T_1 \vdash \forall\overline{x}\forall\overline{y}(s(t(\chi))(\overline{x},\overline{y}) \to (s(t'(\phi))(\overline{x}) \leftrightarrow s(t(\phi))(\overline{y})))$. Notice that since χ is a homotopy from identity to $s \circ t$, by Lemma 70, we have (Lemma C) $T_1 \vdash \forall x_1 \forall x_2 \forall y_1 \forall y_2$
$((\chi(x_1, y_1) \wedge \chi(x_2, y_2)) \to (\chi(x_1, x_2) \leftrightarrow s(t(\chi))(y_1, y_2)))$.

We prove that (Lemma D) $T_1 \vdash \forall x \forall y(\chi(x, y) \leftrightarrow s(t(\chi))(x, y))$ (illustrated by the diagram below). Reason within T_1, and assume that $s(t(\chi))$ holds for y_1, y_2. Now, there must be an x_1 and x_2 such that $\chi(x_1, y_1)$ and $\chi(x_2, y_2)$, since χ is a bijection. Thus, we have $\chi(x_1, x_2)$ by Lemma C. However, since χ is a function, $x_2 = y_1$, and therefore $\chi(y_1, y_2)$ as well. Assume that χ holds for x_2, y_2. Then since $s(t(\chi))$ is a bijection, there must be a y_1 such that $s(t(\chi))(y_1, y_2)$. And since χ is also a bijection, there must be an x_1 such that $\chi(x_1, y_1)$. Thus, by Lemma C, we have $\chi(x_1, x_2)$. But since χ is a function, $x_2 = y_1$, and therefore we have $s(t(\chi))(x_2, y_2)$. Thus, we have shown that $T_1 \vdash \forall x \forall y(\chi(x, y) \leftrightarrow s(t(\chi))(x, y))$.

$$\begin{tikzcd}
x_1 \arrow[r, "\chi"] \arrow[d, "\chi"] & x_2 \arrow[d, "\chi"] \\
y_1 \arrow[r, "s(t(\chi))"] & x_2
\end{tikzcd}$$

Combining Lemma D with Lemma B, we have $T_1 \vdash \forall\overline{x}\forall\overline{y}(\chi(\overline{x},\overline{y}) \to (s(t'(\phi))(\overline{x}) \leftrightarrow s(t(\phi))(\overline{y})))$. By Lemma 2.52, we have that $T_1 \vdash \forall\overline{x}\forall\overline{y}(\chi(\overline{x},\overline{y}) \to (\phi(\overline{x}) \leftrightarrow s(t(\phi))(\overline{y})))$. Thus, we have that $T_1 \vdash \forall\overline{x}(\phi(\overline{x}) \leftrightarrow s(t'(\phi))(\overline{x}))$ for any Σ_1-formula ϕ. Thus, for any model $\mathcal{M}_1$ of T_1, and any $R \in \Sigma_1$, $R^{\mathcal{M}_1} = s(t'(R))^{\mathcal{M}_1} = R^{(t')^*(s^*(\mathcal{M}_1))}$, and similarly for functional symbols and constants. Therefore, $\mathcal{M}_1 = (t')^*(s^*(\mathcal{M}_1))$.

To prove that for any $\mathcal{M}_2$, $\mathcal{M}_2 = (s)^*((t')^*(\mathcal{M}_2))$, we show that s^* is surjective. Notice that for any $\mathcal{M}_2$ of T_2, $\theta(x, y)$ defines a bijection f on the domain of $\mathcal{M}_2$. Let $f^{-1}\mathcal{M}$ be the model defined by $R^{f^{-1}\mathcal{M}} = \{\overline{x} | f(\overline{x}) \in R^{\mathcal{M}}\}$. Since θ is a homotopy from identity to $t \circ s$, $f^{-1}s^*(t^*(\mathcal{M}_2)) = \mathcal{M}_2$. Since f^{-1} and s^* commute, $s^*(f^{-1}(t^*(\mathcal{M}_2))) = \mathcal{M}_2$. Note that $f^{-1}(t^*(\mathcal{M}_2))$ is isomorphic to $t^*(\mathcal{M}_2)$ by the function f, and therefore is also a model of T_1. Thus, s^* is surjective. Hence, for any model $\mathcal{M}_2$ of T_2, $\mathcal{M}_2 = (s)^*(\mathcal{M}_1)$ for some model $\mathcal{M}_1$ of T_1. Therefore, $(s)^*((t')^*(\mathcal{M}_2)) = (s)^*((t')^*(s^*(\mathcal{M}_1))) = s^*(\mathcal{M}_1) = \mathcal{M}_2$. □

Now we show that Morita bi-interpretability implies Morita interpretably equivalence. Recall the following item in Schema 2.28:

> T_1 and T_2 are **bi-interpretable** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1 in T_2 such that for any model $\mathcal{M}_1$ of T_1, $s^\dagger(t^\dagger(\mathcal{M}_1))$ is provably isomorphic to $\mathcal{M}_1$ and for any model $\mathcal{M}_2$ of T_2, $t^\dagger(s^\dagger(\mathcal{M}_2))$ is **provably isomorphic** to $\mathcal{M}_2$.

A strict definition of Morita bi-interpretability has not been given in the literature as it is not clear what one means by saying that that $t^\dagger(s^\dagger(\mathcal{M}_2))$ is **provably isomorphic** to $\mathcal{M}_2$ in the many-sorted case. Nevertheless, the following proof will only use the fact that $t^\dagger \circ s^\dagger$ is naturally isomorphic to the identity functor, and the same holds for $s^\dagger \circ t^\dagger$.

Schema 2.54. We say that T_1 and T_2 are **naturally congruent** if there is an **interpretation** t of T_2 in T_1 and an **interpretation** s of T_1 in T_2 such that $s^\dagger \circ t^\dagger \cong id$, and $t^\dagger \circ s^\dagger \cong id$.

Definition 2.55. We say that T_1 and T_2 are Morita naturally congruent if there is a Morita interpretation t of T_2 in T_1 and a Morita interpretation s of T_1 in T_2 such that there are functors $t^\dagger$ and $s^\dagger$ compatible with t and s respectively, such that $s^\dagger \circ t^\dagger \cong id$, and $t^\dagger \circ s^\dagger \cong id$.

Theorem 2.56. *If two theories are Morita naturally congruent, then they are Morita interpretably equivalent.*

Proof. Let T_1 and T_2 be two theories with signatures Σ_1 and Σ_2 respectively. Let $t = (T_2, T_2^+, T_1)$ be the interpretation of T_1 in T_2 and $s = (T_1, T_1^+, T_2)$ be the interpretation of T_2 in T_1. Let $t^\dagger$ and $s^\dagger$ be the corresponding compatible functors. Let $\chi_\sigma(x^\sigma, y^\sigma)$ be the formula that witnesses the isomorphism from $\sigma^{\mathcal{M}_1}$ to $\sigma^{t^\dagger(s^\dagger(\mathcal{M}_1))}$ for any model $\mathcal{M}_1$ of T_1, and let $\theta_\delta(x^\sigma, y^\sigma)$ be the formula that witnesses the isomorphism from $\sigma^{\mathcal{M}_2}$ to $\sigma^{s^\dagger(t^\dagger(\mathcal{M}_2))}$ for any model $\mathcal{M}_2$ of T_2.

We only need to prove that T_1 and T_2 are Morita interpretably equivalent. As χ gives us a natural isomorphism between functors $t^\dagger \circ s^\dagger$ and id, and θ a natural isomorphism between functors $s^\dagger \circ t^\dagger$ and id, $t^\dagger$ and $s^\dagger$ witness the equivalence of categories $Mod(T_1)$ and $Mod(T_2)$. Using Lemma 2.32, we obtain an isomorphism f between categories $Mod(T_1)$ and $Mod(T_2)$ which respects $s^\dagger$ and $t^\dagger$ modulo isomorphism.

We then show that f is compatible with s. Let $\mathcal{M}_1$ be an arbitrary model of T_1. Let $\mathcal{M}_1^+$ be an arbitrary expansion of $\mathcal{M}_1$ in T_1^+. Notice that $[\mathcal{M}_1^+|_{\Sigma_2}] = [s^\dagger(\mathcal{M}_1)]$, and since f respects $s^\dagger$ modulo isomorphism, we have $f(\mathcal{M}_1) \cong \mathcal{M}_1^+|_{\Sigma_2}$. Substituting $f(\mathcal{M}_1)$ for $\mathcal{M}_1^+|_{\Sigma_2}$ in $\mathcal{M}_1^+$ for the interpretation of symbols in Σ_2, we have a new model $\mathcal{M}'$. Since $f(\mathcal{M}_1) \cong \mathcal{M}_1^+|_{\Sigma_2}$, $\mathcal{M}' \cong \mathcal{M}_1^+$ and hence is a model of T_1^+. And by construction, $\mathcal{M}'|_{\Sigma_1} = \mathcal{M}_1$ and $\mathcal{M}'|_{\Sigma_2} = f(\mathcal{M}_1)$. Thus, f is compatible with s. That f^{-1} is compatible with t is proved similarly.

□

Presumably, any reasonable definition of Morita bi-interpretability should imply Morita natural congruence.

Thesis 2.57. *If two theories are Morita bi-interpretable, then they are Morita naturally congruent.*

Thus we have the result that Morita bi-interpretability implies Morita interpretably equivalence, by Theorem 2.56.

Thus, we answer all open questions in the diagram raised in Meadows (2024) as follows,

$$\begin{tikzcd}
\text{Relatively Interpretably Equiv} & \text{Morita Interpretably Equiv} & \text{Categorical Iso} \\
\text{Bi-interpretable} \arrow[u, "\times"] \arrow[ur, "\checkmark"] & \text{Morita Bi-interpretable} \arrow[u, "\checkmark"'] & \text{Categorical Equiv} \\
\text{Iso-congruence} \arrow[u, "\times"] \arrow[ur, "\times"] \arrow[urr, "\times"] & \text{Morita-congruence} \arrow[u, "\times"'] \arrow[ur, "\times"] & \text{Objective Equiv}
\end{tikzcd}$$

In fact, we can add a further column to the diagram:

$$\begin{tikzcd}
\text{Strictly Interpretably Equiv} & \text{Categorical Iso} \\
\text{Strictly Bi-interpretable} \arrow[u, "\checkmark"] & \text{Categorical Equiv} \\
\text{Strictly Iso-congruence} \arrow[u, "\times"] \arrow[ur, "\times"] & \text{Objective Equiv}
\end{tikzcd}$$

Lemma 2.58. *Two theories are strictly interpretably equivalent iff they are relatively interpretably equivalent.*

Proof. The direction from left to right is trivial. For the other direction. Let t and s be relative interpretations of T_1 in T_2 and T_2 in T_1 respectively. t and s must be constant on domains, and hence are strict interpretations. For if not, then there is a model $\mathcal{M}$ of T_1 such that $t^\dagger(s^\dagger(\mathcal{M}))$ has a smaller domain than $\mathcal{M}$, and hence $\mathcal{M} \neq t^\dagger(s^\dagger(\mathcal{M}))$. This contradicts the fact that T_1 and T_2 are relatively interpretably equivalent. □

Theorem 2.59. *If two theories are strictly bi-interpretable, then they are strictly interpretably equivalent.*

Proof. By Theorem 2.53, Proposition 2.43 and Lemma 2.58. □

Proposition 2.60. *There are strictly iso-congruent theories which are not categorically equivalent.*

Proof. The counterexample in the proof of Proposition 2.48 is not only iso-congruent, but strictly iso-congruent, which, therefore, suffices to prove the proposition. □

Chapter 3

Representation as Definition

3.1 Introduction

We have seen all kinds of different criteria of theoretical equivalence. But there are two distinct, though often conflated, senses in which people use the notion of theoretical equivalence: structural equivalence vs. representational equivalence. Structural equivalence tends to capture the idea that the two theories or models have the same structure, while representational equivalence means that they "say the same thing about the world". All previous criteria can be seen as legitimate specifications of structural equivalence, each of which captures a certain aspect in which theories and models share or differ in their structure. But there is a gap in how such criteria of structural equivalence relate to the notion of representational equivalence: sharing certain structures does not necessarily mean saying the same thing about the world, particularly when it is not clear which part of the theories/models are actually used for representation and how specifically they are used. For instance, Manet's *Olympia* and Titian's *Venus* both depict Venus lying on a bed with the same pose, but the two paintings convey drastically different themes. Similarly, while two theories may have equivalent categories of models, it is not clear how, if purely for this reason, they are representationally equivalent, as it is not clear what role the category of models plays in the process of representation.

The goal of this chapter is to develop the notion of representation in a formal manner and show how different notions of structural equivalence may be seen as notions of representational equivalence for certain methods of representation.

3.2 Representation as Definition

Not only do we have different representational tools, but also different methods of using these tools. Here, I call a method on how we use representational tools to represent a *representational protocol.* A prevalent representational protocol is *representation by definition.*[1]

The basic idea is to draw inspiration from the late logical empiricist proposal (e.g., Carnap, 1958): to represent the observables is equivalent to *defining* symbols *which are assumed to refer to the observables.* The idea can be generalized from observable quantities to all kinds of quantities that we wish to represent. Thus, for a model to represent a quantity is simply for it to define the symbol that we assume to refer to that quantity. (To simplify, I use the phrase "represent a quantity" for "represent facts about a quantity") Thus, claims about the models lead to claims about the quantity through the definition of interpreted symbols.[2] In quantum mechanics, for instance, we assume the symbol $\langle p \rangle$ is interpreted as the expectation value of the momentum of a particle. To represent the

[1] See Suppes, 1957, chapter 8; Suppes, 2002, Section 3.1.

[2] The current framework allows only parts of our models to represent reality, thus accommodating various positions regarding the question of scientific realism (realism, constructive empiricism (van Fraassen, 1980), qualified realism (Dewar, 2015), etc.) whether the acceptance of a scientific theory requires the acceptance of its full truth.

dynamics of the expected momentum, therefore, we simply define this symbol $\langle p \rangle$ in our structure of Hilbert space as (in the one-dimensional case):

$$(*)\ \langle p \rangle = -i\hbar \int (\psi^* \tfrac{\partial}{\partial x} \psi) dx$$

The Schödinger equation which describes the evolution of wave functions in Hilbert space then leads to hypotheses about the dynamics of expected momentum, which can be further tested by experiments. In this manner, the Hilbert space of wave functions is used to represent (the evolution of) the expected momentum, by definition $(*)$.

In gist, given a set of quantities Q, a set of symbols Λ, and a model $\mathcal{M}$, I propose the following equivalence as illustrated by Figure 2:

> **The Triangular Equivalence of Representation and Definition**: Assuming that Λ refers to Q, representation of Q by $\mathcal{M}$ can be equivalently transformed as definitions of Λ by $\mathcal{M}$, and vice versa.

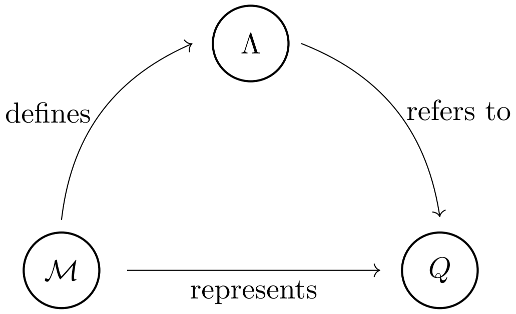

Figure 3.1: The Triangle of Representation.

The Triangular Equivalence of Representation and Definition suggests that every representation ($\mathcal{M}$ represents Q) can be standardized as a two-step process ($\mathcal{M}$ defines Λ which refers to Q). Essentially, we add a mediator between models and reality, that is, a set Λ of symbols which we *assume* refer to the physical quantities Q we wish to represent. Thus, instead of directly discussing how models say things about reality, we can discuss how models *define* symbols that we assume to have factual contents.

If one is not so comfortable with the expression "Λ refers to Q", e.g., due to its realistic flavor, we can also talk instead about how Λ *approximates* or *matches* phenomena or experiences . In general, we only need to assume here that Λ is related to what is represented in a manner that symbols in Λ *obtain factual or cognitive contents*, so that by defining symbols in Λ, the theory or model will be able to make statements with factual contents as well. We simply remain neutral on the question of *through what mechanism* Λ obtains its factual contents.[3]

In the following, I simply make it a working hypothesis that we can identify a set of symbols Λ with factual content (say, *the factual signature*), as is commonly done in physical practices. According to representation as definition, to represent a quantity q is just to define the corresponding symbol in the factual signature.

There are different views about definability, in particular, different views about how many resources a theory/model can use to define new symbols. This essentially corresponds to different notions

[3] For potential answers to this question, see, e.g., Hughes (1997), Suárez (2004), Contessa (2007), Frigg (2022).

of "definitional extension", as we mentioned in the previous chapter, specifically about what kind of structures are free lunches for a theory/model. One's position could vary from the extremely deflationist answer that they are only allowed to use symbols currently available (and primitive) in the signature, to the extremely inflationist answer that they are allowed to use whatever is definable in first-order set theory. In the middle are various positions such as first-order definitional extension, different notions of Morita extension with different sets of permissible operations etc. Here we will first give general schemas about how notions related to representation are defined given a certain notion of definability, and then give examples of how such schemas are used.

Suppose we have a notion of a definitional extension of a theory. The notion of representation can be defined schematically as follows.

Definition 3.1. Let T be a theory with signature Σ. We assume that Σ and Λ are disjoint. A *representation* of a set of factual symbols Λ by a theory T is a tuple $\Lambda \xrightarrow{\delta} T : \langle T, T^+, \Lambda\rangle$, where:

- T^+ is a definitional extension of T with signature Σ^+ and $\Lambda \subseteq \Sigma^+\backslash\Sigma$.
- δ is the *representational context*, which is the set of definitions which specifies the definitions of new symbols of T^+ by T.

The *representational content* of $\Lambda \xrightarrow{\delta} T$ (in notation $(\Lambda \xrightarrow{\delta} T)|_\Lambda$) is defined as $T^+|_\Lambda$ which is the set of semantical consequences of T^+ in the language with signature Λ.

When Λ is clear from the context, we may also write T^δ for $\Lambda \xrightarrow{\delta} T$.

Then we have the notion of equivalence regarding representational contents:

Definition 3.2. Two representations $\Lambda \xrightarrow{\delta_1} T_1$ and $\Lambda \xrightarrow{\delta_2} T_2$ are *equivalent in contents* if $(\Lambda \xrightarrow{\delta_1} T_1)|_\Lambda = (\Lambda \xrightarrow{\delta_2} T_2)|_\Lambda$.

We can also define the notion of equivalence regarding representational capacities.

Definition 3.3. Two theories T_1 and T_2 have *equivalent representational capacities* if for any set of constants Λ (disjoint with both Σ_1 and Σ_2), for any representation $\Lambda \xrightarrow{\delta_1} T_1$, there is a representation $\Lambda \xrightarrow{\delta_2} T_2$ such that $(\Lambda \xrightarrow{\delta_1} T_1)|_\Lambda = (\Lambda \xrightarrow{\delta_2} T_2)|_\Lambda$, and vice versa.

Similar definitions can also be given using models.

Definition 3.4. Let $\mathcal{M}$ be a model with signature Σ. We assume that Σ and Λ are disjoint. A *representation* of a set of factual symbols Λ by a model $\mathcal{M}$ is a tuple $\Lambda \xrightarrow{\delta} \mathcal{M} : \langle \mathcal{M}, \mathcal{M}^+, \Lambda\rangle$, where:

- $\mathcal{M}^+$ is a definitional extension of $\mathcal{M}$ with signature Σ^+ and $\Lambda \subseteq \Sigma^+\backslash\Sigma$.
- δ is the *representational context*, which is the set of definitions which specifies the definitions of new symbols of $\mathcal{M}^+$ by $\mathcal{M}$.

The *representational content* of $\Lambda \xrightarrow{\delta} \mathcal{M}$ (in notation $(\Lambda \xrightarrow{\delta} \mathcal{M})|_\Lambda$) is defined as $\mathcal{M}^+|_\Lambda$ which is the reduct of $\mathcal{M}^+$ in the language with signature Λ.

When Λ is clear from the context, we may also write $\mathcal{M}^\delta$ for $\Lambda \xrightarrow{\delta} \mathcal{M}$.

Definition 3.5. Two representations $\Lambda \xrightarrow{\delta_1} \mathcal{M}_1$ and $\Lambda \xrightarrow{\delta_2} \mathcal{M}_2$ are *equivalent in contents* if $(\Lambda \xrightarrow{\delta_1} \mathcal{M}_1)|_\Lambda \cong (\Lambda \xrightarrow{\delta_2} \mathcal{M}_2)|_\Lambda$.

Definition 3.6. Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ have *equivalent representational capacities* if for any set of constants Λ (disjoint with both Σ_1 and Σ_2), for any representation $\Lambda \xrightarrow{\delta_1} \mathcal{M}_1$, there is a representation $\Lambda \xrightarrow{\delta_2} \mathcal{M}_2$ such that $(\Lambda \xrightarrow{\delta_1} \mathcal{M}_1)|_\Lambda = (\Lambda \xrightarrow{\delta_2} \mathcal{M}_2)|_\Lambda$, and vice versa.

Note that in the above case, some symbols of our theories may not be used for representation. But very often we are interested in a particular case where every symbol in a theory is interpreted as having factual contents. In this case, we can define the full representational content of a theory as follows.

Definition 3.7. Let Σ_1 and Σ_2 be two signatures. We say a function f from Σ_1 to Σ_2 is a *rewriting*, if f sends sort symbols to sort symbols, constants, functional or relational symbols to constants, functional or relational symbols with corresponding arities. We use f^* to denote the induced function from Σ_1 formulas to Σ_2 formulas.

Definition 3.8. Let T be a Σ-theory, T_1 a definitional extension of T with signature Σ_1, and T_2 a definitional extension of T with signature Σ_2. We say that T_1 *is included in* T_2 if there is a rewriting f from Σ_1 to Σ_2 such that:

- f is identity on Σ;
- $T_1 \models \phi$ iff $T_2 \models f^*(\phi)$ for any Σ_1-formula ϕ.

We say that T_1 and T_2 are notational variants of each other if f in the above definition is a bijection.

Definition 3.9. We say that a definitional extension T^+ of T is *maximal*, if every definitional extension of T is included in T^+.

Note that such a maximal definitional extension does exist.

Definition 3.10. We say that a definitional extension T^+ of T is *canonical*, if for any potential definition in T, there is a unique new symbol in T^+ that is defined by that definition.

Thesis 3.11. *Every canonical definitional extension of a theory is maximal.*

Note that this is a thesis, since the term "definitional extension" is still a placeholder at this point. However, it is intuitive why this thesis should be true: Let T^+ be a canonical definitional extension of T, and let T' be an arbitrary definitional extension of T. Now T' can be included by T^+ by mapping a new symbol in T' to the unique symbol in T^+ that are defined by the same potential definition.

Definition 3.12. For a theory T, we say that $T_0, T_1, ...$ is a *sequence of maximal definitional extensions* if $T_0 = T$ and T_{i+1} is a maximal definitional extension of T_i for each natural number i.

We say that T^+ is a *full definitional extension* if there is a sequence of maximal definitional extensions $T_0, T_1, ...$ such that $T^+ \equiv \bigcup_i T_i$.

Intuitively, a full definitional extension of a theory T is simply what we get when we add all possible definitions into T. To justify that indeed we cannot add new definitions to a full definitional extension, we propose the following thesis.

Thesis 3.13. *Let T^+ be a full definitional extension of T. Then any definitional extension of T^+ is included in T^+.*

It is intuitive why this thesis should be true: for any potential formula that defines a new symbol in T^+, it must belong to some T_i, and thus, by construction, the symbol defined by that formula will be included in T_{i+1}.

Definition 3.14. Let T_1 be a Σ_1-theory and T_2 be a Σ_2-theory. We say that T_1 and T_2 are *notational variants* of each other if there is a bijective rewriting f from Σ_1 to Σ_2 such that $T_1 \models \phi$ iff $T_2 \models f^*(\phi)$ for any ϕ in the language of T_1.

Definition 3.15. Let T_1 and T_2 be two theories. We say that T_1 and T_2 are *objectively fully equivalent in contents*, if T_1 has a full definitional extension T_1^* and T_2 a full definitional extension T_2^* such that T_1^* and T_2^* are notational variants of each other.

Full equivalence in contents really means that the two theories reach an objective agreement about the final picture about the world. However, we may also weaken this criterion to require *intersubjective* agreement only.

Definition 3.16. Let T_1 and T_2 be two theories. We say that T_1 and T_2 are *intersubjectivley fully equivalent in contents*, if there is a full definitional extension T_1^* of T_1 (with signature Σ_1^*) and a full definitional extension T_2^* of T_2 (with signature Σ_2^*), such that

- T_1^* represents T_2^* in the sense that, there is a representation $\Sigma_2^* \xrightarrow{\delta_1} T_1^*$ such that $T_2^* \subseteq (\Sigma_2^* \xrightarrow{\delta_1} T_1^*)|_{\Sigma_2^*}$; and T_2^* represents T_1^* in the sense that, there is a representation $\Sigma_1^* \xrightarrow{\delta_2} T_2^*$ such that $T_1^* \subseteq (\Sigma_1^* \xrightarrow{\delta_2} T_2^*)|_{\Sigma_1^*}$.
- The two representations above are *coordinated*, which might be understood as one of the following. Let $t^\dagger$ and $s^\dagger$ denote the functor from $Mod(T_1^*)$ to $Mod(T_2^*)$ and $Mod(T_2^*)$ to $Mod(T_1^*)$ generated by the two representations respectively.
  1. $t^\dagger \circ s^\dagger = id$ and $s^\dagger \circ t^\dagger = id$.
  2. $t^\dagger \circ s^\dagger$ is provably isomorphic to id and $s^\dagger \circ t^\dagger$ is provably isomorphic to id.
  3. $t^\dagger \circ s^\dagger$ is naturally isomorphic to id and $s^\dagger \circ t^\dagger$ is naturally isomorphic to id.
  4. $t^\dagger \circ s^\dagger(\mathcal{M}) \cong \mathcal{M}$ and $s^\dagger \circ t^\dagger(\mathcal{N}) \cong \mathcal{N}$ for any $\mathcal{M}$ and $\mathcal{N}$ in the respective categories.
  5. $t^\dagger \circ s^\dagger(\mathcal{M}) \equiv \mathcal{M}$ and $s^\dagger \circ t^\dagger(\mathcal{N}) \equiv \mathcal{N}$ for any $\mathcal{M}$ and $\mathcal{N}$ in the respective categories.

Now we illustrate the above schema by considering single-sorted first-order theories and adopting standard first-order definitional extension as our notion of definability. We show how previous criteria of theoretical equivalence can be related to different notions of representational equivalence.

We first focus on the case of common-extension criteria. We have the following result.

Lemma 3.17. *(With strict definitional extension) Suppose that T_2 is a definitional extension of T_1 and that T_1 is a definitional extension of T_0. Then T_2 is a definitional extension of T_0.*

Proof. Suppose that T_1 is a definitional extension of T_0, then T_0 strictly interprets T_1 and hence there is a strict translation f from formulas of T_1 to formulas of T_0. Suppose that the new symbols in T_2 are defined by T_1 with the definition δ. Then for any new symbol s of T_2, we can define s using $f(\delta(s))$ by T_0. It is easy to check that T_2 is a definitional extension of T_0 via such definitions. □

Lemma 3.18. *(With strict definitional extension) T^+ is a maximal definitional extension of T iff T^+ is a full definitional extension of T.*

Proof. Assume that T^+ is a maximal definitional extension of T. Let $T_0 = T$ and $T_i = T^+$ for $i > 0$. Note that since T^+ is a maximal definitional extension of T^+ itself, such a sequence $T_0, T_1, ...$ is a sequence of maximal definitional extensions. Thus, $T^+ = \bigcup_{i \in \mathbb{N}} T_i$ is a full definitional extension of T.

Conversely, assume that T^+ is a full definitional extension of T. Let $T^+ = \bigcup_{i \in \mathbb{N}} T_i$ be a sequence of maximal definitional extensions of T. By Lemma 3.17, each T_i is a definitional extension of T. Since $T_i \subseteq T_{i+1}$ for each i, the definitions for a symbol are consistent in all T_i. Then T^+ is also a definitional extension of T by defining all symbols in T_i for each i at once. Note that T^+ is maximal since for any formula ϕ of T^+ that potentially defines a new symbol s, ϕ must be a formula of some T_i, and thus such a definition of s can be included in T_{i+1}. □

Theorem 3.19. *(With standard definitional extension) Assume that T_1 is a Σ_1-theory and T_2 a Σ_2-theory. T_1 and T_2 are fully equivalent in contents iff they are standard definitionally equivalent.*

Proof. Suppose that T_1 and T_2 are fully equivalent in contents. Then there is a full definitional extension T_1^* of T_1 and a full definitional extension T_2^* of T_2 such that T_1^* and T_2^* are notational variants of each other. Let the notational variance be witnessed by the bijective rewriting f from Σ_2 to Σ_1. Without loss of generality, assume that T_2^* and T_1^* have disjoint signatures. Then for each symbol $s \in \Sigma_2$, we define s as $f(s)$ in T_1^* to obtain T_3. Similarly, T_3 can be obtained by defining s as $f^{-1}(s)$ for each $s \in \Sigma_1$ by T_2^*. Thus, T_3 is both a definitional extension of T_1^* and T_2^*. By Lemma 3.18, T_1^* and T_2^* are also (maximal) definitional extensions of T_1 and T_2 respectively. By Lemma 3.17, T_3 is a common definitional extension of T_1 and T_2. Thus, T_1 and T_2 are definitionally equivalent.

Conversely, suppose that T_1 and T_2 are definitionally equivalent. Then there is a definitional extension T_3 of T_1 and T_2. Let T_1^+ be a full definitional extension of T_1 and T_2^+ be a full definitional extension of T_2. Then $T := T_1^+ \cup T_2^+ \cup T_3$ will be a common definitional extension of both T_1 and T_2. Since any definitional extension of T_1 is included in T_1^+, it is also included in T. Thus, T is a maximal definitional extension of T_1. Thus, by Lemma 3.17, T is a full definitional extension of T_1. Similarly, T is a full definitional extension of T_2. Thus, T_1 and T_2 are fully equivalent in contents. □

Now we turn to the case of coordinated-interpretation criteria. We have the following result.

Theorem 3.20. *Assume that T_1 is a Σ-theory and T_2 a Σ'-theory. T_1 and T_2 are equivalent in representational capacities iff they are mutually interpretable (with relative interpretation).*

Proof. First, assume that T_1 and T_2 have equivalent representational capacities, we prove that they are mutually interpretable. By symmetry, it suffices to prove that T_1 is interpretable in T_2.

Let Σ^* be a disjoint copy of Σ witnessed by g. Assume that Σ^* is also disjoint from Σ', and let δ be a representational context which defines $x \in \Sigma^*_{\mathcal{C}}$ as $g(x)$. Then $\Sigma^*_{\mathcal{C}} \xrightarrow{\delta} T_1$ is a representation of $\Sigma^*_{\mathcal{C}}$ by T_1. By assumption, there is a representation $\Sigma^*_{\mathcal{C}} \xrightarrow{\delta'} T_2$ of $\Sigma^*_{\mathcal{C}}$ by T_2 such that $\Sigma^*_{\mathcal{C}} \xrightarrow{\delta} T_1$ and $\Sigma^*_{\mathcal{C}} \xrightarrow{\delta'} T_2$ are equivalent representations, i.e. $(\Sigma^*_{\mathcal{C}} \xrightarrow{\delta} T_1)|_{\Sigma^*_{\mathcal{C}}} \cong (\Sigma^*_{\mathcal{C}} \xrightarrow{\delta'} T_2)|_{\Sigma^*_{\mathcal{C}}}$.

Note that $(\Sigma^*_{\mathcal{C}} \xrightarrow{\delta} T_1)|_{\Sigma^*_{\mathcal{C}}}$ is a disjoint copy of T_1. Therefore, T_1 is interpretable in T_2.

Then assume that T_1 and T_2 are mutually interpretable. We prove that T_1 and T_2 have equivalent representational capacities. By symmetry, it suffices to prove that for any representation $\Lambda \xrightarrow{\delta} T_1$, there is an equivalent representation $\Lambda \xrightarrow{\delta'} T_2$.

By assumption, there is a $\mathcal{L}_{\Sigma^*}$ disjoint copy T_1^* of T_1 and a representation $\Sigma^*_{\mathcal{C}} \xrightarrow{\gamma} T_2$ such that $(\Sigma^*_{\mathcal{C}} \xrightarrow{\gamma} T_2)|_{\Sigma^*_{\mathcal{C}}} \cong T_1^*$. (Note that this ensures that Σ^* and Σ' only differ in constants.) Let g be the function which witnesses that Σ^* is a disjoint copy of Σ.

Now, let $\Lambda \xrightarrow{\delta} T_1$ be an arbitrary representation. Let $\delta' = \delta[x \mapsto g(x)]$ be a representational context which substitutes every constant of Σ that appeared in definitions with its copies. Then $\Lambda \xrightarrow{\delta'} T_1^*$ is an equivalent representation with $\Lambda \xrightarrow{\delta} T_1$, i.e., $(\Lambda \xrightarrow{\delta} T_1)|_\Lambda \cong (\Lambda \xrightarrow{\delta'} T_1^*)|_\Lambda$.

Now δ' defines constants of Λ in $\mathcal{L}_{\Sigma^*}$. And γ defines constants in Σ^* in $\mathcal{L}_{\Sigma'}$. Since Σ^* and Σ' only differ in constants, we can combine δ' and γ to have a representational context $\gamma \circ \delta'$ that defines symbols in Λ by $\mathcal{L}_{\Sigma'}$.

Then we have $(\Lambda \xrightarrow{\gamma\circ\delta'} T_2)|_\Lambda \cong (\Lambda \xrightarrow{\delta'} ((\Sigma_{\mathcal{C}}^* \xrightarrow{\gamma} T_2)))|_\Lambda \cong (\Lambda \xrightarrow{\delta'} T_1^*)|_\Lambda \cong (\Lambda \xrightarrow{\delta} T_1)|_\Lambda$. Therefore, $\Lambda \xrightarrow{\gamma\circ\delta'} T_2$ is an equivalent representation with $\Lambda \xrightarrow{\delta} T_1$.

□

Further, the hierarchy of coordinated interpretation criteria corresponds to the hierarchy of intersubjectively full equivalence as defined in Definition 3.16. We prove the case of strict bi-interpretability for illustration.

Lemma 3.21. *If T_1 and T_2 are strictly bi-interpretable by interpretation s and t, then there is a Σ_1-formula $\chi(x, y)$ such that T_1 proves that $\chi(x, y)$ is a bijection and that for each Σ_1-formula ϕ, we have that*

$$T_2 \vdash \forall x_0, \dots, x_n \forall y_0, \dots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (\phi(x_0, \dots, x_n) \ \leftrightarrow \ s(t(\phi))(y_0, \dots, y_n)) \Big)$$

Proof. Easily proved by induction as s and t preserve conjunction, negation and quantification. □

Theorem 3.22. *T_1 and T_2 are strictly bi-interpretable iff they are intersubjectively fully equivalent in contents (where we choose item 2 in Definition 3.16 as the requirement of coordination).*

Proof. Assume that T_1 and T_2 are intersubjectively fully equivalent in contents. Let T_1^*, T_2^* be the full definitional extensions of T_1 and T_2 respectively, and let $\Sigma_2^* \xrightarrow{\delta_1} T_1^*$ and $\Sigma_1^* \xrightarrow{\delta_2} T_2^*$ be the representations that witness their intersubjective equivalence. Let t and s be the corresponding translations of T_1^* in T_2^* and T_2^* in T_1^*.

By Lemma 3.18, T_1^* will also be a definitional extension of T_1 and consequently $\Sigma_2^* \xrightarrow{\delta_1} T_1^*$ is also a definitional extension of T_1. Since $T_2^* \subseteq (\Sigma_2^* \xrightarrow{\delta_1} T_1^*)|_{\Sigma_2^*}$, and $T_2^*|_{\Sigma_2} \equiv T_2$, we have $T_2 \subseteq (\Sigma_2^* \xrightarrow{\delta_1} T_1^*)|_{\Sigma_2}$. Thus, $\langle T_1, \Sigma_2^* \xrightarrow{\delta_1} T_1^*, T_2 \rangle$ is an interpretation of T_2 in T_1. Similarly, $\langle T_2, \Sigma_1^* \xrightarrow{\delta_2} T_2^*, T_1 \rangle$ is an interpretation of T_1 in T_2.

Let k be the canonical translation from T_1^* to T_1 as in Lemma B.4 in the appendix. Then the translation corresponds to the interpretation $\langle T_1, \Sigma_2^* \xrightarrow{\delta_1} T_1^*, T_2 \rangle$ is $k \circ s$. Similarly, let l be the canonical translation from T_2^* to T_2. Then the translation corresponds to the interpretation $\langle T_2, \Sigma_1^* \xrightarrow{\delta_2} T_2^*, T_1 \rangle$ is $l \circ t$.

By assumption, there is a Σ_1^*-formula $\chi(x, y)$ such that T_1^* proves that $\chi(x, y)$ is a bijection and that for each $R \in \Sigma_1^*$

$$T_1^* \vdash \forall x_0, \dots, x_n \forall y_0, \dots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (R(x_0, \dots, x_n) \ \leftrightarrow \ s(t(R))(y_0, \dots, y_n)) \Big)$$

by Lemma B.5, $T_2^* \vdash \forall \overline{x}(l(t(R)) \leftrightarrow t(R))$. Thus, $T_1^* \vdash \forall \overline{x}(s(l(t(R))) \leftrightarrow s(t(R)))$.

Therefore,

$$T_1^* \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ s(l(t(R)))(y_0, \ldots, y_n)) \Big)$$

Then by Lemma B.5 and Lemma B.6, we have that

$$T_1 \vdash k(\forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ s(l(t(R)))(y_0, \ldots, y_n)) \Big))$$

.

and therefore for any $R \in \Sigma_1$, we have that

$$T_1 \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} k(\chi(x_i, y_i)) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ k(s(l(t(R))))(y_0, \ldots, y_n)) \Big)$$

.

Also by Lemma B.5 and Lemma B.6, we have T_1 proves that $k(\chi)$ is a bijection.

Similarly, there is a Σ_2^*-formula θ such that T_2 proves that $l(\theta)$ is a bijection and for any $R \in \Sigma_2$,

$$T_2 \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} l(\theta(x_i, y_i)) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ l(t(k(s(R))))(y_0, \ldots, y_n)) \Big)$$

.

Thus, we have that T_1 and T_2 are strictly bi-interpretable.

Now supppose that T_1 and T_2 are strictly bi-interpretable. Let $t = \langle T_1, T_1^+, T_2 \rangle$ and $s = \langle T_2, T_2^+, T_1 \rangle$ be the interpretations that witness the bi-interpretability. Let T_1^* be the canonical definitional extension of T_1 and T_2^* be the canonical definitional extension of T_2. By Thesis 3.11 and Lemma 3.18, T_1^* and T_2^* are full definitional extensions of T_1 and T_2 respectively. Let k and l be the canonical translations from T_1^* to T_1 and from T_2^* to T_2 respectively. Then $s \circ k$ and $t \circ l$ are strict translations from T_1^* to T_2^* and from T_2^* to T_1^* respectively.

By Lemma 3.21, there is a Σ_1-formula $\chi(x, y)$ such that T_1 proves that $\chi(x, y)$ is a bijection and that for each Σ_1-formula ϕ:

$$T_1 \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (\phi(x_0, \ldots, x_n) \ \leftrightarrow \ s(t(\phi))(y_0, \ldots, y_n)) \Big)$$

.

For any $R \in \Sigma_1^*$, we have that $T_1^* \vdash \forall \overline{x}(R \leftrightarrow k(R))$.

Therefore, for each $R \in \Sigma_1^*$, we have that:

$$T_1 \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ s(t(k(R)))(y_0, \ldots, y_n)) \Big)$$

.

Since l remains constant on symbols in Σ_2, we have that,

$$T_1 \vdash \forall x_0, \ldots, x_n \, \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \chi(x_i, y_i) \ \rightarrow \ (R(x_0, \ldots, x_n) \ \leftrightarrow \ s(l(t(k(R))))(y_0, \ldots, y_n)) \Big)$$

.

And since T_1 proves that χ is a bijection, T_1^* also proves that χ is a bijection.

Similarly, we can prove that there is a Σ_2^*-formula $\theta(x, y)$ such that T_2^* proves that θ is a bijection and for any $R \in \Sigma_2^*$, we have that:

$$T_2 \vdash \forall x_0, \ldots, x_n \forall y_0, \ldots, y_n \Big(\bigwedge_{i<n} \theta(x_i, y_i) \;\to\; (R(x_0, \ldots, x_n) \;\leftrightarrow\; t(k(s(l(R))))(y_0, \ldots, y_n)) \Big)$$

.

Therefore, we have that T_1 and T_2 are intersubjectively fully equivalent in contents.

□

However, if we adopt the standard Morita extension as our notion of definitional extension, then the notion of structural equivalence (in this case, Morita equivalence) and full equivalence in contents are not equivalent. In particular, full equivalence in contents does not imply Morita equivalence.

Lemma 3.23. *(With standard Morita extension) Every canonical Morita extension is maximal.*

Proof. Easily checked by definition. □

Theorem 3.24. *(With standard Morita extension) There are theories T_1 and T_2 which are fully equivalent in contents but not Morita equivalent.*

Proof. Let T be an empty theory with a single sort σ_0 consisting of only two objects. Let T' be the theory with countably infinite sorts $\sigma_0, \sigma_1, ...$, which says that σ_{i+1} is the product sort of σ_i and σ_i with corresponding projection functions, and σ_0 has precisely two objects. We can prove by induction on the length of Morita descendence that T' cannot be included in any Morita descendence of T. In particular, we can prove by induction that for any sequence of Morita descendence $T, T_1 ..., T_n$ of T, sorts in T_n have at most 2^{2n} objects. But T' has sorts with an arbitrarily finitely large number of objects. Thus, T and T' are not Morita equivalent.

However, T and T' are fully equivalent in content. Let $T_0, ...$ be a canonical sequence of Morita extensions of T, and let T^+ be the resulting full definitional extension of T. Let $T_0', ...$ be a canonical sequence of Morita extensions of T', and let $(T')^+$ be the resulting full definitional extension of T'. We construct a rewriting f from $(\Sigma')^+$ to Σ^+ as follows:

- for symbols in Σ_0 (i.e., Σ), f is identity;
- for symbol σ_{i+1} in $\Sigma_0' - \Sigma_0$, $f(\sigma_{i+1})$ is the symbol in T_{i+1} defined as the product sort of $f(\sigma_i)$ and $f(\sigma_i)$. The projection functions are mapped accordingly.
- for symbol σ in some $\Sigma_{i+1}' - \Sigma_i'$ defined by some potential definition $\langle a, \phi \rangle$. Then there must be some T_n such that $\langle a, \phi \rangle$ is also a potential definition in T_n. Let $f(\sigma)$ be the symbol in T_{n+1} defined by $\langle f(a), f(\phi) \rangle$. The projection functions are mapped accordingly.

It is easy to check that f is a bijection and witnesses that T^+ and $(T')^+$ are notational variants of each other. □

This gives us a case where the intuitive notion of structural equivalence may not capture the equivalence of full representational contents. Indeed, if we take Morita extensions as free lunches and add nothing over and above the original theory, then we should believe that T and T' in the above proof say the same thing about the world even though they are not Morita equivalent.

We prove that if we modify the original definition of Morita extension to allow for *arbitrarily finite* products and coproducts (*generalized Morita extension*), then definitional equivalence indeed corresponds to full equivalence in contents.

Lemma 3.25. *(With generalized Morita extension) If T^+ is a full definitional extension of T, then T^+ and T are Morita equivalent.*

Proof. If T^+ is a full definitional extension of T, then T^+ is implicitly definable over T. By Beth's definability theorem for many-sorted logic (Appendix C), T^+ and T are Morita equivalent. □

Theorem 3.26. *(With generalized Morita extension) T_1 and T_2 are Morita equivalent iff they are fully equivalent in contents.*

Proof. Assume that T_1 and T_2 are fully equivalent in contents. Then there is a full definitional extension T_1^* of T_1 and a full definitional extension T_2^* of T_2 such that T_1^* and T_2^* are notational variants of each other. By Lemma 3.25, there is a Morita descendant $T_1^\dagger$ of both T_1 and T_1^* and a Morita descendant $T_2^\dagger$ of both T_2 and T_2^*. As T_1^* and T_2^* are notational variants of each other, we can construct a Morita descendant T' of $T_1^\dagger$ by defining T_2^* using T_1^* and mimicking the definitions of $T_2^\dagger$ in T_2^*. T' will also be a Morita descendant of $T_2^\dagger$ as it can also define T_1^* using T_2^* and mimicking the definitions of $T_1^\dagger$ in T_1^*. Thus, T' would be a common Morita descendant of T_1 and T_2.

Conversely, assume that T and S are Morita equivalent. Let $T, T_1, ..., T_n$ and $S, S_1, ..., S_m$ be two sequences of Morita extensions of T and S such that $T_n \equiv S_m$.

Let $T^+ = \bigcup_{i\in\mathbb{N}}\{T^i\}$ be a full definitional extensions of T ($T^0 := T$). And let $S^+ = \bigcup_{i\in\mathbb{N}}\{S^i\}$ be a full definitional extensions of S ($S^0 := S$). We may assume that $\Sigma(T_n) \cap \Sigma(T^+) = \Sigma(T)$ and $\Sigma(S_m) \cap \Sigma(S^+) = \Sigma(S)$, and that $\Sigma(T_n) \cap \Sigma(S_n) = \emptyset$, $\Sigma(T^+) \cap \Sigma(S^+) = \emptyset$, $\Sigma(T_n) \cap \Sigma(S^+) = \emptyset$, $\Sigma(S_m) \cap \Sigma(T^+) = \emptyset$.

Let $F_i = T_i \cup T^i$, for $1 \leq i \leq n$. Then F_{i+1} is a maximal definitional extension of F_i. Let $F_{n+j} := T^{n+j} \cup S_j$. Since F_n is a Morita descendant of T which contains S, F_{n+1} is a maximal definitional extension of F_n, and similarly F_{n+j+1} is a maximal definitional extension of F_{n+j}. Thus, $F := \bigcup_{i\in\mathbb{N}}\{F_i\}$ is a complete definitional extension of T. Similarly, let $G_i = S_i \cup S^i$, for $1 \leq i \leq m$. Let $G_{m+j} := S^{m+j} \cup T_j$. Thus, $G := \bigcup_{i\in\mathbb{N}}\{G_i\}$ is a full definitional extension of S. And by construction, $F \equiv G$. Thus, T and S are fully equivalent in contents. □

3.2.1 Application: The Hole Argument

In the general discussion of the hole argument, the principle of Leibniz Equivalence was raised to block the argument which says:

> "Leibniz Equivalence: Isometric models represent the same physical situation"

The formal framework of representation can be used to support this principle. We have the following theorem:

Theorem 3.27. *(With standard definitional extension) For any two isomorphic $\mathcal{L}_\Sigma$-models $\mathcal{M}_1$ and $\mathcal{M}_2$, the representation $\Lambda \xrightarrow{\delta} \mathcal{M}_1$ is equivalent to $\Lambda \xrightarrow{\delta} \mathcal{M}_2$ for any δ.*

Proof. We need to prove that $(\Lambda \xrightarrow{\delta} \mathcal{M}_1)|_\Lambda \cong (\Lambda \xrightarrow{\delta} \mathcal{M}_2)|_\Lambda$. It is enough to show that $\Lambda \xrightarrow{\delta} \mathcal{M}_1 \cong \Lambda \xrightarrow{\delta} \mathcal{M}_2$, or in short, $\mathcal{M}_1^\delta \cong \mathcal{M}_2^\delta$.

By definition, $Th(\mathcal{M}_1^\delta)$ is explicitly definable over $Th(\mathcal{M}_1)$. By Beth's definability theorem, $Th(\mathcal{M}_1^\delta)$ is also implicitly definable over $Th(\mathcal{M}_1)$. Thus, for any model $\mathcal{M}$ of $Th(\mathcal{M}_1^\delta)$ such that $\mathcal{M}|_{\Sigma_1} = \mathcal{M}_1$, we have $\mathcal{M} \cong \mathcal{M}_1^\delta$. By substituting $\mathcal{M}_1$ for $\mathcal{M}_2$ in $\mathcal{M}_2^\delta$, we have a new model $\mathcal{M}'$. Since $\mathcal{M}_1 \cong \mathcal{M}_2$, $\mathcal{M}' \cong \mathcal{M}_2^\delta$. $\mathcal{M}'$ is a model of $Th(\mathcal{M}_2^\delta)$ and hence a model of $Th(\mathcal{M}_1^\delta)$. By construction, $\mathcal{M}'|_{\Sigma_1} = \mathcal{M}_1$. Thus, by implicit definability, $\mathcal{M}' \cong \mathcal{M}_1^\delta$. Therefore, $\mathcal{M}_1^\delta \cong \mathcal{M}_2^\delta$. □

Note that while we only state the theorem in the context of standard definitional extension, the proof shows that it actually holds for any concept of definability that implies implicit definability.

So if, as Weatherall (2018) argues, in the case of general relativity the relevant criterion of isomorphism physicists have in mind is isometry, then by the above theorem, whenever a Lorentzian manifold is used to represent the spacetime in a representational context δ, the isometric manifold in the same representational context will be an equivalent representation. We will give a full discussion of the hole argument in Chapter 6 and Chapter 7.

3.2.2 Application: Representational Capacities

Weatherall and Bradley (2020) relate the structural content of a model with its representational capacities. They hold that "the representational capacities of mathematical objects are precisely those preserved by isomorphism" (p. 1230). The following thesis is reformulated as follows:

> Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ have equivalent representational capacities iff they are isomorphic.

The direction from the right to left can be proved:

Theorem 3.28. *If two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic, then they have equivalent representational capacities.*

Proof. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic, then they are both $\mathcal{L}_\Sigma$-models for some signature Σ. The conclusion easily follows from Definition 3.6, since for any δ, $\mathcal{M}_1^\delta \cong \mathcal{M}_2^\delta$, so there exists a representational context δ' (i.e. $\delta' = \delta$) such that $\mathcal{M}_1^\delta \cong \mathcal{M}_2^{\delta'}$. □

The direction from left to right, however, is not true. First of all, note that there is no guaranteed that two models with equivalent representational capacities share the same signature. For example, the standard model of natural numbers $\langle N, S, 0\rangle$ and its notational variant $\langle N, S', 0'\rangle$ will clearly have the same representaional capacities, but do not share the same signature and hence not isomorphic.

So a more charitable interpretation is to claim that any two $\mathcal{L}_S$ models with equivalent representational capacities are isomorphic. This, however, is false as well.

Theorem 3.29. *There are $\mathcal{L}_\Sigma$ models $\mathcal{M}_1$ and $\mathcal{M}_2$ that are not isomorphic but have equivalent representational capacities.*

Proof. Consider the models of integers $\mathcal{M}_1 = \langle \mathbb{Z}, +^{\mathcal{M}_1}, c^{\mathcal{M}_1}\rangle$ and $\mathcal{M}_2 = \langle \mathbb{Z}, +^{\mathcal{M}_2}, c^{\mathcal{M}_2}\rangle$, where $+^{\mathcal{M}_1} = +^{\mathcal{M}_2}$ is the standard addition on integers, and $c^{\mathcal{M}_1} = 0$ and $c^{\mathcal{M}_2} = 1$. Then $\mathcal{M}_1$ and $\mathcal{M}_2$ are not isomorphic, since they are not elementarily equivalent, $\mathcal{M}_1 \models c + c = c$ while $\mathcal{M}_2 \not\models c + c = c$.

However, they are equivalent in representational capacities. Assume that $\Lambda \xrightarrow{\delta} \mathcal{M}_1$ is an arbitrary representation, we can define an alternative representational context $\delta' = \delta[c \mapsto c - 1]$ (i.e., we substitute every occurence of c appeared in the definitions in δ with $c - 1$), then it is easy to see that $\mathcal{M}_2^\delta|_\Lambda \cong \mathcal{M}_1^{\delta'}|_\Lambda$, so $\mathcal{M}_1^\delta$ and $\mathcal{M}_2^{\delta'}$ are equivalent. □

Now one may argue that "isomorphism" is too strict to serve as the criterion for structural equivalence, and one may ask whether some weakened criterion of structural equivalence will coincide with the equivalence of representational capacities. However, there are actually theories with equivalent representational capacities but are not even categorically equivalent (not to mention definitional equivalent, or Morita equivalent). And thus the equivalence of representational capacities seems to be drastically different from criterion of structural equivalence.

Theorem 3.30. *There are theories T_1 and T_2 which have equivalent representational capacities but are not categorically equivalent (and hence not definitionally equivalent, nor Morita equivalent).*

Proof. Proposition 2.48 gives us an example where two theories are iso-congruent but not categorically equivalent. Since iso-congruence implies mutual interpretability which is equivalent to the equivalence of representational capacities by Theorem 3.20, we have two theories with equivalent representational capacities that are not categorically equivalent. □

3.3 Conclusion

We have seen three types of criteria of theoretical equivalence in chapter 2, the common-extension criteria, the coordinated-interpretation criteria and the categorical criteria. In this chapter, we examine such formal criteria from the perspective of representation based on the idea of representation as definition. We showed that the common-extension criteria can be seen as capturing the notion of "fully equivalent in representational contents". In other words, if we include whatever definable in a theory also as part of its representational contents, then two theories are equivalent in contents if and only if they have a common definitional extension.

The coordinated-interpretation criteria, on the other hand, do not directly correspond to representational equivalence in contents. But they can be related to representation in other ways. Firstly, as shown by Theorem 3.20, mutual interpretability can be seen as capturing the notion of "equivalent in representational capacities". Secondly, as seen in Theorem 3.22, the hierarchy of the coordinated-interpretation criteria can be seen as capturing the hierarchy of *intersubjective* full equivalence (in contents). Both are significant notions about representation that are worth studying.

However, there does not seem to be a direct connection between the categorical criteria and representation. In particular, it is not clear how the category of models of a theory plays a role in representation. In practice, physicists never use the categorical object-arrow structure to represent a physical domain, and the internal structures of models, which are taken to be the objects of the category, almost always matter. Thus, until such a connection can be established, the categorical criteria remain only a group of purely formal equivalence relationships, which do not have a valid influence on our judgment about whether two theories say the same thing about the world.

Chapter 4

Pluralism about Theoretical Equivalence

This chapter serves as a primitive attempt to extract further philosophical implications from previous analysis, particularly on its consequences for the discussion of theoretical equivalence. Due to space constraints, the discussion will be relatively sketchy, and details will be left for future work.

Let us call a theory in which every symbol is assumed to be included in the factual signature a *full theory*. In this section, I will focus on the question of theoretical equivalence between full theories, as is usually assumed in the literature.[1] And I adopt the standard gloss that two theories are theoretically equivalent just in case they say the same thing about the world. Thus, theoretical equivalence for full theories is understood as full equivalence in representational contents, which I shall simply refer to as *representational equivalence* in this section.

Interpreted in this way, the standard narrative in the literature suggests that we should find a way to determine *a uniquely correct* candidate for representational equivalence, by reasoning about the general principles as well as by looking at our intuitions about representational equivalence in specific cases, e.g., different formulations of classical electromagnetism, Hamiltonian v.s. Lagrangian mechanics, etc. I shall call this view *monism about representational equivalence.*

I shall argue that the formal analysis given in previous chapters presents a challenge to this monist view. In particular, as we have seen in Chapter 3, for any two full theories T_1 and T_2, the question whether they are representationally equivalent reduces to the question whether they have a common definitional extension. Thus, monism about representational equivalence reduces to monism about definability: there is a uniquely correct criterion of definability, which characterizes what kinds of structures are free lunches for a theory, and this criterion is the same for all theories. But this seems unlikely. In particular, we have seen a continuum of different notions of definitional extension, including not only the standard definitional extension and the standard Morita extension, but also all variants in between, e.g., Morita^s, Morita^c, Morita^p, Morita^q extensions, and arbitrary combinations of them. And, as one can imagine, there are even more notions of definitional extension that are stronger than the standard Morita extension, e.g., definitional extensions that allow doing products or coproducts of some transfinite size λ, or doing powerset over old domains. Indeed, by taking more and more set-theoretically definable operations as "free lunches", one gains stronger and stronger notions of definitional extension, and finally we would reach a stage where everything set-theoretically definable is taken to be definable. Choosing a particular notion of definability in the middle of this continuum as *the uniquely correct* notion of definability would simply be *arbitrary.*

One might object that, e.g., the standard Morita extension is not an arbitrary choice. In particular, it corresponds to *implicit* definability in many-sorted first-order logic. (See Appendix C.) Therefore, there is a good reason to allow precisely subsorts, product sorts, coproduct sorts and quotient sorts

[1] Most literature about theoretical equivalence does not make the distinction between full and non-full theories. (Though see Dewar (2015) for a discussion.) But the assumption that we focus on full theories should be clear once the distinction is made. Particularly, when people talk about theoretical equivalence between two theories T_1 and T_2, we would usually require that T_1 should be able to "express" every detail of T_2 and vice versa. This would be overkill if either T_1 or T_2 is not a full theory.

in definitional extensions. However, the notion of implicit definability is relative to the semantics of a logic, and there is *a continuum of logics* as well. For instance, powerset is not implicitly definable in many-sorted first-order logic, but *is* implicitly definable in many-sorted *higher-order* logic. Similarly, infinite products or coproducts are not implicitly definable in many-sorted first-order logic, but they *are* implicitly definable in *infinite* many-sorted first-order logic. The more powerful the logic is, the more implicitly definable structures we will have. And if one is willing to go so far as to allow the entire set theory as one's "logic", then one will have all set-theoretically definable structures as implicitly definable. Again, choosing in the middle of this continuum of implicit definability as *the uniquely correct* notion of definability would simply be *arbitrary.* It is particularly the case given the notions of definability resulting from canonical choices of logic do not seem to work well. In particular, first-order logic is known to be too weak to deal with physical theories in practice.

Now one may try to pick the extremes of the continuum to avoid the accusation of arbitrariness. But I shall argue that this strategy fails as well. Presumably, the weakest notion of definability says that we can only define new symbols when they are taken as a copy of an old symbol. This will correspond to the meta-metaphysical view that symbols of theories are supposed to, not only represent, but also represent in a way that *carve the nature at its joints* (Lewis, 1983; Sider, 2011, 2020). Take Goodman's new riddle of induction (1955, 59-83) as an example. Let T_1 be a theory about the colours of emeralds using standard terms for colours such as "green" and "blue". Let T_2 be a theory which uses the terms "grue" and "bleen" instead, where "grue" means "green before 2025 and blue after 2025" and "bleen" means "blue before 2025 and green after 2025". Despite the fact that T_1 and T_2 are, say, standard definitionally equivalent, etc., one may think that they are not representationally equivalent, as the terms in T_2 are not "natural", "fundamental", or do not carve the nature at its joint. Thus, the primitive symbols we use in a theory matter significantly.

If we opt for this weakest notion of definability, two theories are representationally equivalent iff they are notational variants of each other. But this faces a problem. Consider

- $T_1 := \{\forall x \forall y (Rxy \leftrightarrow (x = a_1 \wedge y = a_2))\}$;
- $T_2 := \{\forall x \forall y (Sxy \leftrightarrow (x = a_2 \wedge y = a_1))\}$.

Essentially, both theories say that there are two things a_1 and a_2, and an asymmetric relation holding (only) between the two. However, they differ as to whether this relation places a_1 or a_1 in the first place. Choosing one of the two as the fundamental metaphysical picture of the world seems to make a metaphysical judgement on a purely arbitrary ground.

A further suggestion by Dewar (2019a) is to say that both R and S above are fundamental.[2] However, in this view, we will have that necessarily $\forall x \forall y (Rxy \leftrightarrow Syx)$. This necessary fact is an unpleasant result since it is a brute necessity: R and S are independent components of the world, but they are now dictated as the reverse of each other. Many metaphysicians believe that such brute necessity is to be objected since this seems to be an arbirary metaphysical law without any justification.[3]

Now we turn to the other direction of the continuum. At first glance, it is not clear whether there is even an end in that direction. Perhaps one can have more and more powerful notions of definitional

[2] That is, the strategy of "sophistication" as oppose to the strategy of "simplification".

[3] For similar arguments based on objections to brute necessities, see, for instance, Cameron (2008), Dorr (2008), Kleinschmidt (2015).

extension by devising more and more powerful set theory. But there is no need to go so far as we already have a problem when we reach, say, the implicit definability for higher-order logic. The problem is that natural numbers, real numbers, and mathematical structures defined over them are all implicitly definable in higher-order logic, and thus will be counted as representationally equivalent. For instance, the two dimensional Euclidean space $\mathbb{E}^2$ and the three dimensional Euclidean space $\mathbb{E}^3$ will be counted as representationally equivalent, as they can be put into a common definitional extension where we have both $\mathbb{E}^2$ and $\mathbb{E}^3$.[4] This is possible since both $\mathbb{E}^2$ and $\mathbb{E}^3$ are definable in higher-order logic, and thus, trivially, $\mathbb{E}^2$ can define $\mathbb{E}^3$ in its definitional extension and vice versa. If we go further to set theory, then any two set-theoretically definable structures will be counted as representationally equivalent. It is arguably a criterion of representational equivalence that is too crude.

We have presented the problems for the monist view of representational equivalence: there is a continuum of different notions of definability, and there is no good reason to prefer any one of them. The current analysis seems to suggest another possible way out: it may be the case that there is simply no one uniquely correct criterion for representational equivalence, but rather *multiple* criteria of representational equivalence that are *equally correct*.

Note that this form of pluralism is distinct from many previous pluralist proposals raised in the literature of philosophy of science. Some have been pluralists about scientific methodologies or approaches, which, they argue, inevitably depend on one's goals, values, or perspectives;[5] Some have been pluralists about scientific theories or models, since they believe, for instance, that the world is too complicated to be fully represented by a single model or theory.[6] Here, however, we are not concerned with scientific methodologies or more generally the question how science as a type of social practice should be conducted; nor do we make any assumption that theories can only represent a part of the world. That is to say, the plurality of representational equivalence here does not come from the plurality of subjective values, goals, and perspectives, nor from the partisanhood of scientific representation. Rather, it is resulted from the plurality of views about definability, or in other words, what kinds of structures are free lunches and add nothing over and above the world. Thus, the plurality of representational equivalence is best explained by the plurality of conceptions about *the world* or *reality*. Here is one way the story might go: There are different conceptions of reality, all of which are *equally correct*;[7] according to some, certain structures, say, the product or coproduct of two old domains, are "free lunches" and really add nothing over and above the content of the original theory, since reality does not distinguish whether domains are accompanied by their products or coproducts; but according to other conceptions of reality, products or coproducts are not free lunches, and it is a *factual* question to ask whether there is, say, in extra to domain d_1 and d_2, a product domain $d_1 \times d_2$. Thus, some conceptions of reality will result in more fine-grained criteria of representational equivalence, allowing us to store more information and details, while other conceptions result in more coarse-grained criteria, allowing us to adopt a more abstract perspective. Still, they are *equally correct*. We may call this view *reality-pluralism*.

To illustrate, let T_1 be the formulation of Newtonian spacetime where we have a three-dimensional space and a one-dimentional time, and let T_2 be the formulation of Newtonian spacetime where we

[4] Indeed, by the same reasoning, $\mathbb{E}^2$ will be representationally equivalent to the empty domain.

[5] See Feyerabend (1975), Longino (1987, 1990).

[6] See Kellert, Longino and Waters (2006), Winther (2020).

[7] Of course, one does not need to commit to that every potential interpretation of reality is correct. Some minimal restrictions, such as consistency or empirical adequacy, may be imposed.

have a four-dimensional spacetime. Clearly, there is a sense in which they are equivalent, i.e., that the four-dimensional spacetime can simply be constructed as the product of the three-dimensional space and the one-dimensional time. This sense of equivalence may be captured by the coarse-grained conception of reality corresponding to the standard Morita equivalence, where products are taken as free lunches. But there is also a sense in which they are *not* equivalent. For instance, one may think that T_1 regards space and time as two distinct entities while T_2 unifies them into one, and thus T_1 and T_2 are not equivalent if one adopts a more fine-grained view of reality. Now, reality-pluralism holds that there is really no uniquely correct sense of equivalence, or uniquely correct conception of reality. Both views presented above are *equally correct*, each capturing a perspective of the same and the difference between T_1 and T_2.

This form of realism is particularly backed up by the heterogeneous nature of scientific ontologies and practices of taxonomies of natural kinds. In particular, many people have argued that biological practices suggest that there is really no one objective biological taxonomy, but multiple taxonomies that are equally good,and the choice of which depends on one's research interests and perspectives. "The reality of biological research practices does not seem to support the idea of convergence towards one absolute scientific conception of the biological world." (Ludwig and Ruphy, 2024) Similar examples are abundant in other specific sciences as well, such as astrophysics (Ruphy, 2010), chemistry (Chang, 2012), genetics (Griffiths and Stotz, 2013), etc. The general pluralism about the world thus coheres well with the scientific practices of specific sciences.[8]

Note that reality-pluralism naturally leads to a pluralist answer to the question "What are theories?" as well. Recall that the syntactic notion of scientific theories holds that theories are syntactic objects such as sentences, and supporters of semantic notions of theories hold that theories are semantic models. An important argument against the semantic notion is that the semantic notion of theories fails to provide a sensible criterion of theoretical equivalence. (Halvorson, 2012) And an important argument against the syntactic notion is that scientific theories cannot always be axiomatized. (Suppe, 1977) Now reality-pluralism suggests that there is no uniquely correct conception of reality, and no uniquely correct criterion of theoretical equivalence. Thus, the semantic notion is saved because there is no longer a burden to offer an overarching criterion of theoretical equivalence, and it suffices to hold that the semantic criterion of theoretical equivalence captures *a* conception of reality among the many, say, a visualizable one. Similarly, the syntactic notion is saved because there is no longer a burden to axiomatize every theory, and it suffices to say that axiomatization is necessary for *a* conception of reality— say, the conception where the most fundamental part of the world consists of natural laws to be written down as axioms — and not every scientific theory is meant to capture this conception of reality.[9] Thus, both the syntactic and semantic notions of theories are equally correct, each capturing a legitimate conception of reality.

There is no space to present a thorough defense of reality-pluralism here, but it suffices to conclude that it is a potentially appealing view that avoids the problem of monism indicated above, and coheres well with scientific practices.

[8] Note that reality-pluralism is more general than the pluralism about these specific domains: one may acknowledge the pluralities of specific domains but insist that all these different conceptions of the biological or chemical world are reducible to a fundamental domain, say, the domain of fundamental physics, about which one holds a monist view. This view is compatible with pluralism about specific domains, but not with reality-pluralism. In particular, it is still subject to the problem of monism about representational equivalence presented above.

[9] For instance, as Beatty (1981) argues biological theories may not admit this conception of reality, as biological laws are less central in biology than physical laws are in physics, and such laws are neither universal nor necessary.

Chapter 5

Restricted Set-Theoretical Languages

5.1 Introduction

In Chapter 3, we have developed a framework for representation based on the idea of representation as definition. However, physicists often use bare set-theoretical structures in practice, without assigning explicit object languages to them. However, definitions are only possible within a specific language. Thus, to make sure that the framework is applicable, we develop the formalism called restricted set-theoretical languages below, which allows us to schematically assign a canonical language to any set-theoretical structure.

5.2 Restricted Set-Theoretic Theories/Models

Different representational tools are limited in different ways. For instance, first-order logic is usually thought to be not expressive enough to formulate many physical theories, while set theory is usually thought to be too expressive and hence contains many details that are physically irrelevant. In practice, physicists often simply use set-theoretical structures to represent certain physical domains. But unlike first-order theories, there may not be an explicit axiomatization of such set-theoretical structures, and unlike full first-order set theory, certain details of set-theoretic structures are ignored, sometimes by manually specifying a set of symmetries. Here, I suggest that the way of representation adopted by physicists in practice is best reformulated using restricted set-theoretic languages. The basic idea is that physicists simply take set-theoretic structures as their representational tools, but are free to specify signatures to pick out only certain parts of these structures for representation, and exclude set-theoretic details that are irrelevant to physical representation.

The signature Σ of such restricted set-theoretic languages includes eight essential elements $\Sigma_{\mathcal{B}}$, $\Sigma_{\mathcal{D}}$, $\Sigma_{\mathcal{C}}$, $\Sigma_{\mathcal{R}}$, Σ_{Q}, Def_{Σ}. Intuitively, $\Sigma_{\mathcal{B}}$ is the set of basic domains; $\Sigma_{\mathcal{D}}$ is the set of derivative domains that are constructed from basic domains; $\Sigma_{\mathcal{C}}$ is the set of constant symbols; $\Sigma_{\mathcal{R}}$ is a set of relational symbols that specifies the part of set-theoretical structures to which we want to restrict; Σ_{Q} gives domains that we want to quantify over; and Def_{Σ} is a function which gives the symbols mentioned above their identification information.

Definition 5.1. A *signature* Σ is a tuple $\langle \Sigma_{\mathcal{B}}, \Sigma_{\mathcal{D}}, \Sigma_{\mathcal{C}}, \Sigma_{\mathcal{R}}, \Sigma_{Q}, \mathrm{Def}_{\Sigma} \rangle$ where:

- $\Sigma_{\mathcal{B}}$ is the set of basic domain symbols $b_1, b_2, ...$
- $\Sigma_{\mathcal{D}}$ is the set of derivative domain symbols $d_1, d_2, ...$

We call an element $\theta \in \Sigma_{\mathcal{B}} \cup \Sigma_{\mathcal{D}}$ a domain symbol of Σ, and let $\Sigma_{\mathcal{A}} = \Sigma_{\mathcal{B}} \cup \Sigma_{\mathcal{D}}$ be the set of all domain symbols.

- Σ_{C} is a set of constant symbols: $c_1, c_2, ...$

- Σ_R is a set of relational symbols: $R_1, R_2, ...$
- Σ_Q is a set of domain symbols;
- Def_Σ is a function which
  - assigns to each derivative domain symbol d_j a first-order set-theoretic formula $\phi(v, b_{i_1}, ..., b_{i_n})$ with basic domain symbols as parameters , which specifies how the derivative domain d_j is defined from basic domains;[1]
  - assigns to each constant symbol c_i a domain symbol specifying the domain it belongs to;
  - assigns to each relational symbol R_i a first order set-theoretical formula $\phi(v^{\theta_1}, ..., v^{\theta_j})$ with variables indexed with domain symbols.

  When the context is clear, we write Def for Def_Σ.

The language $\mathcal{L}_\Sigma$ with respect to Σ is defined naturally as follows:

Definition 5.2. $\mathcal{L}_\Sigma$:

- The symbols include: $\neg, \wedge, =, \in, \forall$, and for all domain symbols $\theta \in \Sigma_Q$, countably infinite variables $v_0^\theta, v_1^\theta, ...$
- The formulas can be defined recursively as:
  - Atomic formulas:$R_i x_1, ..., x_n$, where $x_1, ..., x_n$ are either variables or constants with the same corresponding domains as free variables in $\text{Def}(R_i)$.
  - If ϕ is a formula, then $\neg\phi$ is a formula.
  - If ϕ and ψ are formulas, then $\phi \wedge \psi$ is a formula.
  - If ϕ is a formula and v_i^θ is a variable, then $\forall v_i^\theta \phi$ is a formula.

The semantics of $\mathcal{L}_\Sigma$ is given in a routine manner.

Definition 5.3. A *model* $\mathcal{M}$ of a language $\mathcal{L}_\Sigma$ is a quadruple $\langle \mathcal{B}, \mathcal{D}, \mathcal{C}, \cdot^\mathcal{M} \rangle$, where $\cdot^\mathcal{M}$ is a function defined as the union of:

- a bijection from $\Sigma_\mathcal{B}$ to $\mathcal{B}$, which assigns each basic domain symbol $b_i \in \Sigma_\mathcal{B}$, a basic domain $B_i \in \mathcal{B}$.
- a bijection from $\Sigma_\mathcal{D}$ to $\mathcal{D}$, which assigns each derivative domain symbol $d_i \in \Sigma_\mathcal{D}$, a set $D_i \in \mathcal{D}$ such that $x \in D_i$ iff $\phi_{d_i}(x, (b_{i_1})^\mathcal{M}, ..., (b_{i_n})^\mathcal{M})$ holds.
- a function which assigns each constant symbol $c_i \in \Sigma_C$, an element $C_i \in \mathcal{C}$ such that $C_i \in \text{Def}(c_i)^\mathcal{M}$.

Definition 5.4. Let $\mathcal{M}$ be a $\mathcal{L}_\Sigma$-model. The *semantical consequence* is defined as follows:

- s is a variable assignment if it is a function whose domain is the set of variables in $\mathcal{L}_\Sigma$, and, $s(v_i^\theta) \in \theta^\mathcal{M}$.

[1] Here we require that we can prove in our background ZFC $\text{Def}_\Sigma(d_j)$ indeed defines a set, i.e., there is a set including all x such that $\text{Def}(d_j)(x, b_{i_1}, ..., b_{i_n})$ holds, so that we can apply comprehension.

- The denotation of a term t in $\mathcal{M}$ under s (in notation $t_s^{\mathcal{M}}$) is defined as:
  - If t is a variable v_i^θ, then $t_s^{\mathcal{M}} = s(v_i^\theta)$.
  - If t is a constant c_i, then $t_s^{\mathcal{M}} = c_i^{\mathcal{M}}$.
- $\mathcal{M}, s \models \phi$ is defined recursively as:
  - $\mathcal{M}, s \models R_i x_1, ..., x_N$ iff $\text{Def}(R_i)((x_1)_s^{\mathcal{M}}, ..., (x_n)_s^{\mathcal{M}})$ holds.
  - $\mathcal{M}, s \models \neg\phi$ iff $\mathcal{M}, s \not\models \phi$.
  - $\mathcal{M}, s \models \phi \wedge \psi$ iff $\mathcal{M}, s \models \phi$ and $\mathcal{M}, s \models \psi$.
  - $\mathcal{M}, s \models \forall v_i^\theta \phi$ iff for all $a \in \theta^{\mathcal{M}}$, $\mathcal{M}, s[v_i^\theta \mapsto a] \models \phi$.
- And $\mathcal{M} \models \phi$ if for all variable assignments s, $\mathcal{M}, s \models \phi$.

From the above definitions, it should be clear that the essential non-logical symbols in a signature Σ are really constants in $\Sigma_{\mathcal{C}}$ and the basic domain symbols in $\Sigma_{\mathcal{B}}$: once the interpretation of these symbols is determined, the interpretation of all other symbols in Σ is determined as well. In particular, relational symbols in $\Sigma_{\mathcal{R}}$ are not really relational symbols in the usual sense, say, relational symbols in first-order logic: the interpretation of a relational symbol R here is fully fixed by $\text{Def}(R)$. In this sense, relational symbols here are more similar to "logical" rather than "non-logical" symbols.

The isomorphism of models is defined naturally:

Definition 5.5. Two $\mathcal{L}_\Sigma$-models $\mathcal{M}_1 = \langle \mathcal{B}_1, \mathcal{D}_1, \mathcal{C}_1, \cdot^{\mathcal{M}_1} \rangle$ and $\mathcal{M}_2 = \langle \mathcal{B}_2, \mathcal{D}_2, \mathcal{C}_2, \cdot^{\mathcal{M}_2} \rangle$ are *isomorphic* iff for any $\theta \in \Sigma_A$, there is a surjection[2] $f_\theta : \theta^{\mathcal{M}_1} \to \theta^{\mathcal{M}_2}$, such that for $f = \bigcup_{\theta \in \Sigma_A} f_\theta$, we have:

- for any R_i, for any sequence of elements $a_1, ..., a_n$ belonging to corresponding domains, we have $\text{Def}(R_i)(a_1, ..., a_n)$ iff $\text{Def}(R_i)(f(a_1), ..., f(a_n))$;
- for any $c_i \in \Sigma_C$, $f(c_i^{\mathcal{M}_1}) = c_i^{\mathcal{M}_2}$.

Theorem 5.6. *For any two isomorphic $\mathcal{L}_\Sigma$-models, $\mathcal{M}_1$ and $\mathcal{M}_2$, for any $\mathcal{L}_\Sigma$-sentence ϕ, $\mathcal{M}_1 \models \phi$ iff $\mathcal{M}_2 \models \phi$.*

Proof. Let f be the isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$. Let s be an assignment of variables in $\mathcal{M}_1$. Let $f(s)$ be the corresponding assignment of variables in $\mathcal{M}_2$ defined by $f(s)(v_i^\theta) = f(s(v_i^\theta))$. We prove the following stronger claim: for any assignment s, $\mathcal{M}_1, s \models \phi$ iff $\mathcal{M}_2, f(s) \models \phi$.

We proceed by induction.

- for atomic formulas $\phi := R(x_1, ..., x_n)$, we have $\mathcal{M}_1, s \models R(x_1, ..., x_n)$ iff $\text{Def}(R_i)((x_1)_s^{\mathcal{M}_1}, ..., (x_n)_s^{\mathcal{M}_1})$ holds iff $\text{Def}(R_i)(f((x_1)_s^{\mathcal{M}_1}), ..., f((x_n)_s^{\mathcal{M}_1}))$ holds iff $\text{Def}(R_i)((x_1)_{f(s)}^{\mathcal{M}_2}, ..., (x_n)_{f(s)}^{\mathcal{M}_2})$ holds iff $\mathcal{M}_2, f(s) \models R(x_1, ..., x_n)$.
- The inductive cases for $\wedge, \neg$ are trivial.
- The inductive case: $\phi := \forall v_i^\theta \psi$. $\mathcal{M}_1, s \models \forall v_i^\theta \psi$ iff for all $a \in \theta^{\mathcal{M}_1}$, $\mathcal{M}_1, s[v_i^\theta \mapsto a] \models \psi$ iff for all $a \in \theta^{\mathcal{M}_1}$, $\mathcal{M}_2, f(s)[v_i^\theta \mapsto f(a)] \models \psi$ iff for all $b \in \theta^{\mathcal{M}_2}$, $\mathcal{M}_2, f(s)[v_i^\theta \mapsto b] \models \psi$ iff $\mathcal{M}_2, f(s) \models \forall v_i^\theta \psi$.

This completes the induction.

The theorem follows when we take ϕ to be closed sentences. □

[2] We do not require that f is necessarily a bijection, but as long as we include identity of each domain in Σ_R, f must be a bijection.

5.3 Comparison with First-Order Languages

We can systematically translate first-order languages into some restricted set theoretical languages $\mathcal{L}_\Sigma$, and first-order models into some $\mathcal{L}_\Sigma$ models, while preserving the semantic consequence relationship and isomorphism between models. This justifies our claim that representation done using first-order theories/models can be equivalently transformed into representation done using restricted set-theoretic theories/models.

Without loss of generality, we only consider the translation of a first-order theory with a single binary relational symbol R below. Consider a first-order language $\mathcal{L}_{1\Gamma}$ with the signature $\Gamma = \{R\}$. We define Σ as follows:

- $\Sigma_\mathcal{B} = \{b\}$; $\Sigma_\mathcal{D} = \{r\}$; $\Sigma_\mathcal{C} = \{R\}$; $\Sigma_\mathcal{R} = \{E, B\}$; $\Sigma_Q = \{b\}$
- Def_Σ is defined as:
  - Def(r) is defined as the set-theoretical formula $\phi_r(v, b)$ which says that v is a binary relation on b.[3]
  - Def(R) $:= r$.
  - Def(E) $:= v_1^b = v_2^b$.
  - Def(B) $:= (v_1^b, v_2^b) \in v_1^r$.

The first-order formulas are translated into $\mathcal{L}_\Sigma$ by a function $\cdot^*$ as follows:

- For $\phi := v_i = v_j$, we define $\phi^* := v_i^b = v_j^b$.
- For $\phi := R(v_i, v_j)$, we define $\phi^* := B(v_i^b, v_j^b, R)$.
- For $\phi := \neg\psi$, we define $\phi^* := \neg\psi^*$.
- For $\phi := \psi \wedge \chi$, we define $\phi^* := \psi^* \wedge \chi^*$.
- For $\phi := \forall v_i \psi$, we define $\phi^* := \forall v_i^b \psi^*$.

For any first-order model $\mathbb{M} = \langle X, \mathcal{R}\rangle$ where $\mathcal{R}$ is a binary relation on X, we can define a $\mathcal{L}_\Sigma$ model $\mathbb{M}^* = \langle \mathcal{B}, \mathcal{D}, \mathcal{C}, \cdot^{\mathbb{M}^*}\rangle$ as follows:

- $\mathcal{B} = \{X\}$; $\mathcal{D} = \{\mathbf{R}\}$ where $\mathbf{R}$ is the set of all binary relations on X; $\mathcal{C} = \{\mathcal{R}\}$;
- $\cdot^{\mathbb{M}^*}$ is defined as:
  - $b^{\mathbb{M}^*} = \mathbf{X}$, $r^{\mathbb{M}^*} = \mathbf{R}$, $R^{\mathbb{M}^*} = \mathcal{R}$.

The variable assignment s in $\mathbb{M}$ is translated into s^* in $\mathbb{M}^*$ as, for any variable v_i^b in $\mathcal{L}_\Sigma$, $s^*(v_i^b) = s(v_i)$.

Now we have the following theorems:

[3] More explicitly,

∗ $\phi_u := \exists x \exists y (x \in v \wedge y \in v \wedge x \in b \wedge y \in b \wedge \forall z (z \in v \rightarrow (z = x \vee z = y)))$ which defines the set of unordered pairs of b.

∗ $\phi_p := \exists x \exists y (x \in v \wedge y \in v \wedge x \in b \wedge \phi_u(y) \wedge x \in y \wedge \forall z \in v (z = x \vee x = y))$ which defines the set of ordered pairs of b.

∗ $\phi_r := \forall x \in v (\phi_p(x))$ which defines the set of all relations on b.

Theorem 5.7. *For any first-order formula ϕ of $\mathcal{L}_{1\Gamma}$, $\mathbb{M}, s \models \phi$ iff $\mathbb{M}^*, s^* \models \phi^*$.*

Proof. We prove by induction:

- $\phi = v_i = v_j$: $\mathbb{M}, s \models v_i = v_j$ iff $s(v_i) = s(v_j)$ iff $s^*(v_i^b) = s^*(v_j^b)$ iff $\mathbb{M}^*, s^* \models v_i^b = v_j^b$ iff $\mathbb{M}^*, s^* \models (v_i = v_j)^*$.
- $\phi = R(v_i, v_j)$: $\mathbb{M}, s \models R(v_i, v_j)$ iff $(s(v_i), s(v_j)) \in \mathcal{R}$ iff $(s^*(v_i^b), s^*(v_j^b)) \in \mathcal{R}$ iff $(s^*(v_i^b), s^*(v_j^b)) \in R^{\mathbb{M}^*}$ iff $\mathbb{M}^*, s^* \models R^*(v_i^b, v_j^b)$.
- $\phi = \neg\psi$: $\mathbb{M}, s \models \neg\psi$ iff $\mathbb{M}, s \not\models \psi$ iff $\mathbb{M}^*, s^* \not\models \psi^*$ iff $\mathbb{M}^*, s^* \models \neg\psi^*$ iff $\mathbb{M}^*, s^* \models (\neg\psi)^*$.
- $\phi = \psi \wedge \chi$: $\mathbb{M}, s \models \psi \wedge \chi$ iff $\mathbb{M}, s \models \psi$ and $\mathbb{M}, s \models \chi$ iff $\mathbb{M}^*, s^* \models \psi^*$ and $\mathbb{M}^*, s^* \models \chi^*$ iff $\mathbb{M}^*, s^* \models \psi^* \wedge \chi^*$ iff $\mathbb{M}^*, s^* \models (\psi \wedge \chi)^*$.
- $\phi = \forall v_i \psi$: $\mathbb{M}, s \models \forall v_i^b \psi$ iff for all $a \in X$, $\mathbb{M}, s[v_i \mapsto a] \models \psi$ iff for all $a \in X$, $\mathbb{M}^*, s^*[v_i^b \mapsto a] \models \psi^*$ iff $\mathbb{M}^*, s^* \models \forall v_i^b \psi^*$ iff $\mathbb{M}^*, s^* \models (\forall v_i \psi)^*$.

□

We also have:

Theorem 5.8. *For any two first-order models $\mathbb{M}_1$ and $\mathbb{M}_2$, $\mathbb{M}_1$ and $\mathbb{M}_2$ are isomorphic iff their corresponding $\mathcal{L}_\Sigma$ models $\mathbb{M}_1^*$ and $\mathbb{M}_2^*$ are also isomorphic.*

Proof. Assume that $\mathbb{M}_1 = \langle X_1, \mathcal{R}_1 \rangle$ and $\mathbb{M}_2 = \langle X_2, \mathcal{R}_2 \rangle$ are isomorphic, and let f be the isomorphism. We define f^* from $\mathbb{M}_1^*$ to $\mathbb{M}_2^*$ as follows:

- $f_b^* = f$;
- f_r^* is induced naturally by f

Note that $f_r^*(\mathcal{R}_1) = \mathcal{R}_2$ as f is an isomorphism between $\mathbb{M}_1$ and $\mathbb{M}_2$.

It can be easily checked that f^* gives an isomorphism between $\mathbb{M}_1^*$ and $\mathbb{M}_2^*$.

On the other hand, if $\mathbb{M}_1^*$ and $\mathbb{M}_2^*$ are isomorphic, let f^* be the isomorphism. Then $f^*|_X$ will serve as an isomorphism between $\mathbb{M}_1$ and $\mathbb{M}_2$. □

Therefore, restricted set-theoretical languages can be seen as natural generalizations of first order languages. The advantage is that it has greater expressive power and is closer to the actual physical practice. But of course, it will not have completeness and thus cannot have a syntactical characterization of the (semantical) consequence relationship.

5.4 Mathematical Structures in Restricted Set Theoretical Languages

Mathematical structures are often presented as bare set-theoretical structures without an explicit object language. For instance, a topological space is often constructed directly in our set-universe as an ordered pair $\langle X, T \rangle$, where X is the set of base points and T the set of open sets satisfying the axioms of topology. Representation done using topological spaces can be reconstructed in this restricted set-theoretic setting as follows. We have a basic domain symbol θ_1 that denotes the set X and two derivative domain symbols θ_2, θ_3 that denote $\mathcal{P}(X)$ and $\mathcal{P}(\mathcal{P}(X))$, respectively. Then we restrict membership sentences to $v_i^{\theta_1} \in v_j^{\theta_2}$ and $v_i^{\theta_2} \in v_j^{\theta_3}$, and identity sentences to $v_i^{\theta_n} = v_j^{\theta_n}$, for $n = 1, 2, 3$. Thus, we exclude irrelevant set-theoretic details like whether one base point belongs to another, expressed as $v_i^{\theta_1} \in v_j^{\theta_1}$. As a sanity check, we can prove that two topological spaces are homeomorphic iff they are isomorphic when attached with such restricted set-theoretic language. Thus, we accurately capture the topologically significant information we wish to use in representation.

More explicitly, we can assign the following canonical signature to a topological space (X, T):

- $\Sigma_{\mathcal{T}}$ is defined as $\langle \Sigma_{\mathcal{B}}, \Sigma_{\mathcal{D}}, \Sigma_{\mathcal{C}}, \Sigma_{\mathcal{R}}, \Sigma_Q, \mathrm{Def}_\Sigma \rangle$ where:
  - $\Sigma_{\mathcal{B}} = \{\theta\}$; $\Sigma_{\mathcal{D}} = \{p\theta, pp\theta\}$; $\Sigma_{\mathcal{C}} = \{t\}$; $\Sigma_{\mathcal{R}} = \{R_1, R_2, R_3, R_4, R_5\}$; $\Sigma_Q = \{\theta, p\theta, pp\theta\}$;
  - Def_Σ is defined as:
    - ∗ $\mathrm{Def}(px) := v \in \mathcal{P}(\theta)$, $\mathrm{Def}(pp\theta) := v \in \mathcal{P}(\mathcal{P}(\theta))$;
    - ∗ $\mathrm{Def}(t) := pp\theta$;
    - ∗ $\mathrm{Def}(R_1) := v_1^\theta = v_2^\theta$, $\mathrm{Def}(R_2) := v_1^{p\theta} = v_2^{p\theta}$, $\mathrm{Def}(R_3) := v_1^{pp\theta} = v_2^{pp\theta}$, $\mathrm{Def}(R_4) := v_1^\theta \in v_2^{p\theta}$, $\mathrm{Def}(R_5) := v_1^{p\theta} \in v_2^{pp\theta}$
- The restricted set-theoretic language for topological spaces is then defined as $\mathcal{L}_{\Sigma_{\mathcal{T}}}$.

Topological spaces can then be transformed into $\Sigma_{\mathcal{T}}$-models by interpreting the domain θ as the set of base points, the domain $p\theta$ as its powerset, and $pp\theta$ as the powerset of its powerset. The only constant t is interpreted naturally as the set of open sets. We now prove that two topological spaces are homeomorphic iff their transformed logical models are isomorphic.

Theorem 5.9. *Let $S_1 = \langle X_1, T_1 \rangle$, $S_2 = \langle X_2, T_2 \rangle$ be two topological spaces. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be their transformed logical models with the canonical signature. Then S_1 and S_2 are homeomorphic iff $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic as logical models.*

Proof. Assume that f is the homeomorphism from S_1 to S_2. We define the isomorphism f^* from $\mathcal{M}_1$ to $\mathcal{M}_2$ as follows:

- $f^*_\theta = f$
- $f^*_{p\theta}$, $f^*_{pp\theta}$ are defined as the natural extension of f to $\mathcal{P}(X_1)$ and $\mathcal{P}(\mathcal{P}(X_1))$ respectively.

It is easy to check that f^* is an isomorphism:

- for any $x \in X_1$ and $y \in \mathcal{P}(X_1)$, we have $x \in y$ iff $f^*(x) \in f^*(y)$ by our construction of f^*.
- for any $x \in \mathcal{P}(X_1)$ and $y \in \mathcal{P}(X_2)$, we have $x = y$ iff $f^*(x) = f^*(y)$ by our construction of f^*.
- for any $x, y \in X_1$ (or $\mathcal{P}(X_1)$, $\mathcal{P}(\mathcal{P}(X_1))$), $x = y$ iff $f^*(x) = f^*(y)$ since f is a bijection.
- $f^*(T_1) = T_2$ since f is a homeomorphism.

Conversely, assume that g is an isomorphism from $\mathcal{M}_1$ to $\mathcal{M}_2$. Let $g^* := g_x$ be the restriction of g to X_1. Then g^* is a bijection from X_1 to X_2, since we include the identity in domain θ as R_1 in Σ_R. It is easy to check that g^* is a homeomorphism: for any $Y \in \mathcal{P}(X)$, $Y \in T_1$ iff $f^*(Y) \in f^*(T_1)$ iff $f^*(Y) \in T_2$ □

If we continue the procedure, taking into account the domain of $\mathbb{R}^n$ and the domain of functions constructed thereof, we eventually arrive at an object language for Lorentzian manifolds, which physicists use to represent spacetime in general relativity. We shall sketch the details of this construction below.

5.5 An Object Language for Lorentzian Manifolds

Now we sketch how to construct an object language $\mathcal{L}_M$ for Lorentzian manifolds.

First of all, we need the domain of $\mathbb{R}$ (for curves and tangent spaces) and more generally $\mathbb{R}^n$ (for charts and atlas). Both can be added to $\mathcal{L}_M$ as derivative domains, since both can be defined by a single set-theoretic formula. The definition simply follows the routine of textbooks. The set of natural numbers is defined as the set closed under predecessors and whose element is either the empty set ($x = \emptyset$) or a successor $\exists y(x = \{y\} \cup y)$. Then we can define sequences of natural numbers, by which additions and multiplications can be defined by recursion. Integers are then defined as a quotient of pairs of natural numbers by the equivalence relation $(a, b) \sim (c, d)$ iff $a + d = b + c$. Rational numbers are defined as a quotient of the product of integers and positive integers $\mathbb{Z} \times \mathbb{Z}^*$ with the equivalence relation $(a, b) \sim (c, d)$ iff $ad = bc$. The set of real numbers is then defined as the quotient of Cauchy sequences of rational numbers with the equivalence relation $(a_n) \sim (b_n)$ iff for any $\varepsilon > 0$, there is N such that for all $n \geq N$, $|a_n - b_n| < \varepsilon$. Topology over reals is generated by the open intervals (a, b), $a, b \in \mathbb{R}$.

For relational symbols on $\mathbb{R}$ and $\mathbb{R}^n$, we include all common operations in analysis. There should be no difficulties defining such operations following the standard textbooks — Addition, multiplication, or exponentiation are induced from the addition, multiplication, or exponentiation of Cauchy sequences. We may also include the function spaces of reals, which can be defined as a subdomain $\mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m)$. The limits of functions are defined by the standard ε-δ notation.

We simply include the language for topological spaces in $\mathcal{L}_M$. Let θ be the domain of topological points. The domain u of charts (U, ϕ) is then defined as a subdomain of $\mathcal{P}(\theta) \times \mathcal{P}(\theta \times R^n)$, whose elements satisfy the definition of charts, i.e., ϕ is a homeomorphism from U to $\mathbb{R}^n$, and U and the range of ϕ are both open sets. The domain a of maximal atlas is defined as a subdomain of $\mathcal{P}(u)$, whose elements are maximal sets of C^∞ compatible charts which cover the whole space θ. We add a constant symbol $atlas$ to denote the maximal atlas for our target manifold.

Finally, we add the metrical structure. The domain c of smooth curves can be defined as a subdomain of $\mathcal{P}(\mathbb{R} \times \theta)$, whose elements are smooth curves. The tangent bundle TM can be defined as a subdomain of $\theta \times \mathcal{P}(c)$ such that for each of its elements (p, C), C is an equivalent class of curves passing through p, where the equivalence relationship is defined by the sameness of derivatives. Similarly, the binary product of tangent bundles TM^2 can be defined as a subdomain of $\theta \times \mathcal{P}(c) \times \mathcal{P}(c)$.

The domain mt of (pseudo-Riemannian) metric tensor fields is then defined as a subdomain of $\mathcal{P}(TM^2 \times \mathbb{R})$, whose elements are functions from TM^2 to $\mathbb{R}$ which are smooth, bilinear, non-degenerate, and symmetric. The domain lmt can then be defined as a subdomain of mt whose elements are Lorentzian metric tensors, i.e., those whose signature is $(-, +, ..., +)$. This amounts to the requirement that for any point $p \in \theta$, there is a basis over the tangent space of p in which the metric tensor becomes a diagonal matrix with one negative entry and the rest positive entries, or, in formula:

$$\begin{aligned}\phi(y^{mt}) := \forall x^\theta \exists v_1^{pc}, ..., v_n^{pc}((x^\theta, v_1^{pc}, ..., v_n^{pc}) \in TM \wedge \bigwedge_{1\leq i\leq n} (v_i^{pc} \neq 0) \wedge \bigwedge_{1\leq i\neq j\leq n} y^{mt}(v_i^{pc}, v_j^{pc}) = 0 \\ \wedge \bigvee_{1\leq i\leq j} [y^{mt}(v_i^{pc}, v_i^{pc}) < 0 \wedge \bigwedge_{j\neq i, 1\leq j\leq n} y^{mt}(v_j^{pc}, v_j^{pc}) > 0])\end{aligned}$$

Here we assume that the dimension of the manifold is n. And for a domain symbol d, we use pd to denote the derivative domain defined as the powerset of d. To be more rigorous, $v_i^{pc} \neq 0$ above can be further unpacked as $\forall x^c \in v_i^{pc}(\exists y^{\mathbb{R}} x^c(y^{\mathbb{R}}) \neq p)$.

Once the domain lmt is defined, we can add a constant g in lmt to denote the metric field in the Lorentzian manifold.[4]

We may briefly summarize our language $\mathcal{L}_M$ as follows:

- The only basic domain is θ, i.e., the domain for the topological space.
- The derivative domains include $\mathbb{R}$, $\mathbb{R}^n$, u, a, c, TM, TM^2, mt etc.
- There are three constant symbols, t (the set of open sets), $atlas$ (the maximal atlas), and g (the metric tensor field).
- The relational symbols include all common operations in analysis and differential geometry.

It is then easy to see that isomorphisms of $\mathcal{L}_M$ correspond to isometries. In particular, as all derivative domains and relational symbols are defined in terms of the basic domain θ and constants t, $atlas$ and g, isomorphisms between two models $\mathcal{M}_1$ and $\mathcal{M}_2$ in $\mathcal{L}_M$ are simply maps between $\theta^{\mathcal{M}_1}$ and $\theta^{\mathcal{M}_2}$ which preserve the interpretation of t, $atlas$ and g. To show that this simply amounts to the requirement of being an isometry, we reason as follows: the preservation of t is equivalent by definition to the requirement that the map is continuous; the following theorem proves that the further preservation of the maximal atlas is equivalent to the requirement that the map is a diffeomorphism; and finally, it holds by definition that for a diffeomorphism to preserve the metric tensor field g is precisely for it to become an isometry.

Theorem 5.10. *Let $\mathcal{M}_1 = (M_1, \mathcal{A}_1)$ and $\mathcal{M}_2 = (M_2, \mathcal{A}_2)$ be two smooth manifolds, and f a continuous map from M_1 to M_2. Then f is a diffeomorphism iff f preserves maximal atlas, i.e., for any chart (U, ϕ) in $\mathcal{A}_1$, the map $k(U, \phi) = (f(U), \phi \circ f^{-1})$ is a bijection from $\mathcal{A}_1$ to $\mathcal{A}_2$.*

Proof. Suppose that f is a diffeomorphism. Let (U, ϕ) be an arbitrary chart in $\mathcal{A}_1$. We show that $k(U, \phi) := (f(U), \phi \circ f^{-1})$ is pairwise compatible with any arbitrary chart (V, ψ) in $\mathcal{A}_2$. So let (V, ψ) be an arbitrary chart in $\mathcal{A}_2$. If $f(U) \cap V = \emptyset$, they are vacuously compatible. If $f(U) \cap V \neq \emptyset$, we choose p such that $f(p) \in f(U) \cap V$. By assumption, f is smooth at p, therefore, there is a chart (W, χ) in $\mathcal{A}_1$ such that $p \in W$ and $\chi \circ f^{-1} \circ \psi^{-1}$ is smooth. Since (W, χ) is compatible with (U, ϕ), $\phi \circ \chi^{-1}$ is also smooth. Therefore, $\phi \circ \chi{-}1 \circ (\chi \circ f^{-1} \circ \psi^{-1})$ is smooth, and hence $\phi \circ f^{-1} \circ \psi^{-1}$ is smooth. This shows that $(f(U), \phi \circ f^{-1})$ is compatible with (V, ψ), and hence must be included in $\mathcal{A}_2$. Similarly, for any chart (W, χ) in $\mathcal{A}_2$, we can show that $h(W, \chi) := (f^{-1}(W), \chi \circ f)$ must be included in $\mathcal{A}_1$. Finally, it is easy to see that $h \circ k = id$ and $k \circ h = id$, therefore, k is a bijection from $\mathcal{A}_1$ to $\mathcal{A}_2$.

Suppose that f preserves the maximal atlas. Then for any point $p \in M_1$, any chart $(U, \phi) \in \mathcal{A}_1$ covering p, $(f(U), \phi \circ f^{-1})$ is a chart in $\mathcal{A}_2$ covering $f(p)$. And $\phi \circ f^{-1} \circ f \circ \phi^{-1} = id$ is trivially smooth. Therefore, f is smooth at p. Since p is arbitrary, f is a diffeomorphism. □

[4] To state the theory of GR, one may simply add a further constant T of domain $\mathcal{P}(TM^2 \times \mathbb{R})$ for the stree-energy tensor field. As the Einstein tensor can be defined from g, the Einstein field equation can be stated. Or, following the idea of representation as definition, one may simply define a tensor field T from g by the Einstein field equation, and let T denote the stress-energy tensor field.

The restricted set-theoretic language $\mathcal{L}_M$ is clearly not the only way by which one may give an explicit object language and consequently a theory of Lorentzian manifolds. Bradley and Weatherall (2022) identify two previous attempts: Mundy (1992) argues that the Riemannian geometry can be axiomatized in a similar way as Euclidean geometry is axiomatized presumably using higher-order logic, and Shulman (2017) suggests that the classes of Lorentzian manifolds can be defined in HoTT as a specific type whose terms are identified up to isometry. Their points are certainly valid, though they do not bother to give further sketches of their constructions.[5]

Compared to the above approaches, the main advantage of using restricted set-theoretical languages is that it is closest to the actual mathematical practices and hence is most convenient to adopt. As can be seen in the above construction, we simply follow the standard set-theoretical definitions of our targeted mathematical structures (here, Lorentzian manifolds), and include details that are specifically relevant to this type of structures. Additionally, this approach is also more flexible, especially compared to the higher-order approach, as it is applicable to *any* set-theoretical structures, while higher-order logic can be limited in its expressive power.[6]

A mathematically natural but philosophically significant feature of $\mathcal{L}_M$ is that we do not have constants for points in manifolds. The only three constants we have in $\mathcal{L}_M$ are t (the set of open sets), $atlas$ (the maximal atlas) and G (the metric tensor field). This makes $\mathcal{L}_M$ a formalism that corresponds to a variety of what philosophers often call "sophisticated substantivalism" (Belot and Earman, 2001), supported by Maudlin (1988), Butterfield (1989), Stachel (1993, 2002), Rynasiewicz (1994), Hoefer (1996), Pooley (2006), Russell (2014). See Gomes and Butterfield (2023a, 2023b), Cudek (2024), Jacobs (2024) for more recent defenses. Roughly, sophisticated substantivalists say that while spacetime points exist, there cannot be two distinct possible worlds that differ only regarding which spacetime point is which. $\mathcal{L}_M$, therefore, is a specific form of sophisticated substantivalism, that simply rejects that there is any matter of fact about the specific identities of spacetime points (i.e., there are no "haecceitistic facts"), as we simply have no constants for spacetime points[7] Quantificational facts (over spacetime points) are all matters of facts there are about spacetime.

Some believe that sophisticated substantivalism, of one form or another, will need a reformulation of GR. The formalization above suggests, instead, that if one assumes that the isomorphism criterion for the theory of GR is isometry, then the theory of GR must contain no constants for spacetime points, and hence, it must be some form of sophisticated substantivalism. The following two chapters give a more detailed discussion of this point.

Another remark to be added concerns the relationship between $\mathcal{L}_M$ and the well-known method of Ramseyfication (Ramsey, 1929; Carnap, 1966; Lewis, 1966, 1970, 1972). In the context of Ramseyfication, the vocabulary of our language is divided into two parts: the "problematic" part and the "benign" part (Frigg and Votsis, 2011; Dewar, 2019b). Let our theory be $T(, t_1, ..., t_n)$ where $t_1, ..., t_n$ are symbols that belong to the problematic part, which could be first-order terms, relational symbols,

[5] Mundy does prove what he calls the "representation theorem", which essentially says that for any Riemannian manifold $\mathcal{M}$, we can assign a first-order structure $K(\mathcal{M})$ such that K preserves isomorphism classes. It is far from clear how this suffices for a proof of *axiomatizability*, as the class $\{K(\mathcal{M})\}$ is certainly not elementary in first-order logic. Mundy further mentions the fact that the class of Riemannian manifolds is axiomatizable in higher-order logic ("Since **Riem** is definable in higher-order logic..." (p.518)). But no further details are given for this axiomatization.

[6] For instance, higher-order logic presumably cannot deal with set-theoretical structures of ranks higher than $V_{\omega+\omega}$, as Z fails to prove the existence of $\omega + \omega$, but suffices to give a semantics for higher order logic.

[7] Not all varieties of sophisticated substantivalism can be represented by $\mathcal{L}_M$. For instance, certain views of metrical essentialism (Maudlin, 1988) will concede hecceitistic facts but insist that it is metaphysically necessary that two distinct worlds cannot differ merely hecceitisitically.

or symbols in higher-order. Assuming T is finite, the Ramseyfication of T, T^R, is then defined as $\exists x_1, ..., x_n T(x_1, ..., x_n)$. It can be proved that T^R will have the same consequences as T in the benign part of our language, but it is better than T in the sense that it does not contain any problematic symbols. In the current case of GR in $\mathcal{L}_M$, one may take the "problematic" symbols to be constants for spacetime points, and take our final theory as the result of Ramseyfication of a fuller version of GR which does contain constants for spacetime points.

But despite formal similarity, it is worth noting that the motivation behind $\mathcal{L}_M$ is very different from that behind the usual cases of Ramseyfication.

In usual cases of Ramseyfication, the problematic symbols are *by themselves* regarded as problematic. For Ramsey, the problematic symbols consist of non-observational terms, which are also called "secondary terms" or "theoretical terms". Ramsey holds a verificationist view of meaning,[8] which says that a sentence is meaningful iff it is verifiable. Thus, he believes that such non-observational parts of our language are simply meaningless.[9] Carnap (1966) no longer holds such a naive verificationist view of meaning at the time of his writing, but he still believes that only the observable part of our language has cognitive contents or synthetic contents. Thus, Ramseyfication is able to extract the purely synthetical content of a theory. Lewis (1972) applies the method of Ramseyfication in the context of philosophy of mind. For him, the problematic part of our language is no longer non-observable terms, but rather terms that denote mental states and processes. As he supports a materialist view of mind, and in particular the type-type mind-brain identity thesis, it is critical for him to reduce those mental terms to purely physical terms. The way he manages it is by Ramseyfication, in particular, via a functionalist characterization of mental states. Thus, suppose $T(pain, t_1, ..., t_n)$ is our psychological theory of pain, where $t_1, ..., t_n$ are other mental terms, then Lewis offers a definition of pain as $Df(pain) := \iota x \exists x_1, ..., x_n T(x, x_1, ..., x_n)$[10] For humankind, this will pick out a unique neural state that realizes the causal role of pain as specified by T. Of course, $Df(pain)$ is non-rigid, and thus it picks out different neural states for different species of creatures. But the mind-brain identity thesis is still validated in the sense that there are only physical states. Mental terms are used to denote such physical states rather than some distinct types of mental states.

In contrast, in the case of $\mathcal{L}_M$, constants for space-time points are not added not because we regard them as problematic. Here, we share no verificationist view of meaning, nor do we think that spacetime points have to be reduced to something else, as mental states have to be reduced to physical states. The sole reason for not adding constants for spacetime points is that we wish to have isometry as our criterion of isomorphism.

An additional difference is that sometimes Ramseyfication is not only able to eliminate problematic terms, but also our ontological commitment to the objects denoted by such terms. Normally, if we Ram-

[8] "Has it meaning to say that the back of the moon has a surface of green cheese? If our theory allows as a possibility that we might go there or find out in any other way, then it has meaning. If not, not; i.e. our theory of the moon is very relevant, not merely our theory of things in general." (p. 195, 1929)

[9] An interesting tweak here is that by Ramseyfication, although one gets rid of the non-observational terms, one is left with quantifications ranging over them, which by Quine's thesis of ontological commitment, would still commit one to the existence of non-observable things. However, Ramsey himself does not seem to take this as a problem, as he seems to adopt a purely inferential view of quantification:"So far, however, as reasoning is concerned, that the values of these functions are not complete propositions makes no difference, provided we interpret all logical combination as taking place within the scope of a single prefix...For we can reason about the characters in a story just as well as if they were really identified, provided we don't take part of what we say as about one story, part about another." (p.194)

[10] Here ιx means "the x". It is a complicated issue whether such a definition still works if there happens to be no or more than two realizers of x, and Lewis himself changes his view several times on this issue. See Lewis (1972, 1994, 1997) and also Weatherson (2021).

seyfy over a sentence $\phi(c)$, we will still need the thing denoted by c to exist in order for the Ramseyfied sentence $\exists x\phi(x)$ to be true. But in Lewis' definition of pain, $Df(pain) := \iota x \exists x_1, ..., x_n T(x, x_1, ..., x_n)$, it happens to be the case that we can always use specific *physical* states to realize the existential quantification over $x_1, ..., x_n$, and thus Ramseyfication helps us to escape the ontological commitment to mental states distinct from physical states. This, however, is not the case for $\mathcal{L}_M$. In $\mathcal{L}_M$, quantifications are made explicitly over spacetime points, and we do not have substitutes other than spacetime points that can realize such quantifications. Thus, we are still fully committed to the existence, albeit not specific identities, of spacetime points, and hence $\mathcal{L}_M$ remains substantivalist.

A final remark: the most significant objection to Ramseyfication, i.e. the Newman problem (Demopoulos and Friedman, 1985), does not apply to $\mathcal{L}_M$. The general idea behind the Newman problem is that the Ramseyfication RT of a theory T is true iff the benign part of T is true (Ketland, 2004), and this could be an unpleasant result for certain application of Ramseyfication. For instance, in the standard context of structural realism,[11] we take the benign part to be observable terms and the problematic part to be theoretical terms, and one wishes to use RT to capture the structural content of T. However, the Newman Problem shows that RT captures only the *empirical* content of T, and thus collapses structural realism into a form of empiricism.

The Newman problem, however, is not a problem for $\mathcal{L}_M$. Although GR formulated in $\mathcal{L}_M$ contains indeed only propositions that do not mention specific identities of points, this is not a problem in itself. In particular, as we only Ramseyfy over spacetime points, but not over other theoretical terms such as the metric tensor field g or the maximal atlas $atlas$, we will still be able to talk about the structure of spacetime up to isometry. Instead, it simply means that $\mathcal{L}_M$ commits to a form of sophisticated substantivalism, which we call anti-specificsm. And while in the context of structural realism, we may have good reasons to reject collapsing to empiricism, in the current context of substantivalism, I do not see any reason to reject taking anti-specificism as a serious candidate. Indeed, as I shall argue in chapter 7, anti-specificism actually has many independently appealing features over its peers, including blocking the hole argument, and should be taken as the preferred version of (sophisticated) substantivalism.

[11] For application of Ramseyfication in structural realism, see Sneed (1971), Frigg and Votsis (2011). For the debate about the Newman problem, see Votsis (2003), Zahar (2004), Melia and Saatsi (2006), Ainsworth (2009), and Dewar (2019b).

Chapter 6

The Hole Argument and the Formalist Response

6.1 Introduction: the Hole Argument

The hole argument was first formulated by Einstein in 1913, purportedly showing that there cannot be any generally covariant theory of spacetime, which justifies his temporary failure to form such a theory.[1] Earman and Norton (1987) recast it as an argument against the *substantivalist* reading of General Relativity (GR), the view that spacetime points exist independently of matter fields. In particular, they argue that substantivalism leads to indeterminism.

Let $\mathcal{M} = (M, g_{ab})$ be a model of GR, where $\mathcal{M}$ is a four-dimensional manifold, and g_{ab} a Lorentzian metric field. Let d be a diffeomorphism from $\mathcal{M}$ onto itself. Let $d^*\mathcal{M} = (M, d^*g_{ab})$ be the model we obtained after dragging the metric field g_{ab} along d.

Assume that $\mathcal{M}$ can be foliated into space-like hypersurfaces, and Σ one such hypersurface. Consider the case where d is a hole diffeomorphism: it smoothly moves points inside a hole[2] H lying in the future of Σ, but leaves points outside H intact. Then $\mathcal{M}$ and $d^*\mathcal{M}$ seem to witness the indeterminism of GR: they are identical up to Σ, but differ in the future inside the hole H. See Figure 1 for an illustration.

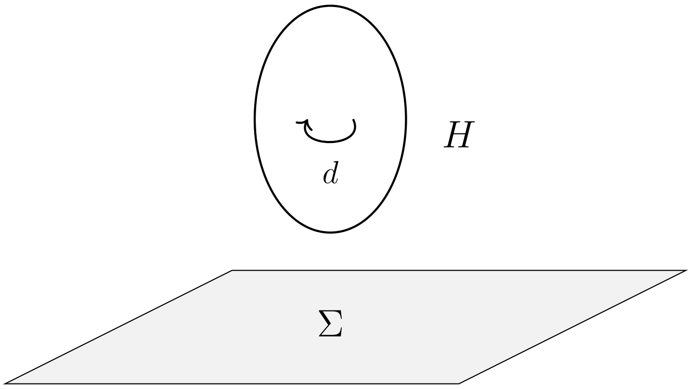

Figure 6.1: Illustration of the hole argument.

The argument can be specified as follows:[3]

- **P1**: $\mathcal{M}$ and $d^*\mathcal{M}$ represent two distinct physical situations S_1 and S_2.

[1] After such a generally covariant theory was found in 1915, Einstein rejected the hole argument by essentially accepting Leibniz equivalence. See Norton, Pooley and Read (2023).

[2] A hole may be defined as a compact open subset of M, see Pooley (2021).

[3] Note that **P1** and **P2** are often combined into a single premise in the literature. We separated them here to better classify different lines of response.

- **P2**: Models of GR represent physically possible situations.
- **P3**: If S_1 and S_2 are physically possible, then indeterminism holds.
- **C**: Indeterminism holds. (**P1 - P3**)

There are mainly two lines of responses to the hole argument: the metaphysical response and the formalist response. The metaphysical response refutes either **P2** (Butterfield, 1989; Maudlin, 1988, 1990; Gomes and Butterfield, 2023a, 2023b), asserting that S_1 and S_2 cannot be both possible, or **P3** (Belot, 1995; Butterfield, 1989; Melia, 1999; Pooley, 2021), arguing that S_1 and S_2 do not witness the indeterminism that we should care about. In contrast, the *formalist* response pays attention to the mathematical formalism used in the representation of spacetime, rejecting **P1**: $\mathcal{M}$ and $d^*\mathcal{M}$ really say the same thing about the world. (Weatherall, 2018; Bradley and Weatherall, 2022; Halvorson and Manchak, 2022).

The core idea of the formalist response is to accept the following principle:

> **Leibniz Equivalence** (**LE**): If two Lorentzian manifolds are isometric, then they represent the same physical situation.

Now, since $\mathcal{M}$ and $d^*\mathcal{M}$ are isometric, as witnessed by the hole diffeomorphism d, they represent the same physical situation according to **LE**. Thus, accepting **LE** will automatically refute **P1**, and hence block the hole argument.

The basic idea behind **LE** is that mathematicians in practice treat isomorphic mathematical structures indistinguishably. Thus, formalists propose the following principle to justify **LE**:

> **The Bradley-Weatherall Principle** (**BWP**): "the representational capacities of mathematical objects are *precisely* those preserved by isomorphism" (Bradley and Weatherall, 2022, p.1230).

And mathematical practice suggests that the relevant criterion of isomorphism for Lorentzian manifolds is isometry. Formalists then conclude from **BWP** that if we stick to Lorentzian manifolds as the standard formalism of GR, anything not preserved by isometry should be representationally irrelevant, and therefore **LE** holds.

The critics of the formalist response (say, *anti-formalists*) raise alleged counterexamples against **LE**. For instance, Roberts (2020) considers two misaligned half-planes $\mathcal{M}_1 = \langle M_1 = \mathbb{R} \times (0, +\infty), g_{ab}\rangle$ and $\mathcal{M}_2 = \langle M_2 = \mathbb{R} \times (1, +\infty), g_{ab}\rangle$ (both with an everywhere identical metric field g_{ab}). He argues that $\mathcal{M}_1$ and $\mathcal{M}_2$ cannot represent the same physical situation, since $\mathcal{M}_2$ only includes points above the line $y = 1$. But $\mathcal{M}_1$ and $\mathcal{M}_2$ are isometric by the map $(x, y) \mapsto (x, y + 1)$. Consequently, he concludes that **LE** is false.

Similarly, Belot (2018) and Fletcher (2020) use the swerve model as a counterexample against **LE**. "The swerve model" is essentially a system containing a single particle that starts to accelerate towards a certain direction at a certain time. Consider a swerve model $\mathcal{M}$, and the model $d^*\mathcal{M}$, where d is a spatial rotation of the manifolds. They argue that, despite $\mathcal{M}$ and $d^*\mathcal{M}$ being isomorphic, they really represent different physical situations (in the substantivalist representational context): the particles in $\mathcal{M}$ and $d^*\mathcal{M}$ are accelerating towards *different directions*.

Formalists respond to such counterexamples that they invoke "semantic metatheory". In particular, in the alleged counterexample to **LE**, to distinguish the two misaligned half-planes $\mathcal{M}_1$ and $\mathcal{M}_2$, one needs to invoke specific identities of plane points — so that we can say, e.g., $\mathcal{M}_1$ contains the point $(0, 1)$ while $\mathcal{M}_2$ does not. Similarly, we need to look into the set-theoretical details of $\mathcal{M}$ and $d^*\mathcal{M}$ in the swerve model in order to distinguish different directions towards which the particle is accelerating. Such specific identities, or specific directions are only expressible in our meta-theory of set theory, but not in the object theory of General Relativity. And meta-theory is representationally irrelevant.

The current chapter aims to give a detailed analysis of this dialogue. I shall argue that while the formalist response to anit-formalists can be formulated as a valid argument, it does not fully answer the anti-formalist challenge, as a key premise is left unjustified.

6.2 Reformulating the Formalist Response

Formalists rely their response on the distinction between meta-theory and object theory. This distinction essentially comes from the distinction between a physical domain and the representational tools which we use to represent that physical domain.

Recall from the syntax-semantics debate, that we can choose different representational tools to talk about the physical domain, among which we have theories (i.e., set of sentences), models ("logical mdoels", i.e., set-theoretical structures which interpret certain signature), or more generally, all kinds of different mathematical structures. But then if we also want to talk about the *representational tools themselves*, we need another level of representation. In this sense, object theories/models are what we use to directly say things about the physical domain, and meta-theories/models are what we use to say things about the object theories/models.

Now, formalists suggest that meta-theory is "representationally irrelevant". There is no doubt that meta-theory is not directly used for representing the physical domain, but instead for representing our representational tools. However, this does not imply that details of meta-theory are irrelevant to representation, especially when our meta-theory already contains the object theory as its part. An extreme case is where the meta-theory/language is just the object theory/language, e.g., when we talk about a natural language in itself. We articulate the truth condition of a sentence by saying, e.g., the sentence "grass is green" is true iff grass is green. In this case, details of the meta-theory are indeed representationally relevant.

The claim, therefore, should better be modified as saying that those details that present *only* in the meta-theory but *not* in the object theory are representationally irrelevant. Note that then the distinction between meta-theory and object theory is not really needed, since all formalists are saying here is that only details of the object theory, which is simply the physical theory in question, are representationally relevant. But this is almost a tautology: only details of a theory T, but not details of some other theory T', contribute to the representational contents of T. To say that some background theory T' is also tacitly assumed to represent in addition to T seems simply to say that one's theory of spacetime is not T but really $T \cup T'$.

The formalist response can then be summarized as pointing out the incompatibility of the following two items for any theory of spacetime T:

- T commits to specific identities of spacetime points, or specific directions in spacetime, etc.
- The isomorphism criterion of T-models is isometry.

The argument can then be summarized as follows.

1. T commits to specific identities of spacetime points (respectively, specific directions in spacetime).

2. Only details of T, but not details of some other theory T', contribute to the representational contents of T.

3. T must include constants for spacetime points (respectively, directions) or mathematical constructions with a similar function. (By 1,2)

4. The isomorphism criterion of T-models is required to preserve such constants or similar mathematical constructions. (By 3)

5. The isomorphism criterion of T-models cannot be isometry. (By 4)

The argument is not formulated in a deductive manner, but I shall argue that each premise is justified and each move is valid. Item 1 is simply our assumption. Item 2 is true by the functional roles theories play in the process of representation, i.e., representational tools. Item 3 follows from item 1 and item 2. Item 4 is true by the concept of isomorphism. The only non-trivial move might be the last move from item 4 to item 5. But as shown at the end of chapter 5, an isometry is *precisely* a map that preserves the set of open sets, the maximal atlas, and the metric tensor field associated with a metric manifold, and hence cannot preserve any extra mathematical constructions such as constants for spacetime points. Indeed, the language $\mathcal{L}_M$ we designed in Section 5.5 whose isomorphism criterion coincides with isometry is a language that does not contain any constants for spacetime points or directions.

If we adjoin the above argument with the following assumption:

> **Isometry as Isomorphism (II)** Lorentzian manifolds are mathematical structures whose isomorphism criterion is isometry.

Then we simply arrive at the conclusion that if we stick to Lorentzian manifolds as our representational tools for spacetime, then there are no specific identities of spacetime points or specific directions in spacetime, and hence the antiformalist counterexamples fail. Note that **II** is supposed to be true simply by what we mean by "Lorentzian manifolds". In mathematical practices, people usually define Lorentzian manifolds as simply set-theoretical structures for the sake of convenience, and then say, in a hand-waving manner, that we only care about their structures up to isometry. The formalist reading of such practices is that the definition of Lorentzian manifolds given as set-theoretical structures is only a working definition, a definition that contains many concrete set-theoretical details that are not inherently part of the structure of Lorentzian manifolds, and the caveat that we only care about their structures up to isometry should be taken more seriously to suggest that Lorentzian manifolds are only defined up to isometry, say, as the structures characterized directly by $\mathcal{L}_M$.[4]

Alternatively, one may suggest that Lorentzian manifolds are literally structures defined in set theory. But then, as we are working in the language of set theory (together with a constant denoting the Lorentzian manifold in question as a set), the isomorphism criterion will not be isometry, but

[4] This is just like first defining groups as the sets of automorphisms of vector spaces, and then say, in a hand-waving manner, that we only care about their structures up to homomorphism. If we follow the formalist interpretation, then groups are really only defined up to homomorphism, as the algebraic structures characterized by the usual axioms of groups. The only difference is that it is much easier to work with an explicit theory of a group whose isomorphism is homomorphism than to work with an explicit theory of Lorentzian manifolds whose isomorphism is isometry. This explains why people use algebraic definitions for groups in practice, but rest assured with a set-theoretical definition for Lorentzian manifolds.

rather set-theoretical isomorphism, and so **II** would be false. To illustrate, in any model U of set theory, two isometric Lorentzian manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ can be found such that $\mathcal{M}_1$ happens to contain the empty set as a point in the manifold, while $\mathcal{M}_2$ does not. Then the sentence "there is an emptyset in the manifold" will be true of $\mathcal{M}_1$ but false of $\mathcal{M}_2$.[5] As isomorphism always implies elementary equivalence, $\mathcal{M}_1$ and $\mathcal{M}_2$ cannot be isomorphic, although they are indeed isometric.

I think there are good reasons to prefer the formalist interpretation over the second interpretation. Statements such as "The emptyset is in the domain of the metric manifold $\mathcal{M}$" seem simply irrelevant to differential geometry, and not included as part of our common understanding of what the structure of Lorentzian manifolds amounts to. It also plays no role when physicists use Lorentzian manifolds to represent the structure of spacetime, as it has no physical meaning to say that a spacetime point is the emptyset.

But in any case, I regard it as a red herring to argue over what the term "Lorentzian manifolds" really means, since the question we should really be asking is what kind of structures the spacetime really possesses, whether it should be called "Lorentzian manifolds" or not. Thus, I simply choose the convention, which I prefer for the above reasons, that the term "Lorentzian manifolds" refers to the mathematical structures characterized by $\mathcal{L}_M$ whose isomorphism criterion is indeed isometry. Thus, **II** is true by convention.

We have seen how counterexamples raised by anti-formalists can be refuted. But this does not mean that the formalist response is simply saved. For while the details about specific points or directions are not part of the structure of Lorentzian manifolds, given the way a term is currently used, there are three lingering questions to be answered:

- If one sticks to Lorentzian manifolds as one's representational tools, is the hole argument blocked?
- Why *should* we stick to Lorentzian manifolds as representational tools to represent spacetime (rather than, say, certain extensions of Lorentzian manifolds)?

Another independent but related question is:

- Do physicists *actually* use Lorentzian manifolds as their representational tools in practice?

We will address these questions in the following sections.

6.3 Does the Formalism Block the Hole Argument?

In this section, we analyze whether the hole argument is blocked if one sticks to Lorentzian manifolds as our representational tools for spacetime. I argued that the formalist strategy of using **BWP** to defend **LE** is invalid, as **BWP** focuses on the wrong target of representational capacities but not representational contents, and **BWP** itself is also only partly correct. Instead, in this section, I will construct a valid justification of **LE** based on an alternative principle, **GLE**. Thus, I conclude that, sticking to Lorentzian manifolds, the formalist response successfully blocks the hole argument— as long as grounded in the correct argument.

[5] More rigorously, true of $(U, \mathcal{M}_1)$ but false of $(U, \mathcal{M}_2)$: since we are working with full-blooded set theory, we always need to keep a model of set theory in the background to interpret the set-theoretical language.

6.3.1 Representational Capacities vs. Representational Contents

Previously, we gave a conceptual framework for representation where we have the crucial component of *representational context.* This element is sometimes ignored in the literature. Recall that **P1** of the hole argument suggests that $\mathcal{M}$ and $d^*\mathcal{M}$ say different things about the world. But a pure mathematical object alone cannot represent anything —it only represents when put into a representational context. Just as the string "red" only denotes red when *we interpret it as such* (we could choose to interpret "red" as denoting *blue* as well).

Once we include the element of representational context, we are able to point out an important distinction, that is, the distinction between representational contents and representational capacities. The representational content of a model is relative to *a specific representational context*, while the representational capacities of a model take into account how a model can be used for representation across *all possible representational contexts* (within a representational protocol).

In view of this distinction, we can see that there is an ambiguity in **LE** that needs to be resolved. Recall, the original version of **LE** says:

> (**LE**) If two Lorentzian models $\mathcal{M}$ and $\mathcal{M}^*$ are isometric, then they are representationally equivalent.

But it is ambiguous whether representational equivalence refers to the equivalence of representational capacities or the equivalence of representational contents in some specific contexts. Indeed, the terminology in the literature can be confusing. Fletcher (2020) and Belot (2018) use the term "representational equivalence" for equivalence of representational *capacities*, but "representational distinctness" for distinctness of representational *contents* in any common representational context.

Accordingly, we can distinguish two theses of Leibniz Equivalence:[6]

- Leibniz Equivalence of Capacities: (**LEcap**) If two Lorentzian manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ are isometric, then, they have the same representational capacities, i.e., for any representational context C_1, if $\mathcal{M}_1$ represents a physical situation S in C_1, then there is a representational context C_2 such that $\mathcal{M}_2$ represents S in C_2, and vice versa.
- Leibniz Equivalence of Contents: (**LEcon**) If two Lorentzian manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ are isometric, then, for any representational context C, if $\mathcal{M}_1$ represents a physical situation S in C, then $\mathcal{M}_2$ also represents S in C, and vice versa.

The premise of the hole argument **P1** should also be specified in terms of representational contexts. Recall,

> **P1**: $\mathcal{M}$ and $d^*\mathcal{M}$ represent two distinct physical situations S_1 and S_2.

The specification of **P1** that is relevant to the hole argument is the one which restricts to the representational context adopted by substantivalists:

> **P1***: For the representational context C adopted by the substantivalists of GR, $\mathcal{M}$ in C and $d^*\mathcal{M}$ in C represent two distinct physical situations S_1 and S_2.

[6] There are many similar specifications of Leibniz Equivalence in the literature. "Strong Leibniz equivalence" vs. "Weak Leibniz equivalence" in Roberts (2020), **RUME** vs. **REME** in Fletcher (2020), and **RUME*** and **MIRD*** in Pooley and Read (2021) etc..

Formalists must refute **P1*** to block the hole argument.

Now it is not hard to see that the only version of Leibniz equivalence that can refute **P1*** is **LEcon**, but not **LEcap**, since the latter does not guarantee that in a *common* representational context, $\mathcal{M}$ and $d^*\mathcal{M}$ will represent the same physical situation. However, the formalist justification of **LE** is to invoke the Bradley-Weatherall principle, which can be paraphrased as:

> **BWP**: Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic iff they have equivalent representational capacities.

Clearly, **BWP** only justifies **LEcap**, but not **LEcon** (which is strictly stronger than **LEcap**), and so cannot block the hole argument. Echoing some anti-formalists (Roberts, 2020; Pooley and Read, 2021), I conclude here that the formalist defense focuses on the wrong target, i.e., representational capacities.

While **BWP** does not help justify the relevant version of Leibniz equivalence, I shall still give it an independent evaluation, as different variants of **BWP** are prevalent in the literature of philosophy of physics, especially in discussions of symmetries and dualities. The next subsection, therefore, is devoted to the task. As we shall see, not only do formalists focus on the wrong target, but their statements about this target are also only partly correct.

6.3.2 On the Bradley-Weatherall Principle

The Bradley-Weatherall principle is a widely held speculative principle that has not been thoroughly evaluated. Various versions of it can be found in the literature of philosophy of physics:[7]

- Two solutions of a classical theory's equation of motion are related by a symmetry if and only if ... they are equally well- or ill-suited to represent any particular physical situation. (Belot, 2013, p. 1)
- Two models of a physical theory are symmetry-related iff they can represent the same possible physical situations. (Luc, 2022, p. 72)

Here we evaluate **BWP** first by informal reasoning and then through formal proofs based on the formal representation framework we developed previously.

I suggest that the direction of **BWP** from the right to the left is true. If two models are isomorphic, then the representational tool we are using is *literally the same mathematical structure* (meta-)represented by the set-theoretic models. Therefore, as we are using the same representational tool, we can only capture the same range of representational contents.

> (Theorem 3.28) If two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic, then they have equivalent representational capacities.

With the assumption that the relevant criterion of isomorphism for Lorentzian manifolds is isometry, the above theorem logically entails **LEcap**. So at least, one version of **LE** is indeed correct, though unable to block the hole argument.

However, the other direction of **BWP**, that equivalence of representational capacities entails isomorphism, is not true in general. First, there is no guarantee that two models with equivalent

[7] See Hall and Ramírez (2024) for more examples.

representational capacities even share the same signature. For example, the standard model of natural numbers $\langle N, S, 0\rangle$ and its notational variant $\langle N, S', 0'\rangle$ will clearly have the same representational capacities but are not isomorphic, since they do not share the same signature.

A more charitable reading is to claim that *assuming the same signature*, equivalent representational capacities entails isomorphism. This, however, is false as well.

> (Theorem 3.29) There are $\mathcal{L}_\Sigma$ models $\mathcal{M}_1$ and $\mathcal{M}_2$ such that they are not isomorphic but have equivalent representational capacities.

One might hope if equivalence of representational capacities does not entail isomorphism, it should at least entail some weaker criterion of *structural equivalence.* Popular candidates raised in the literature include definitional equivalence, Morita equivalence, categorical equivalence, etc. However, as proved in Theorem 3.30, there are models with equivalent representational capacities which are not definitionally equivalent, Morita equivalent or even categorical equivalent.

While the formal proof depends on the details of the representational protocol we adopt, the intuitive idea behind technical results is straightforward and more general. To illustrate, consider Figure 3.

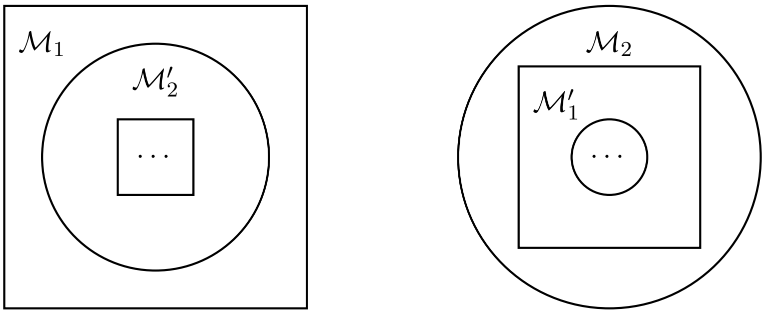

Figure 6.2: Models with equivalent representational capacities but not structurally equivalent.

It is intuitive that $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same representational capacities. After all, no matter how $\mathcal{M}_1$ is used for representation, $\mathcal{M}_2$ can mimic the representation of $\mathcal{M}_1$ using the copy $\mathcal{M}_1'$ contained within $\mathcal{M}_2$, so $\mathcal{M}_2$ will be able to do the same job, and vice versa. However, there is an intuitive sense in which $\mathcal{M}_1$ and $\mathcal{M}_2$ are *not* structurally equivalent: as they still disagree about what the "shape" of the outermost universe is, whether it is to be characterized as $\mathcal{M}_1$ or $\mathcal{M}_2$. Thus, it is unsurprising that the equivalence of representational capacities does not entail popular characterizations of structural equivalence, from isomorphism to Morita equivalence.

In sum, the right-to-left direction of **BWP** (isomorphism implies equivalence of representational capacities) is true, while the left-to-right direction (equivalence of representational capacities implies isomorphism) is false, and false even if we focus on weaker criteria of structural equivalence, such as definitional equivalence or Morita equivalence. Therefore, not only does the formalists' reply to their critics focus on the wrong target, i.e., representational *capacities* instead of representational *contents*, but what they say about the wrong target is also only partly true.

6.3.3 On Leibniz Equivalence of Contents

Now we have seen that the formalist defense of **LE** fails, as **BWP** is only partly correct and at most justifies **LEcap**, which is too weak to block the hole argument. What is needed is the strictly stronger thesis **LEcon**. In this section, I argue that the formalist response can be saved by reconstructing a *valid* justification for **LEcon**, based on the formal framework developed in chapter 3.

I propose the following argument for **LEcon**:

- **Generalized Leibniz Equivalence** (**GLE**): If two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are isomorphic, then, for any representational context C, if $\mathcal{M}_1$ represents a physical situation S in C, then $\mathcal{M}_2$ also represents S in C, and vice versa.

- **Isometry as Isomorphism** (**II**): The relevant criterion of isomorphism for Lorentzian manifolds is isometry.

- **LEcon** (from **GLE** and **II**)

This argument follows the general line of thought in Weatherall (2018) and Halvorson and Manchak (2022), though it concerns representational *contents* rather than representational *capacities*. Indeed, based on the representational protocols we developed previously, **GLE** can be stated and proved as follows.

> (Theorem 3.27) For any two isomorphic $\mathcal{L}_\Sigma$-models $\mathcal{M}_1$ and $\mathcal{M}_2$, the representation $\Lambda \overset{\delta}{\to} \mathcal{M}_1$ is equivalent to $\Lambda \overset{\delta}{\to} \mathcal{M}_2$ for any δ.

Another subtle difference between the above argument and the original argument of formalists is that the concept of isomorphism used here is *logical* isomorphism rather than *categorical* isomorphism. As we commented previously in Chapter 3, it is not clear how the category of physical models relates to the process of representation. And so if the argument is interpreted in categorical terms (see Pooley and Read, 2021), it is would be hard to give an explicit argument for the statement that if two models are isomorphic in some category, they must represent the same content. However, adopting the notion of logical isomorphism and building the formal framework of representation based on the idea of representation as definition, we are able to show how isomorphism influences representational contents, and rigorously prove the statement we need as Theorem 3.27. But since that logical isomorphism for a theory T coincides with categorical isomorphism in $Mod(T)$, the above difference is more about reasoning strategies but not the final conclusion.

Thus, we have vindicated the truth of **GLE**. And another assumption **II** is presumably a conceptual truth about what kind of mathematical structures the term “Lorentzian manifolds” refers to, as used by physical mathematicians. Therefore, **LEcon** is justified based on two plausible assumptions **GLE** and **II**.

6.4 Do Physicists Actually Use Lorentzian Manifolds for Representation?

We have seen that if we stick to Lorentzian manifolds as our representational tools, then the hole argument is indeed blocked. The counterexamples raised by antiformalists invoke resources in the meta-theory extra to the structure of Lorentzian manifolds, and hence do not attack **LE**. However, such counter-examples could also be interpreted as suggesting that we *should not* stick to Lorentzian manifolds as our representational tools, but instead we should use certain extensions of Lorentzian manifolds, e.g., Lorentzian manifolds extended with specific identities of spacetime points or directions.

In this section, I shall first discuss a related but independent question, do physicists *actually* stick to Lorentzian manifolds as their representational tools in practice?

This claim is presumably vindicated by standard textbooks of spacetime physics.[8] But some anti-formalists (Gomes and Butterfield, 2023; Landsman, 2023; Cudek, 2024) challenge this claim by alleged counter-examples. In this subsection, I examine four such examples closely and defend the formalist assumption against each.

6.4.1 Two Misaligned Planes and the Swerve Model

We first consider the examples already discussed previously, i.e., the two misaligned planes (Roberts, 2020) and the swerve model (Belot, 2018). These examples could be interpreted as suggesting that physicists really distinguish between isometric manifolds, and so we need specific constants of either spacetime points or directions in the vocabulary of GR to make sense of such practices. However, this extension is not really needed to explain the practices. As argued by Luc (2022), these types of example used by physicists are actually implicitly considered to represent *sub*systems of our universes. And it is perfectly fine for formalists to distinguish isometric subsystems, since they are isometric only with respect to the quantities within the subsystems, and we can still use the global resources to distinguish them. So, indeed, if we already have, in our background system, some particle that swerves towards a certain direction at a certain time, then we can use such a background particle as our reference frame to distinguish situations in which the particle under current investigation swerves towards different directions. This can be done even though we do not have constants for spatial directions and stick to Lorentzian manifolds as our representational tools, where we only have quantification over but no labeling for spacetime points.

6.4.2 Lie Derivative

Some people (Gomes and Butterfield, 2023; Landsman, 2023) argue that formalists cannot make sense of the definition of Lie derivative:

$$L_X g(x) = \lim_{t \to 0} \frac{1}{t}(g(x) - (\phi_t)_* g(\phi_{-t}(x)))$$

They submit that this definition requires one to fix the spacetime points while drag the field g along the flow ϕ. Arnold (1989, p. 198) gives the metaphor that taking the Lie derivative at some point x, is like having a fisherman sitting still at x and taking derivative with respect to the river flow (the flow of the field g) passing in front of him. They argue that formalists cannot make sense of this picture, since if the identities of spacetime points are fixed by their field values, then the spacetime points cannot stand still while their field values are dragged elsewhere. Rather, the spacetime points always stick to their field values. Thus, $L_X g(x) \equiv 0$ for all x, since the field value attached to a spacetime point does not change.

The genuine difficulty of evaluating this argument is that it is way too ambiguous to be understood. The main part of the argument relies on a metaphor of fishing, and not every element of the metaphor has a clear correspondence in differential geometry. While it is pretty clear what it means to say “the river flows while the stones on the riverbed are fixed”, what does it mean to say “the field g is dragged along the flow ϕ *while the spacetime points stand still*”, or that “a point x would stick to its field

[8] For instance, see Hawking and Ellis, 1973, p. 56 and Wald, 1984, p.260 and p.438

value"? There is some genuine ambiguity in understanding these terms, which creates a difficulty in assessing whether the formalist position really leads to such proposed consequences.

The most charitable interpretation of the argument seems to be this: As formalists identify two isometric spacetimes, if a point x is related to another point y by isometry, then they must be identified, and therefore, if we drag the field g along the flow ϕ, we construct an isometry between a point x and $\phi_t(x)$. Thus, according to the formalist position, x and $\phi_t(x)$ must be the same point, and hence $L_X g(x) \equiv 0$.

But this interpretation relies on a false interpretation of the formalist position: if two points are related by isometry, they must be the same point. This is typically what metrical essentialists would say in spirit, *but not formalists*. Rather, sticking to Lorentzian manifolds identified up to isometry only means that there is simply *no matter of fact* whether it is this point or that point, but *not* that the two points are the same. Consider a Lorentzian manifold that has two distinct points related by isometry. By the formalist response, the only fact of the matter is that there are two distinct points which instantiate the same set of metric properties — which, however, is completely different from saying that, since they share the same metric properties, they are identified as *the same point*. Indeed, $\mathcal{L}_M$ identifies Lorentzian manifolds only up to isometry, but we can still formulate a sentence in $\mathcal{L}_M$ which says that there is a function f from θ to θ such that f is an isometry (i.e., f is continuous, diffeomorphic and preserves the metric value), yet there exists a point x such that $f(x) \neq x$.[9] We can further find $\mathcal{L}_M$-models where this sentence is made true. Thus, interpreted in this way, the anti-formalist argument is based on a false assumption about what the formalist position commits to.

There might be other interpretations of Arnold's metaphor, but here I shall simply argue that the formalist shall have no problem dealing with the Lie derivative in principle. In particular, the Lie derivative can be defined and calculated in $\mathcal{L}_M$ (with the parameter X denoting an arbitrarily chosen smooth vector field), a formalism whose isomorphism criterion is precisely isometry, and does not contain any specific identities of spacetime points. This can be easily seen as follows: the flow ϕ can be defined from X; g is already contained in $L_{\mathcal{M}}$ as a constant, the limit operation is also defined for the reals. Thus, the Lie derivative can be defined as (in sketch):

$$\begin{aligned} L_X g(x^\theta) = k^{\mathcal{R}} := \exists U^{p\theta} \exists \epsilon^{\mathcal{R}} \exists \phi (x^\theta \in U^{p\theta} \wedge (\phi : (-\varepsilon^{\mathcal{R}}, \varepsilon^{\mathcal{R}}) \times U \to \theta) \\ \wedge \forall y^\theta \in U (\phi(0, y^\theta) = y^\theta) \wedge \frac{\partial}{\partial t}\phi(t, y^\theta) = X(\phi(t, y^\theta)) \\ \wedge k^{\mathcal{R}} = \lim_{t \to 0} \frac{1}{t}(g(x^\theta) - (\phi_t)_* g(\phi_{-t}(x^\theta)))). \end{aligned}$$

Here, ϕ is an abbreviation for $\phi^{\mathcal{P}(\mathcal{R} \times \theta \times \theta)}$.

For any Lorentzian manifold $\mathcal{M}$, it is easy to check that the interpretation of the Lie derivative as defined above must coincide with the standard Lie derivative defined directly in set theory, and thus we cannot have the problem that $L_X g(x) \equiv 0$ for all x.

[9] More rigorously, this can be said in some definitional extension of $\mathcal{L}_M$, where we add the domain of functions over θ. Still, the isomorphism criterion remains to be isometry, as the preservation of the set of open sets, the maximal atlas and the metric tensor field automatically guarantees the preservation of things that are definable from them. The same point holds for the definition of Lie derivative below.

6.4.3 Limits of Spacetimes

Gomes and Butterfield (2023) argue that in the definition of the limit of a family of spacetime it is useful to identify (or in their terminology, thread) points by congruence of curves even if they are not related by isometry. Cudek (2024) concludes from this example that the language of GR should include the entire first-order language of set theory, and hence physicists do distinguish between isometric manifolds.

However, the fact that physicists do use congruence of curves other than isometries in physical practices per se does not entail that such congruence of curves must have physical significance, for such talks may only be included for heuristic or practical purposes. And I shall argue that this is indeed the case. Limits of spacetimes are defined in Geroch (1969) as follows:

Definition 6.1. Let $\langle M_\lambda, g_{ab}(\lambda)\rangle$ be a family of Lorentzian manifolds parameterized by λ, which we may assume to be a foliation of a five-dimensional manifold M. A limit space of M is then defined as a tuple $\langle M', g'_{ab}, \lambda', \Psi, \partial M'\rangle$ where:

- M' is a 5-manifold with boundary $\partial M'$,
- λ' is a smooth scalar field on M', and $\partial M'$ is obtained by setting $\lambda' = 0$,
- g'_{ab} is a tensor field on M', with signature $(0, +, -, -, -)$ on $\partial M'$,
- Ψ is a smooth, one-to-one mapping from M onto the interior of $\partial M'$ such that it takes g_{ab} to g'_{ab} and λ to λ'.

Note that, as Gomes and Butterfield also admit, the definition of limit per se does not rely on any threading or identification of points other than isometries. The need for a family of frames is only for the purpose of explicitly *calculating* metrics in some specific limits we wish to study. For instance, if we wish to find a limit of Schwarzschild spacetimes, it is hard to use the above definition directly, say, by putting Schwarzschild spacetimes into a five dimensional manifold with a well-defined boundary. But it would be much easier if we work with explicit coordinate frames. Let the family of Schwarzschild spacetimes index by λ ($\lambda = m^{-\frac{1}{3}}$). Using the standard Schwarzschild coordinates, we can write the indexed family of metrics as:

$$ds^2 = \left(1 - \frac{2}{\lambda^3 r}\right) dt^2 - \left(1 - \frac{2}{\lambda^3 r}\right)^{-1} dr^2 - r^2\Big(d\theta^2 + \sin^2\theta\, d\phi^2\Big).$$

If we work with the family of frames generated by the coordinate transformation $x = r + \lambda^{-4}$ and $\rho = \lambda^{-4}\theta$, the metric will converge to the Minkowski metric when $\lambda \to 0$, and we will obtain the flat spacetime as a limit of Schwarzschild spacetimes. In Gomes and Butterfield's words, we are *threading* any two points a and b lying on the Schwarzschild manifold M_{λ_1} and M_{λ_2} if $r(a) + \lambda_1^{-4} = r(b) + \lambda_2^{-4}$. But this "threading" here only serves as a tool of calculation, and there are infinitely many distinct threadings that deliver the same flat limit. Indeed, any threading induced by $x = r + \lambda^{-n}$, $\rho = \lambda^{-n}$ for any $n \geq 4$ will do. Thus, while it may be argued that whether the flat spacetime is a limit of Schwarzschild spacetimes is a question with physical significance, it is hard to see that the seemly arbitrary choice of threading (e.g., the arbitrary choice of $n \geq 4$ in $x = r + \lambda^{-n}$, $\rho = \lambda^{-n}$) that is used to calculate such a limit can be physically significant as well.

And indeed, the question of whether the flat spacetime is a limit of Schwarzschild spacetimes can presumably be defined and answered using variations of $\mathcal{L}_M$. Here, I give a sketch of the basic idea. Notice that $\mathcal{L}_M$ talks only about a single Lorentzian manifold, say $\mathcal{M}_1$, with the basic domain θ. To talk about limits, we only need to add the domain symbols and constant symbols for another five-dimensional Lorentzian manifold, say $\mathcal{M}_2$. Let γ be its basic domain symbol. We add as our axiom that $\mathcal{M}_2$ can be associated with a foliation such that each leaf is a four-dimensional Schwarzschild spacetime, and that $\mathcal{M}_1$ is Minkowski. The question of whether the Minkowski spacetime is a limit of Schwarzschild spacetimes can then be expressed as a question about *whether there exists* a five-dimensional manifold structure over the domain $\gamma \cup \theta$, together with a smooth scalar field λ, such that:

- its boundary is defined by $\lambda = 0$, and is precisely $\mathcal{M}_1$;
- its interior is precisely $\mathcal{M}_2$, and λ gives a foliation of $\mathcal{M}_2$ such that each leaf is a four-dimensional Schwarzschild spacetime.

It is easy to verify that the definition given above coincides with Definition 6.1 and therefore gives the same answer "yes" to the question. And the extended language which talks about two Lorentzian manifolds (one four-dimensional, another five-dimensional) still characterizes them only up to isometry, without any extra structures such as threadings or specific identities of spacetime points.

6.4.4 Quantum Reference Frames

The last example mentioned by Gomes and Butterfield (2023) is quantum reference frames where "one should anchor the labeling of spacetime points onto the trajectories of the masses involved" (p.24) To be more precise, Giacomini and Brukner (2023) develop the framework of quantum reference frame by adopting an operational view of reference frames, where reference frames are seen as physical systems in general. So spacetime points in different manifolds are operationally identified through their relationship to the physical points of our apparatus. For instance, the locations of a probe particle in different superposed spacetimes will be identified as the same. Thus, we are evaluating the gravitational field "not at an abstract spacetime point but at the location of a particle" (p.5).

Again, we can perform such "identification" within the structure of Lorentzian manifolds. Although we cannot name specific spacetime points in a Lorentzian manifold, there is no issue with talking about spacetime points by the properties they satisfy, and in particular, their relationship with our reference frames as physical systems. So we can uniformly speak about the operationally identified point by the formula $\exists x(\phi(x) \wedge (...))$ where $\phi(x)$ says that x is the position of the probe particle (or, the quantum state of the probe particle in the positional basis has support at x), and (...) the placeholder for whatever one wishes to express about the spacetime point operationally identified. In any case, formalists have no problem with such *operational* identification of spacetime points, which is essentially identification relative to our reference frame as a physical system, in contrast with *absolute* identification, which requires identification of spacetime points independent of any physical systems or spacetime structures they are related to.

6.4.5 Summary of Lessons

Thus, none of the above four examples require deviation from the structure of Lorentzian manifolds, characterized up to isometry. The lessons can be summarized as follows:

- Formalists do not commit to the view that isometric subsystems are not distinguished, but only that isometric universes are not distinguished.
- Formalists do not commit to the view that isometry-related points are identical, but only that isometry-related points are indistinguishable.
- Not all mathematical constructions that appear in physical practices have physical significance. Some may just serve for heuristic or practical purposes.
- Formalists can make perfect sense of operational or physical identification of spacetime points using quantification, but not absolute identification, which identifies spacetime points independently of any physical systems or spacetime structures they are related to.

Thus, we conclude that physicists do stick to Lorentzian manifolds (recalled from Section 6.2, structures characterized up to isometry), as their representational tools in GR. Thus, in the following sections, by *the formalism of GR*, we refer to Lorentzian manifolds, i.e., structures characterized by $\mathcal{L}_M$, or equivalently, objects in the category of Lorentzian manifolds $Mod(\mathcal{L}_M)$.

6.5 Why Should We Stick to Lorentzian Manifolds?

The last section defends that physicists do use Lorentzian manifolds as the standard representational tools for spacetime. But this does not mean that we *should* stick to Lorentzian manifolds as our representational tools. In particular, the examples discussed in the previous sections could also be taken to suggest that we should extend our representational tools by adding further structures, say, specific identities for spacetime points or directions, into the structure of Lorentzian manifolds. In this section, I shall argue that this is a necessary question that one needs to answer in order to refute the hole argument, and that the current strategies offered by formalists are not satisfactory.

Now, at first glance, it seems that formalists take the burden of reasoning to be on the side of anti-formalists. Mundy (1992) says:

> "philosophers are welcome to construct modal extensions of physical theories. However, I claim that nothing in standard physical theory supports such extensions: no scientific problem requires the introduction of any primitive relation extending across different models." (p.522)

Bradley and Weatherall (2024) echo:

> "although one can extend the theory to accommodate the hole argument, there are no empirical or scientific justifications for doing so." (p.1229)

They are definitely correct that there is no physical motivation for extending the formalism of GR. But recall that the hole argument is not really an argument that targets the theory of GR but the *substantivalist interpretation* of GR, which is a philosophical position about spacetime. And from a substantivalist point of view, there are indeed good, and even compelling, *philosophical* motivations to seek such extensions. [10]

[10] This is presumably also recognized by some formalists. For instance, Weatherall (2018) explicitly admits that his aim is not to defend substantivalism, but the formalism of GR per se. And he believes that if additional structures to Lorentzian manifolds are needed by substantivalists, then "would-be" substantivalists, in order to reply effectively to the

Note that while Lorentzian manifolds are now shown to be free from the hole argument, such extensions of Lorentzian manifolds would revive it. For instance, consider the case where we extend the structure of Lorentzian manifolds by constants for spacetime points,[11] and have two models $\mathcal{M}_1$ and $\mathcal{M}_2$ where in $\mathcal{M}_1$, c_1 lying on the future of c_2, yet in $\mathcal{M}_2$, c_1 and c_2 are permuted. Then we can have the following variant form of the hole argument:

- **P1′**: $\mathcal{M}_1$ and $\mathcal{M}_2$ represent two distinct physical situations S_1 and S_2.
- **P2′**: S_1 and S_2 are both physically possible.
- **P3′**: If S_1 and S_2 are both physically possible, then indeterminism holds
- **C′** indeterminism holds.

Now refuting **P1′** along the formalist route is no longer possible, as our representational tools are now extensions of Lorentzian manifolds and hence the relevant criterion of isomorphism will not be isometry. Refuting **P2′** or **P3′** collapses the formalist response into our old metaphysical response. So, the only intelligible and nontrivial move for formalists is to reject such extensions of the formalism of GR. Otherwise, there is still a gap between the formalist conclusion (the formalism of GR is free from indeterminism) and the intended conclusion (the metaphysical reality of spacetime conceived by substantivalists is free from indeterminism). No wonder that some regard the formalist response as irrelevant to the debate of the hole argument (Pooley, 2021; Pooley and Read, 2021; Teitel, 2021).

Intuitively, it is hard to justify such rejections. At the very least, substantivalists should accept that space-time points *exist*. And there is an intuitive appeal that if something really exists, then it can be represented by adding a constant to our domain. As Maudlin (1988) puts it,

> After all, if event locations are fully in the ontology, why should we not be able to refer to them as specific individuals? The restriction to bound variables simply has no reasonable justification within the substantivalist program. (p.84)

So if something exists, then we should be able to talk about it, in one language or another, perhaps in some extension of the language of GR, if needed.

Possible responses currently available in the literature to justify the rejection of such extensions are not really successful. We consider a series of responses below.

First of all, to say that naming or reference, perhaps because it is part of our metatheory, is irrelevant to the representational capacities of Lorentzian manifolds is no rescue: we are no longer talking about Lorentzian manifolds anymore (the formalism of GR per se is saved), but whether substantivalists are committed to certain *extensions* of this formalism that leads eventually to indeterminism. And meta-semantic notions, such as naming, are indeed relevant in the sense that if it is possible, e.g., to name space-time points, then it will also be possible to extend our Lorentzian manifolds with constants.

Secondly, the answer given by Halvorson and Manchak (2022) also fails.

hole argument, need to stipulate what the additional structure might be and why we should think it matters."(p. 344) However, while Weatherall believes that this is not a promising route, the following discussions suggest that it is natural and even *compelling* for substantivalists to extend the formalism of GR, construed as Lorentzian manifolds.

[11] See Rynasiewicz (1996) for the suggestion of adding constants to Lorentzian manifolds. Stachel (1989) also discusses a similar kind of extension by adding individuating fields to Lorentzian manifolds.

> One possibility is that the substantivalist theory includes constant symbols for picking out spacetime points ... But that can hardly be the intention of the substantivalist, because in that case he would be committed to Minkowski spacetime having no symmetries. (p.12)

However, it is not clear whether it is problematic to have no symmetry in our spacetime. After all, one may hold that yes, there are no symmetries "in the strictest sense", but there are plenty of *qualitative* symmetries, that is, maps that preserve all quantities that do not mention specific points. For instance, consider the two-dimensional Euclidean space $\mathbb{E}^2$ with the flat metric. If we label each point with a constant, say $\langle x, y\rangle$, where x, y are reals, then the map $f : \mathbb{E}^2 \to \mathbb{E}^2$ defined by $f(\langle x, y\rangle) = (\langle x+1, y+1\rangle)$ is not a symmetry, as it will map the point with label $\langle 0, 0\rangle$ to the point with label $\langle 1, 1\rangle$. However, one may say that f is still a qualitative symmetry in the sense that it preserves all quantities that do not mention the labels of specific points. In particular, as f is an isometry, it preserves the metric, the set of open sets and the maximal atlas. And as we have seen in Chapter 5, the preservation of these three quantities simply means isomorphism of $\mathcal{L}_M$ guarantees to preserve all properties that can be expressed in the language of $\mathcal{L}_M$, which indeed does not contain constants for specific points. Thus, one may say that while there is no symmetry in the strictest sense, there are plenty of qualitative symmetries, and it is those qualitative symmetries that matter for physical purposes.

Thirdly, one may want to say that the extension of constants is illegitimate simply because it is *impossible*: we cannot refer to specific spacetime points, in particular, we cannot physically identify space-time points as specific individuals. Arledge and Rynasiewicz (2019) consider the case where we tend to identify space-time points using Gaussian normal coordinates. The basic idea is to fix space-time points by "special" coordinates of a hypersurface plus a "time" coordinate specified by geodesics normal to the hypersurface. They conclude that this is impossible since at the common time-slice Σ of $\mathcal{M}$ and $d^*\mathcal{M}$, the physical operations we perform to trace geodesics into the future will be the same, and hence our operation cannot distinguish between geodesics γ and $d^*\gamma$. This indeterminacy of reference leads to the failure of the identification of space-time points as specific individuals.

This reply cannot succeed either. To begin with, notice that this underdetermination of reference is only a special case of Putnam's paradox of reference (1980), as also acknowledged by Arledge and Rynasiewicz. So formalists can only renounce reference to specific space-time points at the cost of renouncing reference *all together*. This may be too high a cost to pay. Second, this argument is vulnerable to a Lewisian reply (1984), just as Putnam's paradox. Lewis argues forcefully:

> Referring isn't just something we do. What we say and think not only doesn't settle what we refer to; it doesn't even settle the prior question of how it is to be settled what we refer to. Meanings –as the saying goes – just ain't in the head. (p. 226)

So reference is partly fixed by the physical world. Take the example from Putnam himself, the physical operations human did on earth before 17th century to fix the reference of "water" may be the same as what human did on a twin earth, where "water" is secretly replaced by XYZ but not H_2O. But still, the referents of "water" on earth and twin earth are each determined, i.e., as H_2O and XYZ respectively. Equally, despite the physical operations we perform to establish Gaussian normal coordinates will be the same in $\mathcal{M}$ and $d^*\mathcal{M}$, the referents of the coordinates will be determined by $\mathcal{M}$ and $d^*\mathcal{M}$ themselves, i.e., as points traced along geodesics γ and $d^*\gamma$ respectively.

Additionally, the point is not even about whether *our* reference is possible. For indeterminism will follow as long as substantivalists acknowledge that there *exist* two possible physical situations S_1, S_2

with a common time-slice and different futures, whether or not we can identify or specify them by our limited physical capacities. As a metaphor, it suffices for *God* to be able to name specific space-time points, imagine dragging the points along a hole transformation, and decree worlds both before and after the transformation as possible.

We have argued that the above three justifications for rejecting the extension of Lorentzian manifolds are not really successful. The final prospective resort we consider here, which I believe is also implicitly suggested by many formalist writings, is to appeal to a specific form of naturalism, which I call "*Matter-of-Fact Naturalism*". The next section gives an analysis of this response and argues that, at least, the naive version of this principle is hard to defend.

6.6 Matter-of-Fact Naturalism

A naturalist motivation can be identified behind most formalist writings. The general idea seems to be that philosophers should trust our best scientific theory, not only what it says, but also *the language* in which it says it.

> As a rough guide, the language of general relativity allows us to say the sort of things that expert users of general relativity say about the external world – for example, 'there is an inextendible geodesic of finite length' or 'if the mass increases beyond a certain bound, then a singularity will form'..... So, when a philosopher starts talking about spacetime points having different properties in different possible worlds, then they have already gone beyond the language of general relativity. (Halvorson and Manchak, 2022)

Weatherall (2018) also echoes that we should stick to the formalism of GR since "the fact that we use the particular mathematical structures we do is the end result of a long process of developing and interpreting general relativity." (p.345)

Formalists suggest here that we should simply adopt the language of our best scientific theory, and any question that goes beyond it would simply not be worth asking. It seems from the above remarks that formalists assume the following "naive" principle:

> **Matter-of-Fact Naturalism**: If a proposition ϕ about a domain cannot be said in the language of our best scientific theory about that domain, then we should believe that there is simply no matter of fact whether ϕ.

For instance, while followers of Aristotle might debate over whether the earth is at the center of the universe, Lorentzian manifolds simply do not include the structure of "a center", not even anything it could be approximately reduced to. Thus, we should simply believe that there is no matter of fact as to which celestial object is at the center of the universe. Similarly, as Lorentzian manifolds do not include the specific identities of spacetime points, we should simply believe that there is no matter of fact whether it is this or that spacetime point that is so-and-so.

However, this principle is subject to the attack of anti-realist arguments against scientific realism. In particular, Laudan (1981) gives a list of scientific theories that were empirically successful at some point, but was later proved to be radically false, some of which include "the crystalline spheres of ancient and medieval astronomy", "the effluvial theory of static electricity", and " the electromagnetic aether". Laudan argues that this historical list inductively suggests that scientific theories that are empirically successful are prone to radical revisions, and thus, we should not believe in what our current best scientific theory literally says despite their empirical success.

While scientific realism focuses on the contents of our best scientific theories, Matter-of-Fact Naturalism focuses on the language in which our best scientific theories are stated. But a similar historical list can be produced to show that there are many things that the once best scientific theories cannot say in their languages, but were later proved to be physically significant. For instance, in Newtonian mechanics, we simply lack the structure to describe the phenomena of quantum entanglement. People who believe in Matter-of-Fact Naturalism at the age of Newton, would then conclude that a particle always has an independent state and there is no matter of fact whether two particles are entangled. Similar examples are abundant. Indeed, every time when scientists discover a phenomenon that requires essential addition of new structures to the existing theory, the principle of Matter-of-Fact Naturalism is disconfirmed.

One may object that Matter-of-Fact Naturalism should really be interpreted as saying that we should believe in the language of our *current* best scientific theory *at this moment*, which is perfectly compatible with changing our views later as science advances. Not so. In particular, the principle of reflection in formal epistemology states that:

> **Principle of Reflection** If a rational agent believes that their future credence in a proposition will be p, then their current credence in the proposition should also be p.

In formulas, we have $Cred^a_t(A|Cred^a_{t+x}(A) = r) = r$ where $Cred^a_t(A|B)$ denotes the credence of an agent a at time t in the proposition A conditioned on the proposition B. Van Fraassen (1984) shows that a Dutch book can be constructed against any agent who violates the principle of reflection. The principle of reflection then suggests that if we believe that it is very likely that we will not believe in the language of our current best scientific theory in the future, then we should not believe in it *now*. While the principle of reflection is certainly not uncontested, violating the principle will need further justification.[12]

Another possible response is to restrict the Matter-of-Fact Naturalism to the language of our best scientific theory *of all time*. While this makes the principle much plausible, we have difficulties applying this principle to the case of the hole argument, as it is hard to believe that GR will be the best theory of spacetime of all time.

I conclude that it is not plausible in general to hold the principle of Matter-of-Fact Naturalism, i.e., to trust what the language of our best scientific theory cannot say. However, I suggest that in this specific case of the hole argument, we *do* have good reasons to believe a specific thing that the language of GR cannot say, which suffice to refute the hole argument. In the next section, I will argue 1. to refute the hole argument, we only need to buy a specific feature about the formalism of GR which I call "anti-specificism" about spacetime; 2. this specific feature actually subsists in the long history of human investigations about the nature of spacetime, surviving all the way through the radical revisions made by scientists, which gives a *positive* inductive argument for buying this feature, and 3. apart from scientific practices, the metaphysical view brought about by this feature is independently plausible as it avoids many problems faced by contemporary versions of spacetime substantivalism.

[12] See e.g. Talbott (1991) and Christensen (1991) for counterexamples to the principle of reflection. Note, however, such counterexamples of reflection contain essentially unusual cases of forgetting or memory altering. Thus, they do not help directly to justify how the violation of reflection can be justified in the case of daily scientific practices where no such unusual cases are involved.

Chapter 7

Anti-specificism about Spacetime

7.1 Anti-specificism about Spacetime

Looking back at how the hole argument is constructed, we notice that the single most important feature about the formalism of GR that blocks the hole argument is that Lorentzian manifolds do not include *the specific identities of spacetime points*, as can also be seen from the explicit language $\mathcal{L}_M$ of Lorentzian manifolds we construct in Chapter 5. Thus, for $\mathcal{M}$ and $d^*\mathcal{M}$ in the hole argument, while we drag the metric field along the hole diffeomorphism d within the hole, there is no danger of indeterminism, since *there is simply no matter of fact as to which point is which*, or *which point is assigned a specific metric value*, and so there is nothing to be indeterminate about. We may call the view that there is no matter of fact about specific spacetime points *anti-specificism* about spacetime.

Anti-specificism is a view that belongs to the general camp of "sophisticated substantivalism" or "anti-haecceitism". "Sophisticated substantivalism", a label introduced by Belot and Earman (2001), is originally used to denote positions which deny that there can be two possible worlds that differ only with respect to specific identities of spacetime points. And many people follow the original sense of the term (Pooley, 2013, 2021; Teitel, 2019). But sometimes it is also used more broadly to refer to any version of substantivalism that rejects indeterminism as the consequence of the hole argument (Norton, Pooley and Read, 2023). "Anti-haecceitism" is another popular label in the literature that does not have a fixed meaning. The term "haecceitistic" is often used to refer to those facts that involve specific identities of some objects, which are sometimes also called "non-qualitative", in contrast to "qualitative" facts that do not mention specific objects. Thus, the most general anti-haecceitism says that there are simply no two distinct possible worlds that differ only with respect to specific identities of objects, i.e., differ only with respect to haecceitistic facts. A more limited version of haecceitism (SP-haecceitism in Teitel, 2019) restricts the targeted domain of the thesis to only spacetime points, saying that there are no two distinct possible worlds that only differ haecceitistically with respect to spacetime points. Thus construed, (SP-)haecceitism is equivalent to sophisticated substantivalism in its original meaning. A even more specific understanding of anti-haecceitism is to say that "spacetime points do not possess trans-world identities" (Norton, Pooley and Read, 2023). In this section, we will use "sophisticated substantivalism" in its original meaning, and "anti-haecceitism" in the more specific sense of denying the trans-world identity of spacetime points. Adopting this convention, anti-haecceitism becomes a specific type of sophisticated substantivalism: if there is no trans-world identity of spacetime points, then there is no difference such as that p in $\mathcal{M}$ has the metric value $g(p)$ while p in $d^*\mathcal{M}$ has metric value $g(d^{-1}(p))$. A typical example of anti-haecceitism is the counterpart theory advocated by Butterfield (1989), Gomes and Butterfield (2024a, 2024b), and Jacobs (2024), which says that there is no identity but only counterpart relationship between spacetime points across possible worlds. And a typical example of sophisticated substantivalism that is not anti-haecceitism is metric essentialism (Maudlin, 1988, 1990), which says that spacetime points possess their metric properties *essentially*. We will discuss these views in Section 7.3 in detail.

With terminologies defined as such, anti-specificism is a form of anti-haecceitism, as it denies any specific identities of spacetime points *whether cross-world or not*, and *a fortiori* it is a form of sophisticated substantivalism. But it is different from the counterpart theory in that anti-specificism not only rejects haecceitistic facts about spacetime points across possible worlds, but also rejects haecceitistic facts about spacetime points within a single world. That is, there are no haecceitistic facts about spacetime points whatsoever.

This view was previously considered by Hoefer (1996), Caulton and Butterfield (2012) and Russell (2014) under different names.[1][2] There are some noteworthy objections to this view, some of which I shall put in later sections. But it is important to deal with some prevalent concerns which, I doubt, prevent people from considering this view seriously from the very beginning. In particular, one may worry that anti-specificism is really relationalism in disguise.

> I tend to think of metaphysical views that reject commitments to absolute spatial positions of particular things as kinds of relationism, broadly speaking, rather than substantivalism... Whatever label we give them, the rejecters face the same challenge as relationists: to find some alternative empirically adequate theory of space without those commitments. (Russell, 2014, p. 74)

While Russell takes anti-specificism to be what some substantivalists are essentially saying in response to the hole argument, he thinks that this is really a form of relationalism, or at least, will require a reformulation of our theory of spacetime just as relationalism does. And he is puzzled by why people who seem to hold an anti-specificist view about spacetime "have not seemed to appreciate the challenge to produce an alternative theory" (p.75) Hoefer (1996) similarly recognizes the need of reformulation, though he argues that a *re*-interpretation of the formalism of GR in an anti-specificist manner can do the same job as well.

Both authors believe that our physical theories of spacetime under literal interpretation *do* commit to specific identities of spacetime points, and thus anti-specificists need to reformulate or reinterpret the theory to avoid such commitments. As argued in Sections 5.5 and 6.2, this is a mistake. Instead, if we believe that the isomorphism criterion for models of GR in its original form is already isometry, then GR has to be anti-specificist theory from the very beginning. The formulation of GR in $L_{\mathcal{M}}$, therefore, should not be taken as a *re*formulation or *re*interpretation of the original formalism of GR,

[1] Russell suggests that haecceitistic propositions about spacetime points are non-factual. He seems to use the term "factual" in a metaphysically loaded sense, and here I opt for a more neutral terminology by using "matter of fact" instead. There is an intuitive sense in which whether something is a matter of fact is intelligible, though unpacking the meaning under the metaphysical context may require future work. Hoeffer also expresses his view as abandoning "the ascription of primitive identity to spacetime points" (1996, p. 14). Though Teitel (2022) reads Hoefer differently, suggesting that he only rejects modal haecceitistic facts about spacetime points, i.e., haecceitistic facts across possible worlds. Caulton and Butterfield (2012) use the term "structuralism" to label a continuum of views that roughly hold that heuristic facts about spacetime points "are grounded in" qualitative facts. I suggest "the weak end" of the continuum (p. 237) corresponds precisely to anti-specificism considered in this paper. Their discussion also shows that anti-specificsm is not only motivated for spacetime points, but likely for particles in quantum mechanics as well.

[2] One may also relate anti-specificism with "generalism", and in particular, "quantifier generalism" in the discussion of ontic structural realism (Dasgupta, 2014, 2016; Glick, 2020). However, a crucial difference is that "generalism", as a specific type of ontic structural realism, is under the burden of eliminating ontological commitments to individuals, or at least individuals as fundamental objects, in order to say that, as the famous slogan goes, "all there is to the world is structure". In the current context, as a form of (sophisticated) substantivalism, anti-specificism faces no such burden, and makes no bones about its commitment to the existence of spacetime points.

but rather as an *explication* or *rational reconstruction* of GR. That is to say, it presents the full formal picture of GR already assumed in physical practice.

And anti-specificism is not only true of the formalism of GR. As we will see later, as part of a positive induction supporting anti-specificism, that this is so for all historical theories of spacetime, from Aristotle's view of spacetime, Newtonian mechanics, special relativity to general relativity. Thus, among all versions of substantivalism, it is not anti-specificism that calls for a revisionary project, either in formalism or in interpretation, but traditional versions of substantivalism which do grant specific identities to spacetime points.

We will also argue that there are many independently good reasons why we should accept anti-specificism, particularly compared to its peers. But whether or not anti-specificism is tenable, one may worry, just as Russell, that it is really a form of relationalism in the first place, so the hole argument is still successful, as it forces us to give up substantivalism. Here, I shall argue that this is not the case.

There are different characterizations of substantivalism. The universal requirement is the commitment to the *existence* of spacetime points. Notice that anti-specificism does satisfy this requirement. It does commit to the existence of spacetime points, only refraining from endowing them with specific identities.

One may wonder what it means or how it is coherent to say that something exists but has no specific identities.[3] But $\mathcal{L}_M$ is just a rigorous and consistent formulation of this view. In fact, cases of anti-specificism are actually much more common than one might have thought. Consider a first order language $\mathcal{L}_1$ with a binary relational symbol R as its only non-logical symbol, and a $\mathcal{L}_1$ theory T_1 with the following axioms:

- $\exists x \exists y (x \neq y \wedge \forall z (z = x \vee z = y))$ (there are exactly two distinct things)
- $\exists x \exists y (x \neq y \wedge xRy \wedge \forall z \forall w (zRw \rightarrow (z = x \wedge w = y)))$ (there is a relation R holding (only) between the two distinct things, and R is asymmetric)

Consider a further language $\mathcal{L}_2$ which adds to $\mathcal{L}_1$ two constants c_0 and c_1, and a $\mathcal{L}_2$ theory T_2:

- $c_0 \neq c_1$
- $\forall x (x = c_0 \vee x = c_1)$
- $\forall x \forall y (xRy \leftrightarrow (x = c_0 \wedge y = c_1))$

In comparison, both T_1 and T_2 say that there are two things, and there is an asymmetric relation R holding between them. However, while T_2 explicitly states that c_0 points to c_1, T_1 is not specific as to *which point* is pointing and *which point* is being pointed. Now, one could still say that T_2 contains more factual information than T_1. But if we further take what can be expressed in the language of $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively as *all matters of fact there are*, then T_1 does not leave anything out. In fact, it is easy to see that T_1 is complete in $\mathcal{L}_1$, i.e., any sentence that can be expressed in $\mathcal{L}_1$ is either provable or disprovable by T_1. Then, according to T_1 in $\mathcal{L}_1$, there exist two things, but there is simply *no matter of fact* as to which point is which, *period.* In this sense, T_1 is anti-specific about its domain.

[3] Teitel (2021), for instance, confesses that “I have very little handle on what these views are meant to be claiming. Like many, I find the glosses above – involving ideology like what has ‘primitive identity,’ or what ‘can be individuated’ across possibilities — opaque and obscure if not just colorful ways to express some precise first-order modal doctrine like anti-haecceitism or no-shifts.” (p.266)

Thus, any first-order theory with existential quantification over a domain but no constants naming objects within, like T_1, will be anti-specific about its domain. But clearly, there is nothing incoherent or unintelligible about T_1 or any such theory in general. The overall lesson is that the commitment to the existence of something and the commitment to the *specific identity* of something that exists are not quite the same thing. One can commit to that there are such and such things, but not to *which* one is *which.*

Now there is often an additional requirement for a position to be qualified as substantivalism, in extra to the mere commitment to the existence of spacetime points. In particular, substantivalists are believed to hold that spacetime not only exists but exists *" independently of the processes occurring within it"* (Norton, Pooley and Read, 2023). That is to say, there is no way in which spacetime can be reduced to relations or properties of material things. There are different senses in which we can understand the notion of "reduction", but I shall argue that, under all interpretations of "reduction", anti-specificism meets this requirement.

The first sort of reduction is *metaphysical* reduction, which is sometimes further unpacked as the relation of grounding (Dasgupta, 2011) or ontological dependence (Cameron, 2023). The general idea is that if A is metaphysically reduced to B, then A is nothing over and above B in the sense that once God has created B, A is automatically created. For instance, the singleton set of Socrates (i.e. $\{Socrates\}$) can be metaphysically reduced to Socrates himself. Or one may hold that the product of domain A and domain B can be metaphysically reduced to A or B. It should be clear that for anti-specificism, spacetime cannot be metaphysically reduced to sets of relations or properties. In particular, anti-specificism allows for non-trivial symmetries of spacetime which relate one spacetime point to *a distinct yet indistinguishable* spacetime point. As such pairs of points share the same metrical properties but are still distinct points, there is no way to reduce spacetime points to merely metrical relations or properties.

The second sort of reduction is *semantic* reduction. To say that T_1 is semantically reduced to T_2 means that there is a translation from theory T_1 to T_2 that preserves meaning. For example, by substituting "bachelor" with "unmarried man" one can semantically reduce a theory of bachelor to a theory without the term "bachelor". It should also be clear that semantic reduction is impossible without metaphysical reduction. Particularly, anti-specificism will still contain (purely) existential quantification of spacetime points. For instance, "there are two distinct spacetime points", or in formula, $\exists x^p \exists y^p (x^p \neq y^p)$ (assuming p as the sort of spacetime points), cannot be semantically reduced, unless we find a way to metaphysically reduce the domain of spacetime points to another sort of things, say, sets of metrical properties or relations. But we have just argued that such a metaphysical reduction is not possible.

The final sort of reduction is *epistemic* reduction, the concept which features in Oppenheim and Putnam (1958). Saying that T_1 is epistemically reduced to T_2 means that the epistemic role of T_1, say, explaining certain phenomena, can be played equally by T_2. Now, there are indeed some alternative formulations of GR, such as Leibniz algebra[4], supported particularly by relationalists (e.g., Earman, 1989). The basic idea of Leibniz algebra is that, instead of looking at a concrete manifold M, we look at the *ring* of continuous or smooth real-valued functions defined over M (in notation, $C(M)$ and $C^\infty(M)$). Smooth tangent vector fields are then defined as derivations on $C^\infty(M)$, which form a module $D(M)$ over the ring $C^\infty(M)$ that satisfies the Leibniz identity. Further constructions needed

[4] See Geroch (1972) for technical details.

to write down field equations are then defined in terms of $D(M)$. A Leibniz algebra or Leibniz model is thus simply a tuple consisting of a ring R, a subring R^{∞} (both contain the real field which corresponds to the set of constant functions over M) and other objects of algebraic types that correspond to different fields presented in the standard formalism of GR, e.g., a Lorentzian metric field.

Thus, one may argue that the essential role the manifold M plays in GR is only to fix the algebraic structure of $C(M)$ and $C^{\infty}(M)$, by which we then write down our field equations. By substituting the Lorentzian manifold $\mathcal{M}$ with the Leibniz algebra that $\mathcal{M}$ realizes, we will lose nothing in terms of explanatory or predictive power of our theory.

First of all, we notice that even if this argument were valid, this should be a challenge to all sorts of views that purportedly claim to be substantivalist, not just anti-specificism, and thus it does not disfavor anti-specificism against its substantivalist peers, e.g., metrical essentialism or counterpart theorists. And we do have reasons to believe that it is *in*valid. In particular, there are some doubts whether Leibniz algebra really eliminates commitments to spacetime points. Rynasiewicz (1992), for instance, submits that a Leibniz algebra is still a substantivalist model in disguise. For one thing, there is a natural process by which one can easily reconstruct the manifold structure given a Leibniz algebra.[5] And secondly, each isometry of a Lorentzian manifold corresponds to a unique homomorphism of the base ring in its Leibniz algebra through pre- and post-composition of the representation maps. So the problem of the hole argument *reappeared* using the formalism of Leibniz algebra.

Finally, we think epistemic irreducibility is not central to how most people view the thesis of substantivalism, as essentially a thesis concerning the fundamental structure of the world. It is surely possible to have metaphysical structures or entities which do not play any specific role in explaining certain phenomena, but adding them in our metaphysical pictures could nevertheless cater to certain prevalent metaphysical principles that win favor among philosophers or add coherency or elegance to our overall theory. In such cases, we do have reasons to add entities or structures that are epistemically reducible. The general lesson is simply that explanatory power is not the only reason by which one can justify a theory over another. Even if spacetime points are epistemically reducible, it does not mean that they do not exist or have inferior metaphysical status.

In summary, we have argued that anti-specificism is a clear and coherent form of substantivalism which is entailed by the formalism of GR. The following sections give in detail the reasons for accepting this view.

7.2 An Optimistic Induction

Laudan's pessimistic induction argues that we should not believe what our best scientific theories say, since they are prone to radical revisions. Indeed, it is perfectly conceivable that physicists update the theory of GR to a radically different theory, just as we updated Newtonian mechanics to Special Relativity and then to General Relativity in the early 20th century. However, in this section, I shall argue that the specific feature of anti-specificism possessed by the formalism of GR is a common feature in all different historical theories of spacetime despite radical changes. This then gives us an optimistic induction for anti-specificism.

[5] More specifically, the set of real maximal ideals of the ring R constitutes a smooth manifold which realizes the structure of the Leibniz algebra.

7.2.1 Spacetimes in History

We argued in Section 6.2 that anti-specificism holds if we stick to Lorentzian manifolds as the formalism of GR, and in Section 6.4 that it is indeed the formalism which physicists actually used in practice. Thus, we have come to the conclusion that anti-specificism holds for the formalism of GR.

The spacetime for special relativity, i.e., Minkowski spacetime, is really a special case when our Lorentzian manifold happens to be a Minkowski space, i.e., a four-dimensional pseudo-Euclidean manifold with a Minkowski metric. The relevant criterion of isomorphism for Minkowski space is again isometry. The only difference is that, due to specific conditions we impose for a Lorentzian manifold to be a Minkowski space, such isometry maps form a Poincaré group. But still, there is no requirement that corresponds to the preservation of labels of spacetime points, and thus anti-specificism holds as well.

While the above conclusion may not be surprising, we are going to see that anti-specificism is true for all kinds of *classical* spacetime structures as well. We give a quick review of such structures below. Though the labeling by the names of physicists may not be fully faithful to their views, we choose to follow the convention in coherence with the literature (Friedman, 1983; Earman, 1989).

The sparsest spacetime structure is presumably what Earman (1989) calls "Machian spacetime" which arguably demonstrates Mach's view about spacetime. There is some vagueness as to what structure "the Machian spacetime" amounts to. The general idea is that we have only an absolute notion of simultaneity, and a family of three-dimensional Euclidean metric spaces as instantaneous spaces. If we follow this idea literally, then spacetime is simply represented by a family of three-dimensional Euclidean spaces. The structure of time would only be literally a set of isolated points, and there is no connection between two points lying on distinct instantaneous spaces. We may call this *Machian-sparse spacetime*. One may not be satisfied and instead think that these different instantaneous spaces are actually not isolated, but "glued together smoothly". In this case, we may require the whole spacetime to be a four-dimensional manifold, with a smooth foliation into a family of three-dimensional Euclidean spaces. The leaf space then represents the structure of time. We may require the foliation to be simple, so that the leaf space also possesses a smooth manifold structure. At this moment, time is "glued together", but still not ordered. If one believes that time is indeed linearly ordered, then we may further require the leaf space to be linearly orderable. One may even want to say that such an order must be a linear continuum, i.e., a linear order which satisfies the following two intuitive conditions.

- Density: for any two distinct points a and b, there is a point c such that $a < c < b$.
- Least upper bound property: for any subset S of the space, if S is bounded above, then there is a least upper bound of S.

which would say that time is dense and there are no "gaps" in time. One can prove that this is equivalent to requiring that the orderable leaf space be connected. Now our leaf space is a one-dimensional, connected (Hausdorff and second-countable) manifold, which then can be proved to be diffeomorphic to either (intervals or rays of) the real line $\mathbf{R}$, or the circle S^1 (Hirsch, 1976). Since the leaf space is orderable, it cannot be diffeomorphic to S^1, and if we add the further conditions that the linear continuum compatible with the leaf space has no maximum nor minimum, which intuitively says that time has no starting nor ending point, then the leaf space will be diffeomorphic to the real line. This final structure, i.e., a four-diemensional spacetime simply foliated into three-dimensional

Euclidean spaces with the leaf space orderable by a linear continuum, seems to be exactly what Earman has in mind when he imposes the set of time symmetries for Machian spacetime as $t' = f(t)$ where $\frac{df}{dt} > 0$.

We may call this Machian-rich spacetime. Though, as we have seen, this structure commits only the mere notion of absolute simultaneity, but also that time is smoothly glued together with space, and ordered in a particular manner. We also notice that since our leaves are all R^3, by Corollary 31 in Meigniez (2002), the simple foliation forms a fiber bundle. And since the base space is contractible, it is globally trivial. Therefore, in fact, this Machian-rich spacetime must be diffeomorphic to R^4. On the other hand, by positing our spacetime as R^4, we precisely commit to the structure of Machian-rich spacetime.

While Machian spacetime does not admit the comparison of "time interval", Leibnizian spacetime enables us to do this by adding a time metric to the above structure. Thus, one may take Leibniz spacetime to be what we obtain by assigning a real number to each of the leaves in a Machian-rich spacetime in a way compatible with the linear order. This method assigns a fixed real number as the time of an instantaneous space, but what actually has physical significance is arguably only the time interval between two instantaneous spaces. So a more popular approach (Friedman, 1983) is to add a smooth non-vanishing covector field dt which is exact, i.e., it is indeed the exterior derivative of some global function t intuitively interpreted as the global time function, which justifies the notation. In this way, dt only determines t up to a constant, and hence gives no significance to the absolute value of real a point receives. Note that since R^4 is simply connected, it suffices to require that dt is closed, from which exactness follows from the Poincaré lemma. At this point, we may also summarize Euclidean metrics for instantaneous spaces into a (2,0)-tensor field h with signature $(1, 1, 1, 0)$ which is symmetric, and compatible with dt in the sense that $h^{ab}dt_a = 0$. The family of instantaneous spaces is then defined as the integral surfaces of dt, which, together with the three-dimensional metric fields induced by h, is required to be three-dimensional Euclidean metric spaces.

Galilean spacetime adds to Leibnizian spacetime a flat affine connection D which is compatible with h and dt in the sense that $D_a h^{bc} = 0$ and $D_a dt_b = 0$. The fixed connection D allows us to introduce the notion of parallel transport and consequently *the absolute notion of acceleration.* If we take the normed tangent field V^a as the velocity field of a particle, then the acceleration of the particle can be evaluated as the deviation of the parallel transport of the original velocity vector along the geodesic, or in notation, $a^a = V^b D_b V^a$.

Newtonian spacetime assumes the further notion of absolute rest. This is done by picking a smooth vector field A to be the state of rest. We require A to be compatible with previous structures we impose by setting $D_a A^b = 0$, $dt_a A^a = 1$. Then, the particles whose worldlines coincide with the integral curves of A are said to be at absolute rest, and the absolute spatial velocity W of a particle is evaluated by subtracting A from the normed velocity field V, in formula, $W^a = V^a - A^a$.

Now it is not hard to see that all classical spacetime structures introduced above respect anti-specificism. Machian-sparse spacetime is only a family of isolated three-dimensional Euclidean spaces. Euclidean spaces are affine spaces, which give no absolute coordinates to points. So for instance, there is no matter of fact as to which point is the original point. Instead, an affine space is simply a set of anonymous points on which a vector space (in this case, Euclidean vector spaces) acts, and there is no specification as to which point is linked to which by which vector, as every vector equally acts on every point. Thus, in a Marchian-lite spacetime, there is simply no matter of fact about a particular point, but only quantificational facts, e.g. the fact that, for any point x, for any vector v, here is some instantaneous point y linked to x by v, where x, y, v are all general variables ranging over the whole domains, but not names for specific points.

Machian-heavy spacetime admits a more substantival structure of time, resulting in essentially a manifold diffeomorphic to R^4. But again, we only move from a family of three-dimensional Euclidean spaces to a four-dimensional Euclidean space, which still keeps the points within anonymous. Leibniz or Galilean space is anti-specificist in a similar manner, as adding a time metric or an affine connection does not add any information about specific points. But even *Newtonian* spacetime is anti-specificist. With Newtonian spacetime, we are indeed able to identify points in different instantaneous spaces by the congruence of integral curves of A. However, there is no specification as to *which integral curve is which.* Indeed, any symmetry of a certain instantaneous space that permutes space points will generate a global symmetry of our spacetime that permutes integral curves passing through the corresponding space points.

Therefore, to break anti-specificism, we need to commit to a spacetime structure even richer than Newtonian spacetime. *Aristotelian spacetime*, e.g., is constructed by adding "the center of the universe" to Newtonian spacetime, which threatens to refute anti-specificism. Formally, we pick a particular integral curve of A to specify the worldline of this center. Now, there could be different interpretations of Aristotelian center of the universe. It could be interpreted as simply a property of spacetime points, just as other metrical properties one assigns to spacetime points in GR. In this case, anti-specificism still holds, as we only commit to a property that happens to be satisfied by a single point in each situation, and we may well insist that there is no matter of fact as to whether it is the same point or not that satisfies this property in different situations. This is just like some metrical properties which can only be satisfied by a single point in each situation, and anti-specificists insist that there is no matter of fact as to *which* point satisfies them. However, if we understand the Aristotelian center as *the very essence* of some particular point which acts as the center of the universe in a metaphysically necessary manner, then *anti-specificism* is indeed broken, and now we have at least one special point in our theory, i.e., *the* center of the universe.

But there is still a long way to go from Aristotelian spacetime to full *specificism*, i.e., the case where the specific identity of *every single point* is added to our structure. A full specificism would require assigning specific identities to each spacetime point, which could be done by adding an extra *ID field* on the manifold whose value at each point in the manifold is the identity or name of the spacetime point to be represented. But no such field has ever been proposed in the literature of spacetime. The only example that presumably validates specificism is the example of *colour* discussed by Riemann himself when the notion of manifold was introduced (Riemann, 2004, p.258). Points in the manifold of colour represent specific colours, and are mutually connected according to their RGB values. Thus, one may take the manifold simply as a closed cube in R^3. We may then add the ID field which assigns specific RGB values to each point in the manifold. Note that what is special about this case is that intuitively there is something intrinsic to specific colours that is not captured by the pure manifold structure. For instance, while permutation of the base R with the base G preserves the manifold structure, it maps red to green, which intuitively are different by their intrinsic nature. It is for this reason that an ID field in the manifold of colour is well-motivated and even necessary. In contrast, there is little motivation for positing specific identities, essence, or intrinsic nature for spacetime points in addition to the metrical properties they instantiate in the manifold. Indeed, as we have seen, none of the classical spacetime structures posit such an ID field for spacetime. And except for Aristotelian spacetime, all classical spacetime structures commit to anti-specificism to a full extent.

7.2.2 Physicists' Views in History

One might worry that the formalisms above are really *re*formulations of classical spacetimes in the hindsight of spacetime in General Relativity as exactly a four-dimensional metric manifold and may not faithfully reflect the physicists' view at that time. However, while physicists like Leibniz or Newton surely do not have the formal notion of manifold in their mind, I shall argue that their writings indeed commit to anti-specificism, just as illustrated by the above formalisms.

Due to the limited space, I will focus mainly on the view of Newton and give only a brief overview of the positions held by other physicists. Indeed, people such as Leibniz or Mach are both famous relationalists in history, and it should not be surprising at all that they are anti-specificist about spacetime points. Leibniz, for instance, raises the *shift* argument against substantivalism of Newtonian spacetime in Leibniz-Clark correspondence:

> Space is something absolutely uniform; and, without the things placed in it, one point of space does not absolutely differ in any respect whatsoever from another point of space. Now from hence it follows ... that 'tis impossible there should be a reason, why God, preserving the same situations of bodies among themselves, should have placed them in space after one certain particular manner, and not otherwise; why every thing was not placed the quite contrary way, for instance, by changing East into West. (Leibniz and Clarke, 1956, p.26)

Leibniz's argument, whether correct or not, clearly shows his commitment to anti-specificism: for if there are indeed specific identities for spacetime points, then precisely by the above argument, we can generate continuum many possible worlds by shifting all the matters in the universe some inches away from their current positions, which would be objectionable to Leibniz. Galileo (1967) and Mach (1893) also show a similar commitment to relationalism and consequently to anti-specificism.[6]

What is more interesting is the view of Newton, whose view of spacetime has been taken as an archetypical form of full-blooded anti-specificism. But in fact, despite the prevalent impression, scholars have emphasized that Newton never treats space or time as a full-fledged substance (Stein, 2002; Hoefer, Huggett, and Read, 2021). The following paragraphs of exegesis tend to show further that Newton's view of spacetime is indeed anti-specificist.

Newton summarizes his view about spacetime mostly in the *Scholium* attached to the *Principia*, and also in an unpublished manuscript *De Gravitatione*. The notion of absolute space is introduced in *Scholium* for which Newton gives six arguments. While the detailed reasoning of each argument varies, the basic idea is the same, i.e., absolute space is necessary in order to define the *true* motion of a body. As argued by Rynasiewicz (2019), Newton really takes for granted the common assumption, also held by Aristotelian and Cartesian philosophers, that "each body has a unique state of true motion (or rest)", and absolute space is needed to make sense of this notion.

[6] Mach (1893) stands for relationalism by arguing that all inertial effects are to be explained by the relative motion of a body with respect to other massive bodies in the universe, but not with respect to the absolute space. Galilean does not address the issue of substanvalism vs. relationalism directly, but is famous for raising the principle of what we now call *Galilean Relativity* (Galileo, 1967), which says that experimenters will observe the very same phenomena whether they are at rest or moving uniformly with a certain velocity. While indistinguishability does not logically entail indifference, it does suggest the redundancy of the absolute space as posited by Newton.

However, as we have also seen in the formalism of Newtonian spacetime presented above, to have the notion of absolute motion or rest, it suffices to posit the absolute space in an *anti-specificist* manner. Specifically, the definition of absolute velocity, i.e., $W^a = V^a - A^a$, does not refer to any *specific* space point. Instead, it is defined without committing to any specific fact about which space point is which, left fully open to all symmetries of the three-dimensional Euclidean space. Thus, Newton's reasons for positing an absolute space really only support an *anti-specificist* form of spacetime, and position of specific identities of spacetime points remains unjustified.

In fact, Newton himself explicitly withdraws from viewing space as "full" substance. In *De Gravitatione*, Newton considers three possible options to classify space, space as substance, as accident (which is essentially another name for properties) or as nothing. Newton argues that they all fail:

> Perhaps now it may be expected that I should define extension as substance, or accident, or else nothing at all. But by no means, for it has its own manner of existing which is proper to it and which fits neither substances nor accidents ... And much less may it be said to be nothing, since it is something more than an accident, and approaches more nearly to the nature of substance. (pp. 21-2, 2004)

Thus, Newton believes that 1. in contrast with relationalism, space exists, as it is not nothing; 2. in contrast with metrical essentialism, space cannot be conceived as simply certain sets of properties; 3. space is not substance neither, though it "approaches more nearly to the nature of substance".

"Substance" as used by Newton has a very peculiar meaning in this context, and does not correspond to the modern thesis of substantivalism, nor directly to the Aristotelian notion of substance. In fact, the two reasons Newton gives for rejecting space as substance are the following.

> On the one hand, because it is not absolute in itself, but is as it were an emanative effect of God and an affection of every kind of being; on the other hand, because it is not among the proper affections that denote substance, namely actions, such as thoughts in the mind and motions in body. (p.21, 2004)

So Newton rejects the view of space as a substance simply because 1. space points cannot act, and 2. they are not "absolute in themselves" but "an emanative effect of God and an affection of every kind of being". Although the first point is easy to understand, the second one is more elusive. Stein (2002) carefully analyzes the meaning of this term, and argues that, despite its theological connotation, the term is really used to mean that space is a *necessary consequence* of positing anything as existing. Thus, in Newton's view, space really exists not as the "characters" of our universe which act (what Newton calls "substance"), but as "backgrounds" or "props" which have to be there in order for any character to be able to act.

To argue for the immobility of space, Newton further explicitly advocates the view that both space and time can only be individuated by their mutual position or order:

> Moreover, the immobility of space will be best exemplified by duration. For just as the parts of duration are individuated by their order, so that (for example) if yesterday could change places with today and become the latter of the two, it would lose its individuality and would no longer be yesterday, but today; so the parts of space are individuated by their positions, so that if any two could change their positions, they would change their individuality at the same time and each would be converted numerically into the other.

> The parts of duration and space are understood to be the same as they really are only because of their mutual order and position; nor do they have any principle of individuation apart from that order and position, which consequently cannot be altered. (2004, p.25)

The most straightforward reading of the above paragraph is a naive form of essentialism which says that the essence of space (or time) points consists merely in their mutual positions (order), and any two points which instantiate the same set of positional properties will be identified as the same.[7] The problem with this reading is that such naive essentialism is blatantly incoherent with Newtonian spacetime. In fact, any two points in the three-dimensional Euclidean space share the same set of positional properties, and yet cannot be identified as one. Here I suggest a more coherent reading is to take the anti-specificist perspective, and read "individuality" of a point above as the collection of *all matters of fact there are about a point.* To individuate two points is simply to find a proposition holds of one point but not of another. Thus, the above-quoted paragraph suggests the following view:

> All matters of fact there are about a space (or time) point are exhausted by the positional (or sequential) properties instantiated by the point;

which is simply a specific version of anti-specificism about Newtonian spacetime. Note that this view avoids the problem of naive essentialism as it only claims that, e.g., points in the three-dimensional Euclidean space are *indistinguishable* with respect to all matters of fact, *but not* that they are to be identified as the same point. It also makes sense of several remarks made by Newton. For instance, the interchange of positions would indeed lead to the interchange of individuality, as the previous facts about one point now hold for the other point.

In sum, we have seen that not only the formalism of GR, but also the formalism of historical theories of spacetime, including many physicists who invent and interpret these theories, advocates anti-specificism. This gives us an optimistic induction to think that, just as it has survived through the radical revisions of our past spacetime theory, it will likely survive through the future as well, and we have good reasons to expect that anti-specificism will be part of the common core of spacetime theories, if they converge at all. The next section gives a more detailed exposition on this argument, particularly in comparison with Laudan's pessimistic induction.

7.2.3 Optimism vs. Pessimism

This section gives a detailed construction of the optimistic induction for anti-specificism, in comparison with Laudan's pessimistic induction. We show that they are not only compatible, but the pessimistic induction actually strengthens the optimistic induction.

Laudan's pessimistic induction has the following form:

- Inductive Premises: Theories T_1,...,T_n were empirically successful but turned out to be false, and not even approximately true.
- Inductive conclusion: The current best scientific theories, despite being empirically successful, are likely to be false and not even approximately true as well.

Thus, Laudan concludes that we should not believe what our best scientific theories literally say.

[7] Note also that it is not clear whether this view really counts as substantivalism, as essences of space points can be reduced to merely mutual positional relationship.

A crucial feature to note about pessimistic induction is that the notion of "empirical success" adopted in the inductive premises is really a *relative* one. To say, e.g., that Newtonian mechanics *is* empirically successful is simply false, as it is unable to explain lots of empirical phenomena such as the photoelectric effect (to be explained by quantum mechanics), or time dilation (to be explained by special relativity). Rather, what is true is that Newtonian mechanics *was* empirically successful during the age of 17-19th centuries, but it is not anymore.

Note that this is to say that the above notion of "empirical success" is a notion relative to time. Rather, it is relative to *the range of experience*, which happens to be available to human society during a specific period of time. Thus, for Newtonian Mechanics to be empirically successful, it only needs to explain the portion of phenomena that were available to the human society during 17-19th century, and phenomena such as the photoelectric effect or time dilation were simply not among them.

Once we recognize the relative nature of "empirical success", we can see that the inductive conclusion of pessimistic induction is not as strong as it seems. It really says that the scientific theories that best explain the range of phenomena that *are currently available to us* are likely not even approximately true. But this should not come as a surprise, since the range of experience that is currently available to us is really very limited, and it will increase predictably and vastly once scientists are able to make larger or more precise experimental instruments. As a folklore, it is said that if we are able to build a particle accelerator with the size of the solar system, we will be able to solve many open questions in high-energy physics. The real lesson we learn from the pessimistic induction is perhaps that as the range of experience increases, the scientific theories that best explain it tend to change radically. But again, it is not clear whether that comes as a surprise, since we always take into account simplicity or elegance of a theory when we talk about best explanation, and it seems natural to expect that the simplest theory that explains a certain range of experience could be very different from the simplest theory that explains a larger range of experience.

In any case, it is important to recognize that the conclusion of pessimistic induction only says that *relative* empirical success (i.e., empirical success with respect to a certain range of experience) does not guarantee truth. And this is natural to expect since a relatively successful theory could fail very badly at explaining a larger range of experience. But then the induction is compatible with the following thesis:

> **Absolute Scientific Realism** (**ASR**): If a theory is *absolutely* empirically successful (i.e., empirically successful with respect to the *maximal* range of potential experience), then it is likely to be true.

A no-miracle argument for **ASR** can be given: it would be a miracle if a theory is able to explain *all potential phenomena*, and yet is false.

Note that **ASR** does not entail scientific realism in the normal sense, sometimes entitled "convergent scientific realism", which says the best scientific theories of different ages will be closer and closer to the final truth.[8] Indeed, one may hold that while all intermediate products of scientific practices are radically false, the final scientific theory, which gives the best explanation for all potential phenomena, is still likely to be true and thus **ASR** shall hold.

ASR then settles a ground for the optimistic induction for anti-specificism sketched above. Since anti-specificism holds for all past spacetime theories, we may inductively conclude that it is likely to

[8] See Popper (1972) for a classical account. See Psillos (1994, 1999, 2009) for a defense of convergent realism by the *divide et impera* strategy, and see Lyons (2006), Cordero (2011) for critique.

hold still for spacetime theories that accommodate a larger and larger range of experience, and in particular, for the theory that accommodates all potential phenomena, which by **ASR**, is likely to be true. Thus, the optimistic induction for anti-specificism assumes only **ASR**, and is compatible with the conclusion of pessimistic induction.

In fact, the lesson we learn from the pessimistic induction can be used to *strengthen* the optimistic induction in the following manner.

- It is very likely that when we update to a larger range of experience, the scientific theory that best explains our experience will be radically different from the previous one.
- However, anti-specificism is a feature that survives through all the updates of experience and radical changes of spacetime theories until now.
- The simplest explanation for this is that anti-specificism is simply a part of the scientific theory that gives the best explanation for all potential phenomena, and thus it is able to survive through previous updates of experience.

Pessimistic induction tells us that scientific theories in general do not converge to the final truth steadily but tend to sway drastically all along the way. But precisely because of this, the fact that anti-speicficism as a feature of our spacetime theories remains stable throughout the updates of experience turns out to be a more significant piece of evidence for its truth. Thus, the optimistic induction we gave is not only compatible with the conclusion of pessimistic induction, but also strengthened by it.

7.3 Anti-speicficism and Its Peers

Apart from the above optimistic induction, I argue that anti-specificism is independently more favourable than its peers from metaphysical considerations. We first introduce popular versions of substantivalism about spacetime in literature, and show why anti-specificism is able to avoid the challenges that plague its peers.

Metric essentialism (Maudlin, 1988, 1990) is a form of sophisticated substantivalism which says that spacetime points possess their metric properties as their *essence*. There are different ways of interpreting this thesis. We may call the position which takes the “metric properties” above as merely qualitative properties as *qualitative* essentialism (Teitel, 2019), and non-qualitative essentialism otherwise. It is also helpful to set the distinction between *strict* and *loose* essentialism: the former believes that there cannot be two objects sharing the same essence, while the latter does not impose such a restriction. Thus, we have four different versions of metric essentialism, and I will argue that all of them have serious difficulties.

Non-qualitative essentialism, whether loose or strict, faces the problem of *cheap determinism* (Teitel, 2022). For if one includes qualitative *and non-qualitative* metric properties a spacetime point satisfies in its essence, then determinism holds *trivially*. For now if any two spacetimes are identical up to a certain moment, then they must be the same about the future, since all facts about the future are already included in the essences of the spacetime points lying in the past. But just as Earman and Norton suggest that determinism should not fail for purely metaphysical reasons, one may similarly believe that determinism also *should not hold* for purely metaphysical reasons. But as we have just seen, Non-qualitative essentialism makes indeterminism simply *metaphysically impossible*

Quanlitative loose essentialism is too weak to block the hole argument, as we can have two spacetimes identical up to a certain moment but differ in the future haecceitistically, i.e., some set of qualitative metric properties is instantiated in the future by one point in a spacetime, but by another point in another spacetime. Qualitative strict essentialism, on the other hand, is incoherent at face value. In any spacetime which allows symmetries, there will be two spacetime points that share the same qualitative metric properties, and yet are not identical, which violates the strictness requirement. One may try to restrict the strictness requirement as applicable only across different spacetimes, i.e., that there cannot be two *distinct* spacetimes where two points share the same qualitative metric properties. It is not entirely clear whether this restriction is backed up by a reasonable metaphysical notion of essence that explains why objects in the same possible world may well share essences but objects across possible worlds cannot. Worse still, it also faces the challenge of cheap determinism: If two distinct spacetimes are identical to a certain moment, then they must be qualitatively the same, as all qualitative properties about the future are already included in the essences of the spacetime points lying in the past. But they also cannot differ haecceitistically, otherwise there will be two distinct spacetimes where the same qualitative properties are instantiated by different points which violates the (restricted) strictness requirement.

Another popular choice of sophisticated substantivalism is *counterpart theory* (Butterfield, 1989; Gomes and Butterfield 2023a, 2023b; Jacobs, 2024). While metric essentialism may not belong to the camp of anti-haecceitism, as spacetime points can still be identical or distinct across different possible worlds, the counterpart theory insists that objects are world-bound, and simply rejects the idea of cross-spacetime identities. Thus, it is anti-haecceitistic in our sense. Instead of identity or distinctness, counterpart theorists suggest that objects in different possible worlds should be compared by the “counterpart” relationship. One object could have zero, one or many counterparts in another possible world, depending on whether there exists zero, one or many objects in that exotic world that are *similar* enough to the original object. Since we are now comparing spacetime points using counterpart relationship, $\mathcal{M}$ and $d^*\mathcal{M}$ will not falsify determinism once we pick d as the counterpart relation which we use to compare the two worlds.

There have been many objections to the counterpart theory in general, all of which would apply to this specific counterpart version of spacetime substantivalism. For instance, the first-order quantificational logic brought about by the counterpart theory has many unwelcome results (Lewis, 1968; Hall, Rabern and Schwarz, 2024). For instance, the intuitive principle of necessity of identity and necessity of distinctness are invalidated (where $\Box$ is the necessity operator):

- $\not\vdash \Box\forall x\forall y(x = y \to \Box(x = y))$;
- $\not\vdash \Box\forall x\forall y(x \neq y \to \Box(x \neq y))$.

On the other hand, many counter-intuitive principles, such as Necessity of Existence and the Converse Barcan Formula, turn out to be valid:

- $\forall y\Box\exists x(x = y)$;
- $\forall x\Box\phi \to \Box\forall x\phi$.

General worries aside, I believe that the specific application of the counterpart theory as a response to the hole argument also faces difficulties. In particular, while it may avoid the metaphysical version of the hole argument by denying cross-world identities, it does not seem to avoid the *epistemic* version.

While metaphysical determinism says that there cannot be two possible worlds that are identical up to a certain moment but differ in the future, the epistemic version says that for an agent with idealized reasoning ability, given all the information about a world up to a certain moment, she will be in a position to know any future facts about the world. We can give an epistemic version of the hole argument, using $\mathcal{M}$ and $d^*\mathcal{M}$ as representing, not metaphysical possible worlds, but *epistemic possibilities* concerning the current *actual world*: for an idealized agent believing in GR, she may know all the information about the current actual world up to a certain moment, say, a moment before which $\mathcal{M}$ and $d^*\mathcal{M}$ still agree, but she cannot know whether the future will be described by $\mathcal{M}$ or $d^*\mathcal{M}$, and thus epistemic determinism fails. Now, just as Earman and Norton argue that metaphysical determinism should not fail for purely metaphysical reasons, one may similarly argue that whether epistemic determinism holds or not should be determined *a posteriori*.

Counterpart theorists cannot answer the epistemic version of the hole argument. Specifically, while one may deny that there are cross-world identities of spacetime points, the epistemic hole argument concerns only one world, i.e., the actual world. That is, $\mathcal{M}$ and $d^*\mathcal{M}$ are not taken to be representations of two possible worlds, but representations of different epistemic possibilities, i.e., conceivable situations in one's mind, of the actual world. And the counterpart theorists do admit that we can talk about identity or distinctness of spacetime points *within one world*. With the standard notion of identity and distinctness available, it seems unjustified to choose instead the alternative counterpart relationship and defend epistemic determinism in terms of it. In other words, it would be unmotivated for counterpart theorists to argue that $\mathcal{M}$ and $d^*\mathcal{M}$, when both are taken to describe the same actual world, should be compared using the counterpart relationship d instead of the standard identity map.

We have seen that both metric essentialism and counterpart theorists have their own problems. But anti-specificism avoids such a problem quite easily. Anti-specificism is free from the challenge of cheap determinism: things about specific identities are not regarded as matters of fact. Thus, being identical up to a certain moment does not require the two worlds to align in the specific identities of spacetime points before that moment, which, according to essentialists, could already entail qualitative or non-qualitative facts about the future that lead to cheap determinism. Thus, whether determinism holds or not is still left completely open by the field equation of our spacetime theory. Additionally, anti-specificism has no problem dealing with the epistemic hole argument. For whether $\mathcal{M}$ and $d^*\mathcal{M}$ are taken to be describing two possible worlds or two epistemic possibilities concerning the same actual world, anti-specificists always hold that there is no matter of fact that involves specific identities of spacetime points, and thus $\mathcal{M}$ and $d^*\mathcal{M}$ will always be taken to have the same representational content.

7.4 Conclusion

We conclude that anti-specificism provides a form of (sophisticated) substantivalism which blocks the hole argument. We have independently good reasons to accept anti-specificism since it is 1. supported by the formalism of GR, 2. supported by the formalism of previous spacetime theories, and 3. is shown to avoid certain challenges (e.g., cheap determinism) that plague other popular versions of sophisticated substantivalism, such as metric essentialism and counterpart theory.

Chapter 8

Conclusion

In this thesis, I have explored the concept of theoretical equivalence from the perspective of representational equivalence. The general lesson is that the interaction between the formal perspectives and the representational or interpretational perspectives of scientific theories proves to be fruitful. On the one hand, by adopting the perspective of representation, we are able to better understand the philosophical significance behind all kinds of formal criteria and extract significant philosophical consequences from formal theorems. On the other hand, the construction of a formal framework of representation provides us with a rigorous tool to clarify and disentangle ambiguous philosophical notions, such as representational contents or representational capacities, and to evaluate speculative philosophical principles and claims, such as the Bradley-Weatherall principle and Leibniz equivalence.

It is undeniable that the discussions presented in this thesis are subject to many limitations. In particular, the following questions are left for future work.

- Chapter 4 discusses how the formal results suggest a pluralist position on theoretical equivalence. The discussion is relatively brief due to space limitations, and more detailed arguments may be constructed in support of this view.

- The notion of definition, translation or interpretation explored in Chapter 2 focuses only on non-logical symbols. According to logical anti-exceptionalism[1], logic is continuous with the sciences and logical symbols do not have a special status in scientific theories. Thus, generalizing current work to include logical symbols would be a natural next step.

- This thesis takes the hole argument as its main example. There are many other discussions in philosophy of physics that concern whether two theories are equivalent or not, e.g., the discussion about the equivalence of Hamiltonian and Lagrangian formulations of classical mechanics, the discussion about the equivalence of different formulations of quantum mechanics, etc. Further applications to these cases may help clarify the discussion and potentially provide new insights.

[1] See Williamson (2013, 2017), Priest (2014), Martin and Hjortland (2021) for representatives of this view.

Appendix A

Non-disjoint Languages

The reason why people restrict themselves to disjoint signatures is clear. If T_1 and T_2 have non-disjoint signatures, then T_1 cannot freely define symbols of T_2, as the definitions can be in contradiction to T_2, and vice versa. Indeed, without restricting to disjoint signatures, definitional equivalence would not even be transitive, as proved in Barrett and Halvorson (2016) Example 5 and Lefever and Székely (2019) Theorem 1. Lefever and Székely (2019) then gives a definition of definitional equivalence that is equivalent to the standard formulation for disjoint signatures, but applies to non-disjoint signatures as well.

Definition A.1. T_1 and T_2 are *definitional equivalent by chain* if there is a finite sequence $T_1, ..., T_2$ such that for any neighboring pair T and T' in the sequence, either T definitionally extends T' or T' definitionally extends T.

It is easy to see that this relation is indeed an equivalence relation. Lefever and Székely (2019, Theorem 4) give a proof that it is equivalent to standard definitional equivalence in disjoint signatures.

Theorem A.2. *Two theories with disjoint signatures are definitional equivalent by chain if and only if they are definitional equivalent in the standard sense.*

While Lefever and Székely give only a syntactical version, the corresponding semantical version can be given in a similar spirit.

Definition A.3. $\mathcal{M}_1$ and $\mathcal{M}_2$ are *definitionally equivalent by chain* if there is a finite sequence $\mathcal{M}_1, ..., \mathcal{M}_2$ such that for any neighboring pair $\mathcal{M}$ and $\mathcal{M}'$ in the sequence, either $\mathcal{M}$ definitionally extends $\mathcal{M}'$ or $\mathcal{M}'$ definitionally extends $\mathcal{M}$.

Similarly, we can prove the following theorem.

Theorem A.4. *Two models with disjoint signatures are definitional equivalent by chain if and only if they are definitional equivalent in the standard sense.*

Here, I propose another conceptually simpler way to solve the issue of non-disjoint signatures.

Definition A.5. We say that Σ is a *disjoint copy* of Σ', if they are disjoint and there is a bijection t from Σ to Σ' that preserves arity.

Note t naturally extends to a bijection t^* from formulas of $\mathcal{L}_\Sigma$ to formulas of $\mathcal{L}_{\Sigma'}$, and hence a bijection from theories of $\mathcal{L}_\Sigma$ to theories of $\mathcal{L}_{\Sigma'}$. Also note that it induces a bijection $t^\dagger$ from models of $\mathcal{L}_\Sigma$ to models of $\mathcal{L}_{\Sigma'}$, where for any symbol $\alpha \in \Sigma$, we set $t(\alpha)^{t^\dagger(\mathcal{M})} = \alpha^\mathcal{M}$.

Definition A.6. Let T_1 be a theory in signature Σ_1, and T_2 a theory in signature Σ_2. We say that T_2 is a *disjoint copy* of T_1, if Σ_2 is a disjoint copy of Σ_1 witnessed by bijection t, and that $t^*(T_1) = T_2$.

Definition A.7. Let $\mathcal{M}_1$ be a model in signature Σ_1, and $\mathcal{M}_2$ a model in signature Σ_2. We say that $\mathcal{M}_2$ is a disjoint copy of $\mathcal{M}_1$, if Σ_2 is a disjoint copy of Σ_1 witnessed by bijection t, and that $t^\dagger(\mathcal{M}_1) = \mathcal{M}_2$.

Thus, we can define the following notion of definitional equivalence *modulo copy.*

Definition A.8. Two theories T_1 and T_2 are *definitional equivalent modulo copy* if there exists a disjoint copy T_2' of T_2 such that T_1 and T_2' are definitional equivalent.

Definition A.9. Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are *definitional equivalent modulo copy* if there exists a disjoint copy $\mathcal{M}_2'$ of $\mathcal{M}_2$ such that $\mathcal{M}_1$ and $\mathcal{M}_2'$ are definitional equivalent.

It is clear that for disjoint signatures, definitional equivalence modulo copy is equivalent to (standard) definitional equivalence.

Theorem A.10. *Two theories/models with disjoint signatures are definitional equivalent modulo copy if and only if they are definitional equivalent in the standard sense.*

It is also easy to show that definitional equivalence modulo copy is equivalent to definitional equivalence by chain.

Theorem A.11. *Two theories/models are definitional equivalent modulo copy if and only if they are definitional equivalent by chain.*

As disjoint copy gives a straightforward manner to generalize definitions and theorems given in situations with disjoint signatures. Therefore, in the maintext of this thesis, we feel free to assume in certain cases that the signatures of theories or models are disjoint.

Appendix B

An Alternative Definition of Morita Extension

We present an alternative definition of Morita extension. A definition for *pure* Morita extension will be given first.

Definition B.1. Let T be a many-sorted first-order theory in signature Σ and T^+ be a many-sorted first-order theory in signature Σ^+ such that $\Sigma \subseteq \Sigma^+$. We say that T^+ is a *atomic pure Morita extension* of T if one of the following holds:

- $\Sigma^+ = \Sigma \cup \{\sigma, \pi_0, \pi_1\}$ and we have $T^+ \equiv T \cup \{\Phi_\sigma^{prod}\}$.
- $\Sigma^+ = \Sigma \cup \{\sigma, p_0, p_1\}$ and we have $T^+ \equiv T \cup \{\Phi_\sigma^{cop}\}$.
- $\Sigma^+ = \Sigma \cup \{\sigma, \pi\}$ and we have $T^+ \equiv T \cup \{\Phi_\sigma^{sub}\}$, and $T \models \exists x^{\sigma_0} \phi_\sigma(x^{\sigma_0})$.
- $\Sigma^+ = \Sigma \cup \{\sigma, \pi\}$ and we have $T^+ \equiv T \cup \{\Phi_\sigma^{quo}\}$, and T proves that ϕ_σ defines an equivalence relation.

We say that T^+ is a *pure Morita extension* of T if there is a set of atomic pure Morita extensions $\{T_\alpha\}_{\alpha<\omega}$, such that:

- for any two distinct $\alpha_1, \alpha_2 < \omega$, $\Sigma_{T_{\alpha_1}} \cap \Sigma_{T_{\alpha_2}} = \Sigma_T$.
- $T^+ = \bigcup\{T_\alpha\}_{\alpha<\lambda}$;

Definition B.2. We say that T^+ is a *pure Morita descendant* of T if there is a finite sequence of pure Morita extensions $T_1, ..., T_n$ such that $T_1 = T$ and $T_n = T^+$.

We notice the following lemma (Lemma 5.13 in Meadows (2024)).

Lemma B.3. *T^+ is a Morita descendant of T iff there is a pure Morita descendant T^* of T such that T^+ is a definitional extension of T^*.*

This lemma then gives us an alternative definition of Morita descendant, and consequently an alternative definition of Morita equivalence.

The original proof's use of the coding lemma (as proved in Andréka, Madarász and Németi,2008 and Barrett and Halvorson, 2016) is not entirely sound. Here we present a different proof which does not rely on the coding lemma.

Lemma B.4. *(Halvorson, 2019, p.124) Let T^+ be a definitional extension of T. Let L_{T^+} be the language of T^+, and let L_T be the language of T. Then we can define a canonical translation t from L_{T^+} to L_T obtained by translating every new symbol s to its definition ϕ_s, and every old symbol s to itself, with no relativization over domains.*

Lemma B.5. *(Lemma 4.6.11 in Halvorson, 2019) Let T^+ be a definitional extension of T, and t the canonical translation. $T^+ \vdash \phi \leftrightarrow t(\phi)$ for all L_{T^+} formulas ϕ.*

Lemma B.6. *(Lemma 4.6.12 in Halvorson, 2019) Let T^+ be a definitional extension of T and t the canonical translation, if $T^+ \vdash t(\phi)$ then $T \vdash t(\phi)$.*

Now we give the proof of Lemma B.2 as follows.

Proof. We prove by induction on the length of the mixed Morita expansion T^*. Suppose the conclusion holds for mixed Morita expansion of length n. Consider a mixed Morita expansion T^*_{n+1} of length $n+1$. By definition, T^*_{n+1} is a mixed Morita successor of T^*_n which is a mixed Morita expansion of T of length n. By inductive hypothesis, there is a pure Morita expansion T^+_n of T and a definitional extension $T^\dagger_n$ of T^+_n such that $T^*_n \equiv T^\dagger_n$.

Now T^*_{n+1} is a mixed Morita successor of T^*_n. Let the set of symbols in T^*_{n+1} but not in T^*_n be Σ (i.e. $\Sigma := \Sigma_{T^*_{n+1}} \backslash \Sigma_{T^*_n}$). We first define T^+_{n+1} as a pure Morita expansion of T^+_n where we add sort symbols in Σ. Note that since definitional extensions do not define new sorts, the set of sort symbols will be the same for T^+_n and $T^\dagger_n$. And since $L_{T^\dagger_n} = L_{T^*_n}$, the set of sort symbols will be the same for T^+_n and T^*_n.

- Suppose $\sigma \in \Sigma$ is defined in T^*_{n+1} as a product of σ_1 and σ_2. Then σ_1 and σ_2 are in T^*_n and hence already in T^+_n. So we can also define σ in T^+_{n+1} as a product of σ_1 and σ_2.

- The case for coproduct is exactly the same as the product case.

- Suppose $\sigma \in \Sigma$ is defined in T^*_{n+1} as a subsort of σ_1, with the domain formula ϕ. Again, σ_1 will already be in T^+_n. And since t preserves the quantification of domains, $t(\phi)$ is a formula in $L_{T^+_n}$ with at most one free variable of sort σ_1. Thus, we can define σ in T^+_{n+1} as a subsort of σ_1 with the domain formula $t(\phi)$.

- Suppose $\sigma \in \Sigma$ is defined in T^*_{n+1} as a quotient of σ_1 by the formula ϕ. Then we define σ in T^+_{n+1} also as a quotient of σ_1 by the formula $t(\phi)$. We check that the admissibility condition holds. Let ψ be the sentence in T^*_n that says ϕ defines an equivalence relation over σ. The admissibility condition gives us that $T^*_n \vdash \psi$. Then $T^\dagger_n \vdash \psi$. By Lemma B.5, we have $T^\dagger_n \vdash t(\psi)$. By Lemma B.6, we have $T^+_n \vdash t(\psi)$. As t preserves boolean connectives and quantifications, $t(\psi)$ says that $t(\phi)$ defines an equivalence relation over σ_1. Thus, the admissibility condition indeed holds for $t(\phi)$.

Along the way, we also define the projection functions in Σ accordingly.

We have defined all the sort symbols and associated projection functions in Σ in T^+_{n+1}. Now we give a definitional extension $T^\dagger_{n+1}$ of T^+_{n+1} where we define other constants, relational or functional symbols in Σ, and also inherit definitions from $T^\dagger_n$.

For any constant, functional or relational symbol defined in $T^\dagger_n$ by formula ϕ, let it be defined in exactly the same way by ϕ in $T^\dagger_{n+1}$. For a constant, functional or relational symbol in Σ, we define it in $T^\dagger_{n+1}$ as follows. It suffices to show the case for functional symbols.

Suppose T^*_{n+1} defines a new functional symbol f by a $L_{T^*_n}$-formula ϕ with arity $\langle \sigma_1, ..., \sigma_n \rangle \to \sigma$, where $\sigma_1, ..., \sigma_n, \sigma$ are sorts in $L_{T^*_n}$. Then we may define f in $T^\dagger_{n+1}$ by the $L_{T^+_n}$-formula $t(\phi)$. Again, we check the admissibility condition. Let ψ be the sentence in T^*_n that says ϕ defines a functional symbol with arity $\langle \sigma_1, ..., \sigma_n \rangle \to \sigma$. The admissibility condition gives us that $T_n^* \vdash \psi$. Then $T_n^\dagger \vdash \psi$. By Lemma B.5, we have $T_n^\dagger \vdash t(\psi)$. By Lemma B.6, we have $T_n^+ \vdash t(\psi)$. Since T_{n+1}^+ is a pure Morita expansion of T_n^+, T_{n+1}^+ includes T_n^+, and hence $T_{n+1}^+ \vdash t(\psi)$. As t preserves boolean connectives and quantifications, $t(\psi)$ says that $t(\phi)$ defines a functional symbol with arity $\langle \sigma_1, ..., \sigma_n \rangle \to \sigma$. Thus the admissibility condition indeed holds for $t(\phi)$.

Now we have defined $T_{n+1}^\dagger$ as a definitional extension of a Morita expansion T_{n+1}^+ of T. We then show that $T_{n+1}^* \equiv T_{n+1}^\dagger$.

We first show that $T_{n+1}^\dagger$ entails T_{n+1}^*. Note that $T_{n+1}^* \equiv T_n^* \cup \Phi_1 \cup \Phi_2$, where Φ_1 is the set of definitions for new sort symbols and Φ_2 the set of definitions for new constant, functional or relational symbols.

By construction, $T_{n+1}^\dagger$ includes $T_n^\dagger$, and since $T_n^\dagger \equiv T_n^*$, we have for any $L_{T_{n+1}^*}$ sentence $\phi \in T_n^*$, $T_{n+1}^\dagger \vdash \phi$.

Let δ_σ be the definition of a new sort symbol in Φ_1. If it is a definition of a new product or coproduct, it is already included in T_{n+1}^+ and hence in $T_{n+1}^\dagger$. If it is a definition of a new quotient sort by a formula ϕ, then T_{n+1}^+ contains a definition of the same quotient sort by $t(\phi)$. By Lemma B.5, we have $T_n^\dagger \vdash \phi \leftrightarrow t(\phi)$, and thus $T_{n+1}^\dagger \vdash \phi \leftrightarrow t(\phi)$. Therefore, $T_{n+1}^\dagger \vdash \delta_\sigma$. The case for subsorts is similar. Thus, for any sentence $\phi \in \Phi_1$, $T_{n+1}^\dagger \vdash \phi$.

Let δ_s be the definition of a new constant, functional or relational symbol in Φ_2 by the T_n^*-formula ϕ. By construction, $T_{n+1}^\dagger$ includes the definition of the same constant, functional or relational symbol by the T_n^+-formula $t(\phi)$. By Lemma B.5, we have $T_n^\dagger \vdash \phi \leftrightarrow t(\phi)$, and thus $T_{n+1}^\dagger \vdash \phi \leftrightarrow t(\phi)$. Therefore, $T_{n+1}^\dagger \vdash \delta_s$. Thus, for any sentence $\phi \in \Phi_2$, $T_{n+1}^\dagger \vdash \phi$.

Therefore, for any formula ϕ, if $\phi \in T_{n+1}^*$, then $T_{n+1}^\dagger \vdash \phi$. We then prove the other direction, i.e. for any formula ϕ in $T_{n+1}^\dagger$, we have $T_{n+1}^* \vdash \phi$.

$T_{n+1}^\dagger$ is defined as the definitional extension of T_{n+1}^+ which is a pure Morita extension of T_n^+. For any formula $\phi \in T_n^+$, since $T_n^\dagger$ is a definitional extension of T_n^+, $T_n^\dagger \vdash \phi$; and since $T_n^\dagger \equiv T_n^*$, $T_n^* \vdash \phi$, and consequently $T_{n+1}^* \vdash \phi$.

T_{n+1}^+ is obtained by adding new sort symbols to T_n^+. Let δ_σ be the definition of a new sort symbol σ in T_{n+1}^+. It suffices to demonstrate the case where σ is a quotient sort defined by a formula $t(\phi)$. Note that this means that T_{n+1}^* contains a definition of σ as a quotient sort by ϕ. By Lemma B.5, we have $T_n^\dagger \vdash t(\phi) \leftrightarrow \phi$. Again, since $T_n^\dagger \equiv T_n^*$, and T_{n+1}^* includes T_n^*, we have $T_{n+1}^* \vdash t(\phi) \leftrightarrow \phi$. Therefore, $T_{n+1}^* \vdash \delta_\sigma$ as well.

Finally, $T_{n+1}^\dagger$ is defined as a definitional extension of T_{n+1}^+. For the part of $T_{n+1}^\dagger$ which coincide with $T_n^\dagger$, as $T_n^\dagger \equiv T_n^*$, we have $T_{n+1}^* \vdash \phi$ for any formula ϕ in $T_n^\dagger$. Now consider the part which consists of definitions of new constant, functional or relational symbols in Σ. It suffices to demonstrate the case where $T_{n+1}^\dagger$ contains a definition δ_f of a functional symbol f by the formula $t(\phi)$. This means that f is defined in T_{n+1}^* by ϕ. By Lemma B.5, we have $T_n^\dagger \vdash t(\phi) \leftrightarrow \phi$. Again, since $T_n^\dagger \equiv T_n^*$, and T_{n+1}^* includes T_n^*, we have $T_{n+1}^* \vdash t(\phi) \leftrightarrow \phi$. Therefore, $T_{n+1}^* \vdash \delta_f$ as well.

Thus, for any sentence $\phi \in T_{n+1}^\dagger$, we have $T_{n+1}^* \vdash \phi$. □

Appendix C

Beth's Definability Theorem for Many-Sorted Logic

In this section, we prove Beth's definability theorem for many-sorted logic. The proof essentially follows the general line in Andréka, Madarász, and Németi (2008) who prove the theorem for many-sorted theories with countable languages (for the use of omitting type theorem), finitely many sorts (for the translation to first-order logic) and have at least one sort with more than one element (for technical reasons). The following proof generalizes the result in the sense that we will assume none of these conditions. Instead, we will assume the many-sorted version of some standard model-theoretical results, in particular, the fundamental theorem and the Keisler-Shelah theorem of ultraproducts. Following standard model-theoretical textbooks (e.g., Hodges (1993)), it can be checked that the single-sorted proofs of such theorems also work for their many-sorted analogs by straightforwardly changing the notation. For other standard results about many-sorted logic, e.g., completeness and compactness, see Manzano (1996).

Unless otherwise noted, theories in this section concern first-order many-sorted logic. For convenience, we focus only on relational theories. Morita extension in this section is allowed to do arbitrarily finite products and coproducts. We use $\overline{x^{\sigma}}$ for sequences of variables of sorts $\overline{\sigma}$, and we use the notation $(\overline{x^{\sigma}})^i$ to denote the i-th element of $\overline{x^{\sigma}}$. And we assume in the following that T^+ is a Σ^+-theory and $T = T^+|_{\Sigma}$, for some $\Sigma \subseteq \Sigma^+$.

Definitions C.1-C.3 are generalized from Hodges (1993, Chapter 12).

Definition C.1. We say that T^+ is implicitly definable in T if for any model of $\mathcal{M}$ of T, and any two expansions $\mathcal{M}_1$ and $\mathcal{M}_2$ of $\mathcal{M}$ which are models of T^+, we have that there is a unique isomorphism from $\mathcal{M}_1$ to $\mathcal{M}_2$ which fixes $\mathcal{M}$.

Definition C.2. We say that T^+ has the uniform reduction property over T if for any Σ^+-formula $\phi(\overline{x^{\sigma}})$, where $\overline{x^{\sigma}}$ are Σ-variables, there is a Σ-formula $tr(\phi)$ with free variables among $\overline{x^{\sigma}}$ such that $T^+ \vdash \phi(\overline{x^{\sigma}}) \leftrightarrow tr(\phi)(\overline{x^{\sigma}})$.

Definition C.3. We say that T^+ is coordinatised over T if:

- T^+ has the uniform reduction property over T;
- For any $\mathcal{M}^+$ of T^+, every element in $\mathcal{M}^+$ is in the definable closure of $\mathcal{M}^+|_{\Sigma}$.

The following definition generalizes the notion of coding in Andréka, Madarász, and Németi (2008) and Barrett and Halvorson (2016), which is essentially an analog of "piecewise" translation in Visser (2009): we allow a new sort to be "coded" by many different sequences of sorts in different contexts.

The idea is intuitive: let $\Sigma^+ = \Sigma \cup \Sigma^*$, and we want to "code" a new Σ^+ theory into an old Σ theory. We will allow a new sort σ to be constructed by taking the union of a set $f_0(\sigma)$ of sequences of old sorts. So whenever we want to translate a new formula with new variables, we need to use

$k \in \Pi_{\sigma_i \in \Sigma^*} f_0(\sigma_i)$ to record the old sorts from which the new objects are supposed to be constructed. The formal definition is as follows.

Definition C.4. Let $\Sigma^+ = \Sigma \cup \Sigma^*$ and let T^+ be a Σ^+-theory. We use $x, y, z...$ as meta-variables for variables in Σ^* and p, q, r for variables in Σ.

A coding of of T^+ in T is a function $f = f_0 \cup f_1 \cup f_2 \cup f_3$ defined as follows:

- f_0 (the map of sorts): for each sort σ in Σ^*, $f_0(\sigma) = \{\langle 0, \overline{\sigma_0}\rangle, ..., \langle n, \overline{\sigma_n}\rangle\}$ where $\overline{\sigma_0}, ..., \overline{\sigma_n}$ are sequences of sorts in Σ;

- f_1 (the map for variables): for each variable x^σ in Σ_1, $f_1(x^\sigma)$ is a set of sequences of Σ_2-variables $\{\overline{x}^{\overline{\sigma}}_{\langle i, \overline{\sigma}\rangle} | \langle i, \overline{\sigma}\rangle \in f_0(\sigma)\}$;

    - We require that $\overline{y}^{\overline{\beta}}_{\langle i, \overline{\beta}\rangle}, \overline{z}^{\overline{\gamma}}_{\langle j, \overline{\gamma}\rangle} \in f_1(x^\sigma)$ are disjoint if $i \neq j$.
    - We require that variables appeared in $f_1(x^\sigma)$ and $f_1(y^\gamma)$ are disjoint if x^σ and y^γ are distinct variables;
    - For convenience, we call $k \in \Pi_{x^\sigma \in \Sigma_1} f_1(x^\sigma)$ an assignment of f. We write $A(f) := \Pi_{x^\sigma \in \Sigma_1} f_1(x^\sigma)$. And we write $\tilde{k}(x^\sigma) := \langle i, \overline{\sigma}\rangle$ for $k(x^\sigma) = \overline{x}^{\overline{\sigma}}_{\langle i, \overline{\sigma}\rangle}$.

- f_2 (the coding formulas): for each variable x^σ in Σ^*, $f_2(x^\sigma)$ is a function from $A(f)$ to Σ^+-formulas, such that for each assignment k, $f_2(x^\sigma)(k)$ is a Σ_2-formula with free variables among $x^\sigma, k(x^\sigma)$;

    - We require that if $\tilde{k}(x^\sigma) = \tilde{k'}(y^\sigma)$, then $f_2(x^\sigma)(k) = f_2(y^\sigma)(k')[y^\sigma \mapsto x^\sigma, k'(y^\sigma) \mapsto k(x^\sigma)]$;

- f_3 (the map for formulas): for each assignment k, and each Σ^+-formula $\phi(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}})$, $f_3(\phi)(k)$ is a Σ-formula with free variables among $k(x^{\sigma_1})(k), ..., k(x^{\sigma_n}), \overline{p}^{\overline{\gamma}}$

    - if $\tilde{k}(x^{\sigma_i}) = \tilde{k'}(y^{\sigma_i})$ for all $1 \leq i \leq n$,
      $f_3(\phi(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}}))(k) = f_3(\phi(y^{\sigma_1}, ..., y^{\sigma_n}, \overline{q}^{\overline{\gamma}}))(k)[k(\overline{y}^{\overline{\sigma}}) \mapsto k(\overline{x}^{\overline{\sigma}}), \overline{q}^{\overline{\gamma}} \mapsto \overline{p}^{\overline{\gamma}}]$

Let $[x^{\sigma_1}, ..., x^{\sigma_n}]$ be the quotient of $A(f)$ by the equivalence relation $k \sim k'$ if $k(x^{\sigma_i}) = k'(x^{\sigma_i})$ for all $1 \leq i \leq n$. We write $\forall k(\overline{x}^{\overline{\sigma}})$ for $\forall k(x^{\sigma_1})...\forall k(x^{\sigma_n})$, and $f_2(\overline{x}^{\overline{\sigma}})(k)$ for $f_2(x^{\sigma_1})(k) \wedge ... \wedge f_2(x^{\sigma_n})(k)$.

We further require that

1. $T^+ \vdash \forall x^\sigma (\bigvee_{[k] \in [x^\sigma]} \exists k(x^\sigma)[(f_2(x^\sigma))(k)])$,
2. $T^+ \vdash \forall k(x^\sigma) \exists_{\leq 1} x^\sigma (f_2(x^\sigma)(k))$
3. for any two k, k', if $k(x^\sigma) \neq k'(x^\sigma)$, then $T \vdash \forall x^\sigma \neg(\exists k'(x^\sigma)[(f_2(x^\sigma))(k')] \wedge \exists k(x^\sigma)[(f_2(x^\sigma))(k)])$.
4. For any Σ^+-formula $\phi(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}})$, we have that
   $T^+ \vdash \phi(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}}) \leftrightarrow \bigwedge_{[k] \in [\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})(f_2(\overline{x}^{\overline{\sigma}})(k) \to f_3(\phi)(k))$

Lemma C.5. *If T^+ is implicitly definable in T, then for any model $\mathcal{M}_1, \mathcal{M}_2$ of T, and any of their expansions $\mathcal{M}_1^+$ and $\mathcal{M}_2^+$ in $Mod(T^+)$, if there is an isomorphism f from $\mathcal{M}_1$ to $\mathcal{M}_2$, then there is a unique isomorphism f^+ from $\mathcal{M}_1^+$ to $\mathcal{M}_2^+$ that extends f.*

Proof. Let h be the isomorphism from $\mathcal{M}_1$ to $\mathcal{M}_2$. Substituting $\mathcal{M}_1$ for $\mathcal{M}_2$ in $\mathcal{M}_1^+$, we have a model $\mathcal{M}_2^*$. Since $\mathcal{M}_1 \cong \mathcal{M}_2$, there is an isomorphism $k : \mathcal{M}_2^* \cong \mathcal{M}_1^+$ that extends h. Since $\mathcal{M}_2^*$ and $\mathcal{M}_2^+$ are both expansions of $\mathcal{M}_2$, there is a unique isomorphism between $\mathcal{M}_2^*$ and $\mathcal{M}_2^+$ fixing $\mathcal{M}_2$. Thus, $\mathcal{M}_1^+ \cong \mathcal{M}_2^+$.

Suppose (towards a contradiction) that there are two different isomorphisms f, g from $\mathcal{M}_2^+$ to $\mathcal{M}_1^+$ that extends h^{-1}. Then, combined with k, $k \circ f$ and $k \circ g$ are two different isomorphisms from $\mathcal{M}_2^+$ to $\mathcal{M}_2^*$ (Note that k is a mono in $Mod(T^+)$). Since k extends h, and both f, g extend h^{-1}, we have that $k \circ f$ and $k \circ g$ are two different isomorphisms from $\mathcal{M}_2^+$ to $\mathcal{M}_2^*$ both fixing $\mathcal{M}_2$. Contradiction.

Therefore, there is a unique isomorphism f^+ from $\mathcal{M}_1^+$ to $\mathcal{M}_2^+$ which extends h. □

The following theorem is a variant of Hodges (1993), Lemma 12.5.1, and the proof is a modification of Theorem 3.3.3 in Andréka, Madarász and Németi (2008). Both of the proofs essentailly use the completeness of theories.[1] Here, with the assumption of implicit definability, the assumption of completeness can be dropped. We assume the many-sorted version of the Keisler-Shelah theorem, the fundamental theorem and the expansion theorem of ultraproducts.

Theorem C.6. *If T^+ is implicitly definable in T, then T^+ has the uniform reduction property over T.*

Proof. Suppose (towards a contradiction) that this is not the case. Then there is a Σ^+-formula $\phi(x^{\sigma_1}, ..., x^{\sigma_n})$ such that for any Σ-formula ψ with free variables among $x^{\sigma_1}, ..., x^{\sigma_n}$, $T \not\models \phi(\overline{x}^{\overline{\sigma}}) \leftrightarrow \psi(\overline{x}^{\overline{\sigma}})$.

We claim that there are models $\mathcal{M}_1$ and $\mathcal{M}_2$ of T^+, their submodels $\mathcal{N}_1 := (\mathcal{M}_1)|_\Sigma$, $\mathcal{N}_2 := (\mathcal{M}_2)|_\Sigma$, and sequences of elements, $\overline{a}^{\overline{\sigma}}$ in $\mathcal{N}_1$, $\overline{b}^{\overline{\sigma}}$ in $\mathcal{N}_2$, such that $tp_{\mathcal{N}_1}(\overline{a}^{\overline{\sigma}}) = tp_{\mathcal{N}_2}(\overline{b}^{\overline{\sigma}})$, and yet $\mathcal{M}_1 \models \phi(\overline{a}^{\overline{\sigma}})$ and $\mathcal{M}_2 \not\models \phi(\overline{b}^{\overline{\sigma}})$.

Suppose (towards a contradiction) that this is not the case, then if $tp_{\mathcal{N}_1}(\overline{a}^{\overline{\sigma}}) = tp_{\mathcal{N}_2}(\overline{b}^{\overline{\sigma}})$, we must have $\mathcal{M}_1 \models \phi(\overline{a}^{\overline{\sigma}})$ if and only if $\mathcal{M}_2 \models \phi(\overline{b}^{\overline{\sigma}})$. Then there is a subset $\Phi \subseteq S_{\overline{\sigma}}T$ such that for any model $\mathcal{M}$ of T^+, and any sequence of elements $\overline{a}^{\overline{\sigma}}$ in $\mathcal{M}$, we have that $\mathcal{M} \models \phi(\overline{a}^{\overline{\sigma}})$ if and only if $tp_{\mathcal{M}|_\Sigma}(\overline{a}^{\overline{\sigma}}) \in \Phi$. For any $p \in \Phi$, $T^+ \cup p \cup \{\neg\phi\}$ is inconsistent. By compactness, we can find a finite subset of p, take its conjunction as ψ_p and we have that $T^+ \vdash \psi_p \to \phi$. Then we may list all the types in Φ as $\{p_i\}_{i<\lambda}$. Then for any model $\mathcal{M}$ of $T^+, \overline{a}^{\overline{\sigma}}$ in $\mathcal{M}$, $\mathcal{M} \models \phi(\overline{a}^{\overline{\sigma}})$ if and only if $\mathcal{M} \models \psi_{p_i}(\overline{a}^{\overline{\sigma}})$ for some $i < \lambda$. Then $T^+ \cup \{\neg\psi_{p_i} | i < \lambda\} \cup \{\phi\}$ is inconsistent. By compactness, there is a finite subset $\{\psi_1, ..., \psi_n\}$ such that $T^+ \vdash (\bigwedge_{1<i<n} \neg\psi_i) \to \neg\phi$. As $T^+ \vdash \psi_p \to \phi$ for all p, we also have $T^+ \vdash \neg\phi \to (\bigwedge_{1<i<n} \neg\psi_i)$. Therefore, $T^+ \vdash \neg\phi \leftrightarrow (\bigwedge_{1<i<n} \neg\psi_i)$. Contradiction.

Then $\mathcal{N}_1, \overline{a}^{\overline{\sigma}} \equiv \mathcal{N}_2, \overline{b}^{\overline{\sigma}}$. By the Keisler-Shelah theorem, there is an ultrafilter $\mathcal{U}$ over some index set I such that $\Pi_D\mathcal{N}_1, \overline{a}^{\overline{\sigma}} \cong \Pi_D\mathcal{N}_2, \overline{b}^{\overline{\sigma}}$. By the expansion theorem (p.216, Chang and Keisler, 1973),

[1] Theorem 3.3.3 in Andréka, Madarász and Németi (2008) says if "For any model B of T, every automorphism of B_P extends to an automorphism of B" then T has the uniform reduction property (for the notation adopted here, see Hodges (1993, chapter 12)). Despite claiming, the theorem still essentially requires the completeness of T. This can be shown as follows. Let L be the language with only one unary predicate symbol Q, and let L^+ contain two more unary predicate symbols S and P. Let T contain the following two sentences

- $\forall x P(x)$
- $\forall x \forall y [S(x) \leftrightarrow S(y)] \leftrightarrow [Q(x) \leftrightarrow Q(y)]$

And we relativize over P (i.e. we do trivial relativization). Then for any model B of T, an automorphism f of $B|_L$ maps an element in/not in Q to another element in/not in Q, by (2), these two elements must be in S or not in S at the same time, so f extends to an automorphism of B. But there is no uniform reduction of $S(x)$, as in some models it corresponds to $Q(x)$, while in others, it corresponds to $\neg Q(x)$. This point is confirmed by the authors in private correspondence, to which I am very thankful.

$\Pi_D\mathcal{M}_1, \overline{a}^{\overline{\sigma}}$ and $\Pi_D\mathcal{M}_2, \overline{b}^{\overline{\sigma}}$ are expansions of $\Pi_D\mathcal{N}_1, \overline{a}^{\overline{\sigma}}$ and $\Pi_D\mathcal{N}_2, \overline{b}^{\overline{\sigma}}$. However, by the fundamental theorem of ultraproducts, $\Pi_D\mathcal{M}_1, \overline{a}^{\overline{\sigma}} \models \phi$ but $\Pi_D\mathcal{M}_2, \overline{b}^{\overline{\sigma}} \not\models \phi$. Thus, there is an isomorphism f from $\Pi_D\mathcal{N}_1$ to $\Pi_D\mathcal{N}_2$ which maps $\overline{a}^{\overline{\sigma}}$ to $\overline{b}^{\overline{\sigma}}$ but this isomorphism cannot be extended to an isomorphism from $\Pi_D\mathcal{M}_1$ to $\Pi_D\mathcal{M}_2$. This contradicts Lemma C.5. □

To prove the next major lemma (Lemma C.14), we need to transform Lemma C.7 and Lemma C.8 below to the many-sorted context (Lemma C.13).

Lemma C.7. *(Hodeges, 1993, p.279, a corollary of Theorem 10.2.1) Every single-sorted model $\mathcal{M}$ has a λ-big elementary extension for arbitrarily large cardinal λ.*

Lemma C.8. *(Theorem 6.3.2 in Hodges, 1993) Let $\mathcal{M}$ be a single-sorted model, X a set of elements of $\mathcal{M}$, and $\overline{a}$, $\overline{b}$ two sequences of elements of $\mathcal{M}$. Write $G_{(X)}$ for the group of all automorphisms of $\mathcal{M}$ which pointwise fix X. If $\mathcal{M}$ is λ-big, then the following are equivalent:*

- *There is an automorphism g in $G_{(X)}$ such that $g(\overline{a}) = \overline{b}$;*
- *$tp_{\mathcal{M}}(\overline{a}\backslash X) = tp_{\mathcal{M}}(\overline{b}\backslash X)$.*

Definition C.9. Let Σ be a many-sorted signature with sorts $\{\sigma_i\}_{i<\lambda}$. Let $\lfloor\Sigma\rfloor$ be a single-sorted signature which contains a un-sorted copy of each relational symbol and variable in Σ and unitary predicates U_i for each sort symbol σ_i in Σ. We use $\lfloor s\rfloor$ to denote the un-sorted copy of symbol s.

Definition C.10. Let $\mathcal{M}$ be a model of a many-sorted signature Σ. Let $\lfloor\mathcal{M}\rfloor$ be a model of $\lfloor\Sigma\rfloor$ constructed as follows:

- The domain of $\lfloor\mathcal{M}\rfloor$ is the disjoint union of domains of $\mathcal{M}$. That is, the domain consists of, for each sorted element a^{σ_i} of $\mathcal{M}$, a corresponding element $\lfloor a^{\sigma_i}\rfloor = (a, \sigma_i)$ in $\lfloor\mathcal{M}\rfloor$;
- The interpretation of relational symbols in $\lfloor\mathcal{M}\rfloor$ is the same as in $\mathcal{M}$ modulo the disjoint union. That is, for each relational symbol R with arity $\langle\sigma_1, ..., \sigma_n\rangle$, we have that $\lfloor R\rfloor^{\lfloor\mathcal{M}\rfloor} = \{((a_1, \sigma_1), ..., (a_n, \sigma_n)) | \mathcal{M} \models R(a_1, ..., a_n)\}$;
- The interpretation of each unitary predicate U_i in $\lfloor\mathcal{M}\rfloor$ is the set $\{(a, \sigma_i) | a \in (\sigma_i)^{\mathcal{M}}\}$.

Definition C.11. Let ϕ be an arbitrary formula in a many-sorted signature Σ with sorts $\{\sigma_i\}_{i<\lambda}$. Let $\lfloor\phi\rfloor$ be the formula in the single-sorted signature $\lfloor\Sigma\rfloor$ constructed inductively as follows:

- If $\phi = R(x^{\sigma_1}, ..., x^{\sigma_n})$, $\lfloor\phi\rfloor := \lfloor R\rfloor(\lfloor x^{\sigma_1}\rfloor, ..., \lfloor x^{\sigma_n}\rfloor)$;
- If $\phi = \chi_1 \wedge \chi_2$, then $\lfloor\phi\rfloor := \lfloor\chi_1\rfloor \wedge \lfloor\chi_2\rfloor$;
- If $\phi = \neg\psi$, then $\lfloor\phi\rfloor := \neg\lfloor\psi\rfloor$;
- If $\phi = \forall x^{\sigma_n}\psi$, then $\lfloor\phi\rfloor := \forall\lfloor x^{\sigma_n}\rfloor(U_n(\lfloor x^{\sigma_1}\rfloor) \rightarrow \lfloor\psi\rfloor)$.

Lemma C.12. *Let $\mathcal{M}$ and $\lfloor\mathcal{M}\rfloor$ be as above. Then for any sequence of sorted elements $\overline{a}^{\overline{\sigma}}$, $\mathcal{M} \models \phi(\overline{a}^{\overline{\sigma}})$ if and only if $\lfloor\mathcal{M}\rfloor \models \lfloor\phi\rfloor(\lfloor\overline{a}^{\overline{\sigma}}\rfloor)$.*

Proof. By straightforward induction on the structure of ϕ. □

Lemma C.13. *Let $\mathcal{M}$ be a Σ-model and X a set of sorted elements in $\mathcal{M}$. Suppose that $tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}}\backslash X) = tp_{\mathcal{M}}(\overline{b}^{\overline{\sigma}}\backslash X)$. Then there is an elementary extension $\mathcal{M}^+$ of $\mathcal{M}$ such that there is an automorphism g which sends $\overline{a}^{\overline{\sigma}}$ to $\overline{b}^{\overline{\sigma}}$ and pointwise fixes X.*

Proof. Assume that $tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}}\backslash X) = tp_{\mathcal{M}}(\overline{b}^{\overline{\sigma}}\backslash X)$. By straightforward induction, one proves that for any $\lfloor\Sigma\rfloor$-formula ϕ with parameters from X, we have $\lfloor\mathcal{M}\rfloor, s \models \phi(\lfloor\overline{a}^{\overline{\sigma}}\rfloor)$ for any assignment s iff $\lfloor\mathcal{M}\rfloor, s \models \phi(\lfloor\overline{b}^{\overline{\sigma}}\rfloor)$ for any assignment s. Thus, $tp_{\lfloor\mathcal{M}\rfloor}(\lfloor\overline{a}^{\overline{\sigma}}\rfloor\backslash\lfloor X\rfloor) = tp_{\lfloor\mathcal{M}\rfloor}(\lfloor\overline{b}^{\overline{\sigma}}\rfloor\backslash\lfloor X\rfloor)$.[2]

By Lemma C.7 and C.8, there is an elementary extension $\lfloor\mathcal{M}\rfloor^+$ of $\lfloor\mathcal{M}\rfloor$ such that there is an automorphism g of $\lfloor\mathcal{M}\rfloor^+$ which pointwise fixes $\lfloor X\rfloor$ and sends $\lfloor\overline{a}^{\overline{\sigma}}\rfloor$ to $\lfloor\overline{b}^{\overline{\sigma}}\rfloor$.

Let $\mathcal{M}^+$ be the many-sorted model of signature Σ constructed from $\lfloor\mathcal{M}\rfloor^+$ as follows:

- $\sigma_i^{\mathcal{M}^+} := \{\chi(p) | p \in (U_i)^{\lfloor\mathcal{M}\rfloor^+}\}$, where $\chi(p) = a$ if $p = (a, \sigma_i) \in (U_i)^{\mathcal{M}}$, and identity otherwise;
- $R^{\mathcal{M}^+} := \{(\chi(p_1), ..., \chi(p_n)) | (p_1, ..., p_n) \in (\lfloor R\rfloor)^{\lfloor\mathcal{M}\rfloor^+}\}$; Note that this is well-defined, since for any relational symbol $R \in \Sigma$ with arity $\langle\sigma_1, ..., \sigma_n\rangle$, $\lfloor\mathcal{M}\rfloor \models \forall x_1 ... \forall x_n \lfloor R\rfloor(x_1, ..., x_n) \to \bigwedge U_i(x_i)$, and hence $\lfloor\mathcal{M}\rfloor^+ \models \forall x_1 ... \forall x_n \lfloor R\rfloor(x_1, ..., x_n) \to \bigwedge U_i(x_i)$.

Now we check that $\mathcal{M}^+$ is an elementary extension of $\mathcal{M}$. It is an extension of $\mathcal{M}$ by construction. To see that it is elementary, we note that for any Σ-formula $\phi(\overline{x}^{\sigma})$, we have that $\mathcal{M} \models \phi(\overline{a}^{\sigma})$ iff $\lfloor\mathcal{M}\rfloor \models \lfloor\phi\rfloor(\lfloor\overline{a}^{\sigma}\rfloor)$ (by Lemma C.12) iff $\lfloor\mathcal{M}\rfloor^+ \models \lfloor\phi\rfloor(\lfloor\overline{a}^{\sigma}\rfloor)$ (by elementary extension) iff $\mathcal{M}^+ \models \phi(\overline{a}^{\sigma})$ (by straightforward induction).

The autormophism g on $\lfloor\mathcal{M}\rfloor^+$ induces a automorphism $g^\dagger$ on $\mathcal{M}^+$, defined as $g^\dagger(a) = b$ iff $g(\chi^{-1}(a)) = \chi^{-1}(b)$.[3] By construction of χ, $g^\dagger$ pointwise fixes X and sends $\overline{a}^{\overline{\sigma}}$ to $\overline{b}^{\overline{\sigma}}$. □

Lemma C.14. *Suppose T^+ is implicitly definable in T. Let $\mathcal{M}$ be an arbitrary model of T^+, and let $\mathcal{N} := \mathcal{M}|_{\Sigma}$. Let $\overline{a}^{\overline{\sigma}}$ be a sequence of elements in $\mathcal{M}$. Then $tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}}\backslash\mathcal{N})$ is isolated in $S^{\mathcal{M}}_{\overline{\sigma}}(\mathcal{N})$.*

Proof. Suppose (towards a contradiction) that this is not the case. Then there is a model $\mathcal{M}$ of T^+, and a sequence $\overline{a}^{\overline{\sigma}}$ in $\mathcal{M}$ such that for any formula $\phi(\overline{x}^{\overline{\sigma}}) \in tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}}\backslash\mathcal{N})$, there is a formula $\psi(\overline{x}^{\overline{\sigma}}) \in tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}}\backslash\mathcal{N})$ such that $Th_{\mathcal{N}}(\mathcal{M}) \not\models \phi(\overline{x}^{\overline{\sigma}}) \to \psi(\overline{x}^{\overline{\sigma}})$. Since $Th_{\mathcal{N}}(\mathcal{M})$ is complete, $Th_{\mathcal{N}}(\mathcal{M}) \models \exists\overline{x}^{\overline{\sigma}}(\phi(\overline{x}^{\overline{\sigma}}) \wedge \neg\psi(\overline{x}^{\overline{\sigma}}))$. Therefore, there is a sequence $\overline{b}^{\overline{\sigma}}$ in $\mathcal{M}$ such that $\mathcal{M} \models \phi(\overline{b}^{\overline{\sigma}}) \wedge \neg\psi(\overline{b}^{\overline{\sigma}})$. Then we have $\overline{b}^{\overline{\sigma}} \neq \overline{a}^{\overline{\sigma}}$.

[2] We demonstrate the base cases of the induction for relational theories as follows. The inductive steps are trivial. And the proof for non-relational theories is similar.

- $\phi := x = y$, $\phi := x = \lfloor p^{\delta}\rfloor$ or $\phi := \lfloor p_1^{\delta_i}\rfloor = \lfloor p_2^{\delta_j}\rfloor$: conclusion holds rivially.
- $\phi := x = \lfloor a^{\sigma_i}\rfloor$. Then, $\lfloor\mathcal{M}\rfloor, s \models x = \lfloor a^{\sigma_i}\rfloor$ for any assignment s iff there is only one sort σ_i with one object a^{σ_i} in $\mathcal{M}$ iff $\lfloor\mathcal{M}\rfloor, s \models \forall x(x = \lfloor b^{\sigma_i}\rfloor)$ for any assignment s.
- $\phi := \lfloor p^{\delta}\rfloor = \lfloor a^{\sigma}\rfloor$. Then $\lfloor\mathcal{M}\rfloor \models \lfloor p^{\delta}\rfloor = \lfloor a^{\sigma}\rfloor$ iff $\delta = \sigma$ and $\mathcal{M} \models p^{\delta} = a^{\sigma}$ iff $\delta = \sigma$ and $\mathcal{M} \models p^{\delta} = b^{\sigma}$ iff $\lfloor\mathcal{M}\rfloor \models \lfloor b^{\sigma}\rfloor = \lfloor p^{\delta}\rfloor$.
- $\phi := \lfloor a_1^{\sigma_i}\rfloor = \lfloor a_2^{\sigma_j}\rfloor$. Then $\lfloor\mathcal{M}\rfloor \models \lfloor a_1^{\sigma_i}\rfloor = \lfloor a_2^{\sigma_j}\rfloor$ iff $\sigma_i = \sigma_j$ and $\mathcal{M} \models a_1^{\sigma_i} = a_2^{\sigma_j}$ iff $\sigma_i = \sigma_j$ and $\mathcal{M} \models b_1^{\sigma_i} = b_2^{\sigma_j}$ iff $\lfloor\mathcal{M}\rfloor \models \lfloor b_1^{\sigma_i}\rfloor = \lfloor b_2^{\sigma_j}\rfloor$.
- $\phi := \lfloor R\rfloor(\lfloor x_1^{\sigma_1}\rfloor, ..., \lfloor x_n^{\sigma_n}\rfloor, \lfloor\overline{a}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$. Then $\lfloor\mathcal{M}\rfloor, s \models \lfloor R\rfloor(\lfloor x_1^{\sigma_1}\rfloor, ..., \lfloor x_n^{\sigma_n}\rfloor, \lfloor\overline{a}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$ for any assignment s iff there is only one sort σ and $\mathcal{M}, s \models \forall x_1^{\sigma} ... x_n^{\sigma} R(x_1^{\sigma} ... x_n^{\sigma}, \overline{a}^{\overline{\sigma}}, \overline{p}^{\overline{\delta}})$ iff there is only one sort σ and $\mathcal{M}, s \models \forall x_1^{\sigma} ... x_n^{\sigma} R(x_1^{\sigma} ... x_n^{\sigma}, \overline{b}^{\overline{\sigma}}, \overline{p}^{\overline{\delta}})$ iff $\lfloor\mathcal{M}\rfloor, s \models \lfloor R\rfloor(\lfloor x_1^{\sigma_1}\rfloor, ..., \lfloor x_n^{\sigma_n}\rfloor, \lfloor\overline{b}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$ for any assignment s.
- $\phi := \lfloor R\rfloor(\lfloor\overline{a}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$. Then $\lfloor\mathcal{M}\rfloor \models \lfloor R\rfloor(\lfloor\overline{a}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$ iff $\mathcal{M} \models R(\overline{a}^{\overline{\sigma}}, \overline{p}^{\overline{\delta}})$ iff $\mathcal{M} \models R(\overline{b}^{\overline{\sigma}}, \overline{p}^{\overline{\delta}})$ iff $\lfloor\mathcal{M}\rfloor \models \lfloor R\rfloor(\lfloor\overline{b}^{\overline{\sigma}}\rfloor, \lfloor\overline{p}^{\overline{\delta}}\rfloor)$.

[3] It can be checked by straightforward induction that $g^\dagger$ is indeed an automorphism on $\mathcal{M}^+$.

Thus, for any formula $\phi(\overline{x}^{\overline{\sigma}}) \in tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}} \backslash \mathcal{N})$, $T \cup \{\phi(\overline{x}^{\overline{\sigma}}) \wedge \phi(\overline{y}^{\overline{\sigma}}), \overline{x}^{\overline{\sigma}} \neq \overline{y}^{\overline{\sigma}}\}$ is consistent. By compactness, $T \cup \{\phi(\overline{x}^{\overline{\sigma}}) \wedge \phi(\overline{y}^{\overline{\sigma}}) | \phi \in tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}} \backslash \mathcal{N})\} \cup \{\overline{x}^{\overline{\sigma}} \neq \overline{y}^{\overline{\sigma}}\}$ is consistent. Then there is a model $\mathcal{M}$ of T, and sequences $\overline{a}^{\overline{\sigma}}$, $\overline{b}^{\overline{\sigma}}$ in $\mathcal{M}$ such that $tp_{\mathcal{M}}(\overline{a}^{\overline{\sigma}} \backslash \mathcal{N}) = tp_{\mathcal{M}}(\overline{b}^{\overline{\sigma}} \backslash \mathcal{N})$ and $\overline{a}^{\overline{\sigma}} \neq \overline{b}^{\overline{\sigma}}$. By Lemma C.13, we can find an elementary extension $\mathcal{M}^+$ of $\mathcal{M}$, and an automorphism f of $\mathcal{M}^+$ which pointwise fixes $\mathcal{N}$ and maps $\overline{a}^{\overline{\sigma}}$ to $\overline{b}^{\overline{\sigma}}$. Since $\overline{a}^{\overline{\sigma}}$ and $\overline{b}^{\overline{\sigma}}$ are distinct, f cannot be the identity. This contradicts the fact that T^+ is implicitly definable in T. □

Theorem C.15. *If T^+ is implicitly definable in T, then T^+ is coordinatised over T.*

Proof. We have already proved that T^+ has the uniform reduction property over T. It remains to show that for any model $\mathcal{M}$ of T^+, every element in $\mathcal{M}$ is in the definable closure of $\mathcal{N} := \mathcal{M}^+|_\Sigma$.

Let $\mathcal{M}$ be an arbitrary model of T^+, and $\mathcal{N} := \mathcal{M}|_\Sigma$ and let a^σ be an arbitrary element of $\mathcal{M}$. By Lemma C.14, $tp_{\mathcal{M}}(a^\sigma \backslash \mathcal{N})$ is isolated in $S_\sigma^{\mathcal{M}}(\mathcal{N})$. Let it be isolated by a formula $\phi(x^\sigma) \in tp_{\mathcal{M}}(a^\sigma \backslash \mathcal{N})$. We show that $\mathcal{M} \models \exists! x^\sigma \phi(x^\sigma)$. Suppose (towards a contradiction) that there are two distinct elements b^σ and c^σ in $\mathcal{M}$ such that $\mathcal{M} \models \phi(b^\sigma) \wedge \phi(c^\sigma)$. Since $tp_{\mathcal{M}}(a^\sigma \backslash \mathcal{N})$ is isolated, we have that $tp_{\mathcal{M}}(b^\sigma \backslash \mathcal{N}) = tp_{\mathcal{M}}(c^\sigma \backslash \mathcal{N})$. By Lemma C.13, we can find an elementary extension $\mathcal{M}^+$ of $\mathcal{M}$ and there is an automorphism f of $\mathcal{M}^+$ which pointwise fixes $\mathcal{N}$ and maps b^σ to c^σ. Since b^σ and c^σ are distinct, f cannot be the identity. This contradicts the fact that T^+ is implicitly definable in T. Therefore, we have that $\mathcal{M} \models \exists! x^\sigma \phi(x^\sigma)$. □

Theorem C.16. *Assume that T^+ is coordinatised over T. Then T^+ is coded in T.*

Proof. We construct a coding f as follows.

Let $\Sigma^+ = \Sigma \cup \Sigma^*$. We use $x, y, z, ...$ as meta-variables for variables in Σ^* and p, q, r for variables in Σ. And let $\Psi(\sigma) = \{\phi(x^\sigma, \overline{p}^{\overline{\gamma}}) | \phi \in Form_{\Sigma^+}(x^\sigma, \overline{p}^{\overline{\gamma}})\}$.

We first prove that for any new sort σ in Σ^*, there is a finite set of formulas $\Phi(\sigma) \subseteq \Psi(\sigma)$ such that for any model $\mathcal{M}$ of T^+, every element in $\sigma^{\mathcal{M}}$ is defined by some formula in $\Phi(\sigma)$, with $\overline{p}^{\overline{\gamma}}$ filled by some elements in $\mathcal{M}|_\Sigma$ of sorts $\overline{\gamma}$.

Suppose (towards a contradiction) that this is not the case. Then there is a new sort σ such that for any finite set of formulas Φ, $T \cup \bigwedge_{\phi \in \Phi} \neg \exists \overline{p}^{\overline{\gamma}} \phi(x^\sigma, \overline{p}^{\overline{\gamma}})$ is consistent. By compactness, $T \cup \{\neg \exists \overline{p}^{\overline{\gamma}} \phi(x^\sigma, \overline{p}^{\overline{\gamma}}) | \phi \in \Psi(\sigma)\}$ is also consistent. But then there is a model $\mathcal{M}$ of T, and an element a^σ in $\mathcal{M}$ that is not in the definable closure of $\mathcal{M}|_\Sigma$. Contradiction.

Let σ be an arbitrary sort in Σ^*. We may enumerate $\Phi(\sigma)$ as $\{\phi_1(x^\sigma, p_1^{\overline{\sigma_1}}), ..., \phi_n(x^\sigma, p_n^{\overline{\sigma_n}})\}$. Let $\theta_m := \phi_m \wedge \bigwedge_{1 \leq i \leq m} \neg \exists \overline{p_i}^{\overline{\sigma_i}}(\phi_i)$, and let $\Theta(\sigma) := \{\theta_m | 1 \leq m \leq n\}$. Then for arbitrary $\mathcal{M}$, every element in $\sigma^{\mathcal{M}}$ is defined by precisely one formula in $\Theta(\sigma)$.

Then let $f_0(\sigma) := \{\langle i, \overline{\sigma_i} \rangle | \theta_i(x^\sigma, \overline{p_i}^{\overline{\sigma_i}}) \in \Theta(\sigma)\}$. Let f_1 be chosen arbitrarily.

Let k be an arbitrary assignment. We define $f_2(x^\sigma)(k) := \theta_{\pi_0(\tilde{k}(x^\sigma))}(x^\sigma, k(x^\sigma))$. Then by our construction of Θ, T^+ satisfies item 1-3 in Definition C.4.

To prove item 4, let k be an arbitrary assignment. And let $\phi(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}})$ be an arbitrary Σ^+-formula. Since T^+ has the uniform reduction property, let $f_3(\phi)(k)$ be the Σ-formula with free variables among $\overline{p}^{\overline{\gamma}}, k(\overline{x}^{\overline{\sigma}})$ such that:

$$T^+ \vdash [\exists \overline{x}^{\overline{\sigma}} (\phi(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}) \wedge f_2(\overline{x}^{\overline{\sigma}})(k))] \leftrightarrow f_3(\phi)(k).$$

Now we have:

$T^+ \vdash \phi(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}})$

$\leftrightarrow \bigwedge_{[k]\in[\overline{x^{\sigma}}]} \forall k(\overline{x^{\overline{\sigma}}})((f_2(\overline{x^{\overline{\sigma}}})(k)) \to \phi(\overline{x^{\overline{\sigma}}}, \overline{p^{\overline{\gamma}}}))$
$\leftrightarrow \bigwedge_{[k]\in[\overline{x^{\sigma}}]} \forall k(\overline{x^{\overline{\sigma}}})((f_2(\overline{x^{\overline{\sigma}}})(k)) \to [\exists \overline{x^{\overline{\sigma}}}(\phi(\overline{x^{\overline{\sigma}}}, \overline{p^{\overline{\gamma}}}) \wedge f_2(\overline{x^{\overline{\sigma}}})(k))])$
$\leftrightarrow \bigwedge_{[k]\in[\overline{x^{\sigma}}]} \forall k(\overline{x^{\overline{\sigma}}})((f_2(\overline{x^{\overline{\sigma}}})(k)) \to f_3(\phi)(k))$ □

Theorem C.17. *If T^+ is coded in T, then T and T^+ are Morita equivalent.*

Proof. Let T^+ be a Σ^+-theory, and $\Sigma^+ = \Sigma \cup \Sigma^*$, and let $T := T^+|_{\Sigma}$. Let f be a coding of T^+ in T. We construct a Morita descendant $T^\dagger$ of T as follows. For each $\sigma \in \Sigma^*$, we define σ as follows:

- Step 1: for each $\langle i, \overline{\sigma_i}\rangle \in f_0(\sigma)$, define a new product sort of $\overline{\sigma_i}$, which we denote as $\langle\overline{\sigma_i}\rangle$.
- Step 2: for each $\langle\overline{\sigma_i}\rangle$, define σ_i^* as a new subsort of $\langle\overline{\sigma_i}\rangle$ by the domain formula $f_3(\exists x^{\sigma}(f_2(x^{\sigma})(k)))$ where $k(\sigma) = \langle i, \overline{\sigma_i}\rangle$.
- Step 3: for each σ_i^*, define $\sigma_i^\dagger$ as a new quotient sort of σ^* by the equivalence formula which corresponds to $f_3(\exists x^{\sigma}(f_2(x^{\sigma})(k)(\overline{y^{\overline{\sigma_i}}}) \wedge f_2(x^{\sigma})(k)(\overline{z^{\overline{\sigma_i}}})))$, where $\tilde{k}(x^{\sigma}) = \langle i, \overline{\sigma_i}\rangle$.
- Step 4: define σ as the coproduct of all $\sigma_i^\dagger$.
- Step 5: Let $\pi_1, ..., \pi_n$ be the projection functions in step 1, *inc* be the inclusion function in step 2, h be the quotient function in step 3 and $q_1, ..., q_m$ the injection functions in step 4. For each $k \in \Pi_{\sigma_i\in\Sigma^*} f_0(\sigma_i)$, we define $F(x^{\sigma})(k)$ as the formula with free variables among $x^{\sigma}, k(x^{\sigma})$ as follows:
  $F(x^{\sigma})(k) := \exists x^{\langle\overline{\sigma_i}\rangle} \exists x^{\sigma^*} \exists x^{\sigma_i^\dagger}(\bigwedge_{1\leq j\leq n}[\pi_j(x^{\langle\overline{\sigma_i}\rangle}) = (k(x^{\sigma}))^j] \wedge inc(x^{\sigma_i^*}) = x^{\langle\overline{\sigma_i}\rangle} \wedge h(x^{\sigma_i^*}) = x^{\sigma_i^\dagger} \wedge q_i(x^{\sigma_i^\dagger}) = x^{\sigma})$, where k is such that $\tilde{k}(x^{\sigma}) = \langle i, \overline{\sigma_i}\rangle$.
  It can be checked that by construction, we have:
  (*): $T^\dagger \vdash f_3(\exists x^{\sigma} \exists y^{\sigma}(f_2(x^{\sigma})(k) \wedge f_2(y^{\sigma})(k) \wedge x^{\sigma} = y^{\sigma}) \leftrightarrow \exists x^{\sigma} \exists y^{\sigma}(F(x^{\sigma})(k) \wedge F(y^{\sigma})(k) \wedge x^{\sigma} = y^{\sigma}))$
  (**): $T^\dagger \vdash f_3(\exists x^{\sigma}(f_2(x^{\sigma})(k)) \leftrightarrow \exists x^{\sigma}(F(x^{\sigma})(k)));$.
- Step 6: we define $R \in \Sigma^*$ in $T^\dagger$ by adding the definition

$$\delta_R := R(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p^{\overline{\gamma}}}) \leftrightarrow \bigwedge_{[k]\in[\overline{x^{\sigma}}]} \forall k(\overline{x^{\overline{\sigma}}})(F(\overline{x^{\overline{\sigma}}})(k) \to f_3(R)(k))$$

Let $\Sigma^\dagger$ be the signature of $T^\dagger$ constructed above. We define a $\Sigma^\dagger$-Morita extension (we may denote as T') of T^+ as follows:

- Repeat Step 1-3.
- Define the injection functions as follows:
  $q_i(x^{\sigma_i^\dagger}, x^{\sigma}) \leftrightarrow \exists k(x^{\sigma}) \exists x^{\langle\overline{\sigma_i}\rangle} \exists x^{\sigma^*} \exists x^{\sigma_i^\dagger}(f_2(x^{\sigma})(k) \wedge \bigwedge_{1\leq j\leq n}[\pi_j(x^{\langle\overline{\sigma_i}\rangle}) = (k(x^{\sigma}))^j] \wedge inc(x^{\sigma_i^*}) = x^{\langle\overline{\sigma_i}\rangle} \wedge h(x^{\sigma_i^*}) = x^{\sigma_i^\dagger})$, where $k(\sigma) = \langle i, \overline{\sigma_i}\rangle$. It is easy to check that T' proves that $\{q_i\}_{1\leq i\leq m}$ defined above form a family of injection functions from $\sigma_i^\dagger$ to σ.
- Repeat Step 5.

Notice that by construction, $T' \vdash F(x^{\sigma})(k) \leftrightarrow f_2(x^{\sigma})(k)$, and thus by item 4 in Definition C.4, definitions added to $T^\dagger$ in step 6 are proved by T'. Therefore, for any $\Sigma^\dagger$-formula $\phi \in T^\dagger$, $T' \vdash \phi$.

We then prove that for any Σ^+-formula ϕ if $T' \vdash \phi$, then $T^\dagger \vdash \phi$. Note that T' is obtained by adding a set of definitions to T^+. By construction, these definitions are provable in $T^\dagger$. Thus, it suffices to show that T^+ is entailed by $T^\dagger$. We prove by induction that for any Σ^+-formula ϕ, if $T^+ \vdash \phi$, then $T^\dagger \vdash \phi$ and if $T^+ \vdash \neg\phi$, then $T^\dagger \vdash \neg\phi$.

There are two base cases. Consider the case where $R \in \Sigma^*$. For convenience, we abbreviate $\bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]}$ as $\bigwedge_{[k]}$.

$$
\begin{aligned}
&T^+ \vdash R(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}})\\
\Rightarrow &T^+ \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})((f_2(\overline{x}^{\overline{\sigma}})(k)) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[4]}\\
\Rightarrow &T^+ \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})((\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k))) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[5]}\\
\Rightarrow &T^+ \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})(f_3(\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k))) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[6]}\\
\Rightarrow &T^\dagger \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} k(\overline{x}^{\overline{\sigma}})(f_3(\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k))) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[7]}\\
\Rightarrow &T^\dagger \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} k(\overline{x}^{\overline{\sigma}})(\exists \overline{x}^{\overline{\sigma}}(F(\overline{x}^{\overline{\sigma}})(k)) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[8]}\\
\Rightarrow &T^\dagger \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})((F(\overline{x}^{\overline{\sigma}})(k)) \to f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[9]}\\
\Rightarrow &T^\dagger \vdash R(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}}).\text{[10]}
\end{aligned}
$$

Similarly, we have:

$$
\begin{aligned}
&T^+ \vdash \neg R(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p}^{\overline{\gamma}})\\
\Rightarrow &T^+ \vdash \bigvee_{[k]\in[\overline{x}^{\overline{\sigma}}]} \exists k(\overline{x}^{\overline{\sigma}})((f_2(\overline{x}^{\overline{\sigma}})(k)) \wedge \neg f_3(R(\overline{x}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k));\text{[11]}\\
\Rightarrow &T^+ \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})(\neg\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k))\\
&\qquad\qquad \vee \exists k(\overline{y}^{\overline{\sigma}})(\neg f_3(R(\overline{y}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k) \wedge \exists \overline{y}^{\overline{\sigma}} \exists \overline{x}^{\overline{\sigma}}(f_2(\overline{y}^{\overline{\sigma}})(k) \wedge f_2(\overline{x}^{\overline{\sigma}})(k) \wedge \overline{y}^{\overline{\sigma}} = \overline{x}^{\overline{\sigma}})));\text{[12]}\\
\Rightarrow &T^+ \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})(\neg f_3(\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k)))\\
&\qquad\qquad \vee \exists k(\overline{y}^{\overline{\sigma}})(\neg f_3(R(\overline{y}^{\overline{\sigma}}, \overline{p}^{\overline{\gamma}}))(k) \wedge f_3(\exists \overline{y}^{\overline{\sigma}} \exists \overline{x}^{\overline{\sigma}}(f_2(\overline{y}^{\overline{\sigma}})(k) \wedge f_2(\overline{x}^{\overline{\sigma}})(k) \wedge \overline{y}^{\overline{\sigma}} = \overline{x}^{\overline{\sigma}}))));\text{[13]}\\
\Rightarrow &T^\dagger \vdash \bigwedge_{[k]\in[\overline{x}^{\overline{\sigma}}]} \forall k(\overline{x}^{\overline{\sigma}})(\neg f_3(\exists \overline{x}^{\overline{\sigma}}(f_2(\overline{x}^{\overline{\sigma}})(k)))
\end{aligned}
$$

[4] By item 4 in Definition C.4.

[5] By item 2 in Definition C.4.

[6] By item 4 in Definition C.4.

[7] Note that this is a Σ-sentence, and $T^+|_\Sigma = T \subseteq T^\dagger$.

[8] By (**).

[9] By logic.

[10] By δ_R.

[11] By item 4 in Definition C.4.

[12] By logic.

[13] By item 4 in Definition C.4.

[14] Note that this is a Σ-sentence, and $T^+|_\Sigma = T \subseteq T^\dagger$.

$$\vee\, \exists k(\overline{y^\sigma})(\neg f_3(R(\overline{y^\sigma}, \overline{p^\gamma}))(k) \wedge f_3(\exists \overline{y^\sigma} \exists \overline{x^\sigma}(f_2(\overline{y^\sigma})(k) \wedge f_2(\overline{x^\sigma})(k) \wedge \overline{y^\sigma} = \overline{x^\sigma})))); \text{[14]}$$

$$\Rightarrow T^\dagger \vdash \bigwedge_{[k]\in[\overline{x^\sigma}]} \forall k(\overline{x^\sigma})(\neg \exists \overline{x^\sigma}(F(\overline{x^\sigma})(k))$$

$$\vee\, \exists k(\overline{y^\sigma})(\neg f_3(R(\overline{y^\sigma}, \overline{p^\gamma}))(k) \wedge \exists \overline{y^\sigma} \exists \overline{x^\sigma}(F(\overline{y^\sigma})(k) \wedge F(\overline{x^\sigma})(k) \wedge \overline{y^\sigma} = \overline{x^\sigma}))); \text{[15]}$$

$$\Rightarrow T^\dagger \vdash \bigvee_{[k]\in[\overline{x^\sigma}]} \exists k(\overline{x^\sigma})((F(\overline{x^\sigma})(k)) \wedge \neg f_3(R(\overline{x^\sigma}, \overline{p^\gamma}))(k)); \text{[16]}$$

$$\Rightarrow T^\dagger \vdash \neg R(x^{\sigma_1}, ..., x^{\sigma_n}, \overline{p^\gamma}).\text{[17]}$$

For the case where $R \in \Sigma$, the conclusion holds since $T^+|_\Sigma = T \subseteq T^\dagger$. And the inductive cases are all straightforward.

Now, since $T' \equiv T^\dagger$, T^+ and T have a common Morita descendant, and hence are Morita equivalent. □

Theorem C.18. *Let T^+ be a Σ^+-theory, $\Sigma \subseteq \Sigma^+$, and $T := T^+|_\Sigma$. The following are equivalent:*

1. *T^+ is implicitly definable in T;*
2. *T^+ and T are Morita equivalent;*

Proof. By Theorem C.15, Theorem C.16, and Theorem C.17, item 1 implies item 2.

For the other direction, assume that T^+ and T have a common Morita descendant T^*. Suppose (towards a contradiction) that T^+ is not implicitly definable in T. Then there is a model $\mathcal{M}$ of T which has two non-isomorphic expansions $\mathcal{M}_1$ and $\mathcal{M}_2$ in $Mod(T^+)$. By Theorem 4.2 in Barrett and Halvorson (2016), every model of T and of T^+ has a unique expansion in $Mod(T^*)$ up to isomorphism. Let $\mathcal{M}_1^*$ be the expansion of $\mathcal{M}_1$ in $Mod(T^*)$, and $\mathcal{M}_2^*$ be the expansion of $\mathcal{M}_2$ in $Mod(T^*)$. Then $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$ are two non-isomorphic expansions of $\mathcal{M}$ in $Mod(T^*)$. Contradiction. Therefore, T^+ is implicitly definable in T. □

[15] By (*) and (**).

[16] Item 1 in Definition C.4 also holds with F in place of f_2.

[17] By δ_R.

Bibliography

Ainsworth, Peter M., 2009, “Newman’s Objection.” *British Journal for the Philosophy of Science* 60, no. 1: 135-171.

Andréka, Hajnal, Judit X. Madarász, and Istvan Németi , 2008, *Mathematical Institute of the Hungarian Academy of Sciences*, Budapest, 93. URL =
<https://old.renyi.hu/pub/algebraic-logic/kurzus10/amn-defi.pdf>

Arledge, Christopher, and Robert Rynasiewicz, 2019, “On Some Recent Attempted Non-Metaphysical Dissolutions of the Hole Dilemma.” URL =
<http://philsci-archive.pitt.edu/16343/>.

Arnold, V.I., 1989, *Mathematical Methods of Classical Mechanics*, 2nd edition. New York: Springer.

Barrett, Thomas William, and Hans Halvorson, 2016, “Morita Equivalence.” *The Review of Symbolic Logic* 9, no. 3 (September 2016): 556-82.

Barrett, Thomas William, and Hans Halvorson, 2022, “Mutual Translatability, Equivalence, and the Structure of Theories.” *Synthese* 200, no. 3: 240.

Beatty, John, 1981, “What’s Wrong with the Received View of Evolutionary Theory?” In *Philosophy of Science* (Proceedings), vol. 2, edited by P. Asquith and R. Giere, 397-426. East Lansing: Philosophy of Science Association.

Belot, Gordon, 1995, “Determinism and Ontology.” *International Studies in the Philosophy of Science* 9, no. 1 (March 1, 1995): 85-101.

Belot, Gordon, 2018, “Fifty Million Elvis Fans Can’t Be Wrong.” *Noûs* 52, no. 4: 946-81.

Beni, Majid Davoody, 2019, *Cognitive Structural Realism: A Radical Solution to the Problem of Scientific Representation.* Vol. 14. Studies in Brain and Mind. Cham: Springer International Publishing.

Born, Max, 1926, “Quantenmechanik der Stobvurgnge”, in *Zeitschrift für Physik*, 38, 803-827.

Bradley, Clara, and James Owen Weatherall, 2022, “Mathematical Responses to the Hole Argument: Then and Now.” *Philosophy of Science* 89, no. 5 (December 2022): 1223-32.

Butterfield, Jeremy, 1989, “The Hole Truth.” *British Journal for Philosophy of Science* 40 (1):1-28.

Cameron, Ross P., 2008, “Truthmakers and Ontological Commitment: Or How to Deal with Complex Objects and Mathematical Ontology Without Getting into Trouble.” *Philosophical Studies* 140 (1): 1-18.

Carnap, Rudolf, 1958, “Beobachtungssprache und theoretische Sprache”, *Dialectica*, 12(3-4): 236-248; translation: Rudolf, Carnap, 1975, “Observational Language and Theoretical Language”, in Rudolf Carnap, *Logical Empiricist*, J. Hintikka (ed.), Dordrecht: D. Reidel Publishing Company, Dordrecht: 75-85.

Carnap, Rudolf, 1966, *Philosophical Foundations of Physics: An Introduction to the Philosophy of Science.* Edited by Martin Gardner. New York: Basic Books.

Caulton, Adam, and Jeremy Butterfield, 2012, "Symmetries and Paraparticles as a Motivation for Structuralism." *British Journal for the Philosophy of Science* 63 (2): 233-285.

Chang, C. C., and H. J. Keisler. 1990. *Model Theory.* 3rd ed. Amsterdam: North-Holland.

Chang, Hasok. 2012. *Is Water H O? Evidence, Realism and Pluralism.* Dordrecht: Springer.

Christensen, David, 1991, "Clever Bookies and Coherent Beliefs." *The Philosophical Review*, 100 (2): 229-47.

Coffey, Kevin, 2014, "Theoretical Equivalence as Interpretative Equivalence." *The British Journal for the Philosophy of Science* 65, no. 4: 821-44.

Cordero, Alberto, 2011, "Scientific Realism and the Pessimistic Meta-Modus Tollens." In *Recent Themes in the Philosophy of Science: Scientific Realism and Commonsense*, edited by S. Clarke and T. D. Lyons, 63-90. Dordrecht: Springer.

Cudek, Franciszek, 2024, "Counterparts, Determinism, and the Hole Argument." *The British Journal for the Philosophy of Science*, in press.

Contessa, Gabriele, 2007, "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74 (1): 48-68.

Dasgupta, Shamik, 2011, "The Bare Necessities." *Philosophical Perspectives* 25, no. 1: 115-60.

Dasgupta, Shamik, 2014, "On the Plurality of Grounds." *Philosophers' Imprint* 14: 1-28.

Dasgupta, Shamik, 2016, "Quality and Structure." In *Current Controversies in Metaphysics*, edited by Elizabeth Barnes, 7–23. London: Routledge.

De Haro, Sebastian, and Jeremy Butterfield, 2021, "On Symmetry and Duality", *Synthese* 198, no. 4: 2973-3013.

De Haro, Sebastian, and Jeremy Butterfield, forthcoming, *The Philosophy and Physics of Duality.* Oxford: Oxford University Press.

Demopoulos, William and Michael Friedman, 1985, "Critical Notice: Bertrand Russell's The Analysis of Matter: Its Historical Context and Contemporary Interest." *Philosophy of Science*, 52(4): 621-639.

Dewar, Neil, 2015, "Symmetries and the philosophy of language." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 52 (Part B):317-27.

Dewar, Neil, 2019a, "Sophistications about Symmetries." *The British Journal for the Philosophy of Science* 70, no. 2: 485-521.

Dewar, Neil, 2019b, "Ramsey Equivalence." *Erkenntnis* 84, no. 1: 77-99.

Dorr, Cian, 2008, "There Are No Abstract Objects." In *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman, 32-63. Oxford: Blackwell.

Earman, John, 1989, *World Enough and Space-Time: Absolute versus Relational Theories of Space and Time.* Cambridge, MA: MIT Press.

Earman, John, and John Norton, 1987, "What Price Spacetime Substantivalism? The Hole Story." *The British Journal for the Philosophy of Science* 38, no. 4: 515-25.

Feyerabend, Paul, 1975, *Against Method*, New York: New Left Books.

Fine, Kit, 2011, "What is metaphysics?" In Tuomas E. Tahko (ed.), *Contemporary Aristotelian Metaphysics.* Cambridge: Cambridge University Press, 8-25.

Fletcher, Samuel C, 2020, "On Representational Capacities, with an Application to General Relativity." *Foundations of Physics* 50, no. 4: 228-49.

French, Steven, and Décio Krause, 2006, *Identity in physics: a historical, philosophical, and formal analysis.* New York: Oxford University Press.

French, Steven, 2020, *There Are No Such Things As Theories.* Oxford: Oxford University Press.

Friedman, Michael, 1983, *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science.* Princeton, NJ: Princeton University Press.

Frigg, Roman, 2022, *Models and Theories: A Philosophical Inquiry.* London: Routledge.

Frigg, Roman, and Ioannis Votsis, 2011, "Everything You Always Wanted to Know About Structural Realism but Were Afraid to Ask." *European Journal for Philosophy of Science* 1 (2): 227–276.

Galilei, Galileo, 1967, *Dialogue Concerning the Two Chief World Systems: Ptolemaic and Copernican.* Translated by Stillman Drake. 2nd ed. Berkeley: University of California Press.

Geroch, Robert, 1969, "Limits of spacetimes," *Communications in Mathematical Physics*,13(3): 180-93.

Giacomini, Flaminia, and Časlav Brukner, 2023, "Einstein's Equivalence Principle for Superpositions of Gravitational Fields and Quantum Reference Frames." *arXiv*,
URL = <http://arxiv.org/abs/2012.13754>.

Gomes, Henrique, and Jeremy Butterfield, 2023a, "The Hole Argument and Beyond: Part I: The Story so Far." *Journal of Physics: Conference Series 2533*, no. 1: 012002.

Gomes, Henrique, and Jeremy Butterfield, 2023b, "The Hole Argument and Beyond: Part II: Treating Non-Isomorphic Spacetimes." *Journal of Physics: Conference Series 2533*, no. 1: 012003.

Goodman, Nelson, 1955, *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press.

Glick, David, 2020, "Generalism and the Metaphysics of Ontic Structural Realism." *The British Journal for the Philosophy of Science* 71 (2): 751-72.

Glymour, Clark, 1970, "Theoretical Realism and Theoretical Equivalence." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*: 275-88.

Glymour, Clark, 1977, "The Epistemology of Geometry." *Noûs* 11, no. 3: 227-51.

Glymour, Clark, 1980, *Theory and Evidence.* Princeton: Princeton University Press.

Glymour, Clark, 2013, "Theoretical Equivalence and the Semantic View of Theories." *Philosophy of Science* 80, no. 2: 286-97.

Griffiths, Paul and Karola Stotz, 2013, *Genetics and Philosophy: An Introduction*, Cambridge: Cambridge University Press.

Hall, Ned, Brian Rabern, and Wolfgang Schwarz, 2024, "David Lewis's Metaphysics." In *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), edited by Edward N. Zalta and Uri

Nodelman. Stanford, CA: Metaphysics Research Lab, Stanford University. URL = <https://plato.stanford.edu/archives/spr2024/entries/lewis-metaphysics/>.

Hall, Geoffrey, and Sebastián Murgueitio Ramírez, 2024, “Symmetries and Representation.” *Philosophy Compass* 19, no. 3: e12971.

Halvorson, Hans, 2012, “What Scientific Theories Could Not Be.” *Philosophy of Science* 79, no. 2: 183-206.

Halvorson, Hans, 2013, *The semantic view, if plausible, is syntactic*, *Philosophy of Science*, 80(3): 475-478.

Halvorson, Hans, 2019, *The Logic in Philosophy of Science*, Cambridge: Cambridge University Press.

Halvorson, Hans, and John Byron Manchak, 2022, “Closing the Hole Argument.” *The British Journal for the Philosophy of Science*, in press.

Hawking S.W., Ellis, G.F.R., 1973, *The Large Scale Structure of Space-Time.* Cambridge: Cambridge University Press.

Hirsch, Morris W., 1976, *Differential Topology.* Graduate Texts in Mathematics, vol. 33. New York: Springer.

Hodges, Wilfrid, 1993, *Model Theory.* Encyclopedia of Mathematics and its Applications, vol. 42. Cambridge: Cambridge University Press.

Hoefer, Carl, 1996, “The Metaphysics of Space-Time Substantivalism.” *The Journal of Philosophy* 93, no. 1: 5-27.

Hoefer, Carl, Nick Huggett, and James Read, 2021, “Absolute and Relational Theories of Space and Motion.” In *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), edited by Edward N. Zalta and Uri Nodelman. URL =
<https://plato.stanford.edu/archives/sum2021/entries/spacetime-theories/>.

Hudetz, Laurenz, 2019, “The Semantic View of Theories and Higher-Order Languages.” *Synthese* 196, no. 3 (March 1, 2019): 1131-49.

Hughes, R. I. G, 1997, “Models and Representation.” *Philosophy of Science* 64 (Supplement): S325-S336.

Jacobs, Caspar, 2023, “Are Models Our Tools Not Our Masters?” *Synthese* 202 (4): 121.

Jacobs, Caspar, 2024, “Some Neglected Possibilities: A Reply to Teitel.” *The Journal of Philosophy* 121, no. 2: 108-20.

Kellert, Stephen H., Helen E. Longino, and C. Kenneth Waters, eds., 2006, *Scientific Pluralism.* Minnesota Studies in the Philosophy of Science, vol. 19. Minneapolis: University of Minnesota Press.

Ketland, Jeffrey, 2004, “Empirical Adequacy and Ramsification.” *The British Journal for the Philosophy of Science* 55, no. 2: 287-300.

Kitcher, Philip, 1993, *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*, Oxford: Oxford University Press.

Kleinschmidt, Shieva, 2015, “Brute Facts.” *Philosophical Studies* 172 (1): 1-17.

Ladyman, James, and Don Ross, 2007, *Every Thing Must Go: Metaphysics Naturalized.* New York: Oxford University Press.

Ladyman, James, and Stuart Presnell, 2020, "The Hole Argument in Homotopy Type Theory." *Foundations of Physics* 50, no. 4: 319-29.

Laudan, Larry, 1981, "A Confutation of Convergent Realism." *Philosophy of Science*, 48(1): 19-49.

Landsman, Klaas, 2023, "Reopening the Hole Argument." *Philosophy of Physics*, no. 1.

Leibniz, Gottfried Wilhelm, and Samuel Clarke, 1956, *The Leibniz-Clarke Correspondence: Together with Extracts from Newton's Principia and Opticks.* Edited with introduction and notes by H. G. Alexander. Manchester: Manchester University Press.

Lefever, Koen, and Gergely Székely, 2019, "On Generalization of Definitional Equivalence to Non-Disjoint Languages." *Journal of Philosophical Logic* 48 (6): 709-729.

Lewis, David, 1966, "An Argument for the Identity Theory." *Journal of Philosophy*, 63: 17-25.

Lewis, David, 1968, "Counterpart Theory and Quantified Modal Logic." *The Journal of Philosophy* 65 (5): 113-126.

Lewis, David, 1970, "How to Define Theoretical Terms." *Journal of Philosophy*, 67: 427-46.

Lewis, David, 1972, "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50: 249-58.

Lewis, David, 1983, "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343-77.

Lewis, David, 1984, "Putnam's Paradox." *Australasian Journal of Philosophy* 62 (3): 221-236.

Lewis, David, 1994, "Reduction of Mind." In *A Companion to Philosophy of Mind*, edited by Samuel Guttenplan, 412-31. Oxford: Blackwell Publishers.

Lewis, David, 1997, "Naming the Colours." *Australasian Journal of Philosophy*, 75: 325-342.

Libkin, Leonid, 2004, *Elements of Finite Model Theory.* Texts in Theoretical Computer Science. An EATCS Series. Berlin: Springer.

Longino, Helen E., 1987, "Can There Be A Feminist Science?", *Hypatia*, 2(3): 51-64.

Longino, Helen E., 1990, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*, Princeton: Princeton University Press.

Luc, Joanna, 2022, "Arguments from Scientific Practice in the Debate about the Physical Equivalence of Symmetry-Related Models." *Synthese* 200, no. 2: 72.

Luc, Joanna, 2024, "The Hole Argument without the Notion of Isomorphism." *Synthese* 203, no. 3: 101.

Ludwig, David, and Stéphanie Ruphy, 2024, "Scientific Pluralism." In *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), edited by Edward N. Zalta and Uri Nodelman. URL= <https://plato.stanford.edu/archives/fall2024/entries/scientific-pluralism/>.

Lutz, Sebastian, 2017, "What Was the Syntax-Semantics Debate in the Philosophy of Science About?" *Philosophy and Phenomenological Research* 95, no. 2: 319-52.

Lyons, Timothy D., 2006, ”Scientific Realism and the Stratagema de Divide et Impera.” *British Journal for the Philosophy of Science* 57 (3): 537-60.

Mach, Ernst, 1893, *The Science of Mechanics: A Critical and Historical Exposition of Its Principles.* Translated by Thomas J. McCormack. Chicago: The Open Court Publishing Co.

Manzano, María, 1996, *Extensions of First-Order Logic.* Cambridge Tracts in Theoretical Computer Science 19. Cambridge: Cambridge University Press.

Martin, Ben, and Ole Thomassen Hjortland, 2021, “Logical Predictivism.” *Journal of Philosophical Logic* 50 (2): 285-318.

Maudlin, Tim, 1988, “The Essence of Space-Time.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1988*: 82-91.

Maudlin, Tim, 1990, “Substances and Space-Time: What Aristotle Would Have Said to Einstein.” *Studies in History and Philosophy of Science Part A* 21 (4): 531-561.

Meadows, Toby, 2024, “Beyond Linguistic Interpretation in Theory Comparison.” *The Review of Symbolic Logic* 17, no. 3: 819-59.

Melia, Joseph, 1999, “Holes, Haecceitism and Two Conceptions of Determinism.” *The British Journal for the Philosophy of Science* 50, no. 4: 639-64.

Meigniez, Gaël, 2002, “Submersions, Fibrations and Bundles.” *Transactions of the American Mathematical Society* 354 (9): 3771-87.

Melia, Joseph, and Juha Saatsi, 2006, “Ramseyfication and Theoretical Content.” *British Journal for the Philosophy of Science* 57, no. 3: 561-585.

Mundy, Brent, 1992, “Space-Time and Isomorphism.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992*: 515-27.

Newton, Isaac, 2004, *Philosophical Writings.* Edited by Andrew Janiak. Cambridge: Cambridge University Press.

Newton-Smith, W.H., 1981, *The Rationality of Science*, London: Routledge & Kegan Paul.

Nguyen, James, 2017, “Scientific Representation and Theoretical Equivalence.” *Philosophy of Science* 84, no. 5: 982-95.

Norton, John, Oliver Pooley, and James Read, 2023, “The Hole Argument”, *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman,
URL = <https://plato.stanford.edu/archives/sum2023/entries/spacetime-holearg/>.

Papineau, David, 2023, “Naturalism”, *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman,
URL = <https://plato.stanford.edu/archives/fall2023/entries/naturalism/>.

Pooley, Oliver, 2006, “Points, particles, and structural realism.” In *The Structural Foundations of Quantum Gravity*, edited by Dean Rickles, Steven French and Juha T. Saatsi, 83-120. Oxford: Oxford University Press.

Pooley, Oliver, 2013, “Substantivalist and Relationalist Approaches to Spacetime.” In *The Oxford Handbook of Philosophy of Physics*, edited by Robert Batterman, 522–586. Oxford: Oxford University Press.

Pooley, Oliver, 2021, “The Hole Argument”, in E. Knox and A. Wilson (eds.), *The Routledge Companion to Philosophy of Physics*, London: Routledge, 145-59.

Pooley, Oliver, and James Alexander Mabyn Read, 2021, “On the Mathematics and Metaphysics of the Hole Argument.” *The British Journal for the Philosophy of Science*, in press.

Popper, Karl R., 1972, *Conjectures and Refutations: The Growth of Knowledge.* 4th edition, London: Routledge & Kegan Paul.

Priest, Graham, 2014, “Revising Logic.” In *The Metaphysics of Logic*, edited by Penelope Rush, 211-223. Cambridge: Cambridge University Press.

Psillos, Stathis, 1994, “A philosophical study of the transition from the caloric theory of heat to thermodynamics: Resisting the pessimistic meta-induction.” *Studies in the History and Philosophy of Science*, 25(2): 159-90.

Psillos, Stathis, 1999, *Scientific Realism: How Science Tracks Truth*, London: Routledge.

Psillos, Stathis, 2009, *Knowing the Structure of Nature: Essays on Realism and Explanation*, London: Palgrave/MacMillan.

Putnam, Hilary, 1980, “Models and Reality.” *The Journal of Symbolic Logic* 45 (3): 464-482.

Quine, Willard van Orman, 1975, “On empirically equivalent systems of the world.” *Erkenntnis* 9 (3):313-28.

Ramsey, Frank Plumpton, 1929, “Theories.” In *The Foundations of Mathematics and Other Logical Essays*, edited by R. B. Braithwaite, with a preface by G. E. Moore, 212-236. London: Kegan Paul, Trench, Trubner & Co.

Rynasiewicz, Robert. 1994. ”The Lessons of the Hole Argument.” *The British Journal for the Philosophy of Science* 45 (2): 407-436.

Rynasiewicz, Robert, 1996, “Is There a Syntactic Solution to the Hole Problem?” *Philosophy of Science* 63, no. 5: S55-62.

Rynasiewicz, Robert, 2019, “Newton’s Scholium on Time, Space, Place and Motion.” In *The Oxford Handbook of Isaac Newton*, edited by Eric Schliesser and Chris Smeenk, 478–500. Oxford: Oxford University Press.

Roberts, Bryan, 2020, “Regarding ‘Leibniz Equivalence.’” *Foundations of Physics* 50, no. 4 (April 1, 2020): 250-69.

Ruphy, Stéphanie, 2010, “Are Stellar Kinds Natural Kinds? A Challenging Newcomer in the Monism/Pluralism and Realism/Antirealism Debates.” *Philosophy of Science*, 77(5): 1109-1120.

Russell, Jeffrey Sanford, 2014, “On Where Things Could Be.” *Philosophy of Science* 81, no. 1: 60-80.

Shulman, Michael, 2017, “Homotopy Type Theory: A Synthetic Approach to Higher Equalities.” In *Categories for the Working Philosopher*, edited by Elaine Landry, 36-57. Oxford: Oxford University Press.

Sider, Theodore, 2011, *Writing the Book of the World.* Oxford: Oxford University Press.

Sider, Theodore, 2020, *The Tools of Metaphysics and the Metaphysics of Science.* Oxford: Oxford University Press.

Sneed, Joseph D., 1971, *The Logical Structure of Mathematical Physics.* Dordrecht: D. Reidel Publishing Company.

Sklar, Lawrence, 1982, “Saving the Noumena.” *Philosophical Topics* 13 (1): 89-110.

Stachel, John, 1989, “Einstein’s Search for General Covariance, 1912-1915”, in Don Howard and John Stachel (eds), *Einstein and the History of General Relativity*, Boston, MA: Birkhauser, 62-100.

Stachel, John, 1993, “The Meaning of General Covariance: The Hole Story.” In *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grünbaum*, edited by John Earman, Allan I. Janis, Gerald J. Massey, and Nicholas Rescher, 129-160. Pittsburgh: University of Pittsburgh Press.

Stachel, John, 2002, “‘The Relations Between Things’ versus ‘The Things Between Relations’: The Deeper Meaning of the Hole Argument.” In *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*, edited by David B. Malament, 231-266. Chicago: Open Court.

Stein, Howard, 2002, “Newton’s Metaphysics.” In *The Cambridge Companion to Newton*, edited by I. Bernard Cohen and George E. Smith, 256-307. Cambridge: Cambridge University Press.

Suárez, Mauricio, 2004, “An Inferential Conception of Scientific Representation.” *Philosophy of Science* 71 (5): 767-79.

Suppe, Frederick, 1977, “The Search for Philosophical Understanding of Scientific Theories.” In *The Structure of Scientific Theories*, edited by Frederick Suppe, 3-241. Urbana and Chicago: University of Illinois Press.

Suppes, Patrick, 1957, *Introduction to Logic.* New York: D. Van Nostrand. Reprint, Mineola, NY: Dover Publications, 1999.

Suppes, Patrick, 2002, *Representation and Invariance of Scientific Structures.* Stanford, CA: CSLI Publications.

Talbott, William J., 1991, “Two Principles of Bayesian Epistemology.” *Philosophical Studies*, 62 (2): 135-50.

Taylor, John R., 2005, *Classical Mechanics.* Sausalito, CA: University Science Books.

Teitel, Trevor, 2019, “Holes in Spacetime: Some Neglected Essentials.” *The Journal of Philosophy* 116, no. 7: 353-89.

Teitel, Trevor, 2021, “What Theoretical Equivalence Could Not Be.” *Philosophical Studies* 178 (12): 4119-49.

Teitel, Trevor, 2022, “How to Be a Spacetime Substantivalist.” *The Journal of Philosophy* 119, no. 5: 233-78.

Tu, Loring W, 2011, *An Introduction to Manifolds.* New York, NY: Springer New York.

Tsementzis, Dimitris, 2017, “A Syntactic Characterization of Morita Equivalence.” *The Journal of Symbolic Logic* 82, no. 4: 1181-98.

van Fraassen, Bas C., 1970, “On the Extension of Beth’s Semantics of Physical Theories.” *Philosophy of Science* 37, no. 3: 325-39.

van Fraassen, Bas C., 1972, “A Formal Approach to the Philosophy of Science.” In *Paradigms and Paradoxes: The Philosophical Challenge of the Quantum Domain*, edited by R. Colodny, 303-66. Pittsburgh: University of Pittsburgh Press.

van Fraassen Bas, C., 1980, *The scientific image*, New York: Oxford University Press.

Van Fraassen, Bas C., 1984, “Belief and the Will.” *The Journal of Philosophy* 81 (5): 235-56.

van Fraassen, Bas C., 2014, “One or Two Gentle Remarks about Hans Halvorson’s Critique of the Semantic View.” *Philosophy of Science* 81, no. 2: 276-83.

Visser, Albert, 2004, “Categories of theories and interpretations”, Logic Group Preprint Series, 228.

Visser, Albert, 2009, “Why the theory R is special ”, *Logic Group preprint series*, 279.

Visser, Albert, 2021, “Extension and Interpretability.” In *Mathematics, Logic, and their Philosophies*, edited by Mojtaba Mojtahedi, Shahid Rahman, and Mohammad Saleh Zarepour, 57 - 87. Cham: Springer.

Votsis, Ioannis, 2003, “Is structure not enough?” *Philosophy of Science*, 70(5), 879-890.

Weatherall, James Owen, 2016a, “Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent?” *Erkenntnis* 81 (5):1073-1091.

Weatherall, James Owen, 2016b, “Understanding Gauge”, *Philosophy of Science* 83, no. 5: 1039-49.

Weatherall, James Owen, 2017, “Categories and the Foundations of Classical Space-Time Theories.” In *Categories for the Working Philosopher*, edited by Elaine M. Landry, 329-48, Oxford: Oxford University Press.

Weatherall, James Owen, 2018, “Regarding the ‘Hole Argument.’” *The British Journal for the Philosophy of Science* 69, no. 2: 329-50.

Weatherson, Brian, 2021, “David Lewis.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Winter 2021 Edition. Stanford University. URL =
<https://plato.stanford.edu/archives/win2021/entries/david-lewis/.>

Wald, Robert M, 1984, *General Relativity.* Chicago: University of Chicago Press.

Williamson, Timothy, 2007, *The Philosophy of Philosophy.* Malden, MA: Wiley-Blackwell.

Williamson, Timothy, 2013, *Modal Logic as Metaphysics.* Oxford: Oxford University Press.

Williamson, Timothy, 2017, “Semantic Paradoxes and Abductive Methodology.” In *Reflections on the Liar*, edited by Bradley P. Armour-Garb, 325-46. Oxford: Oxford University Press.

Winther, Rasmus Grønfeldt, 2020, *When Maps Become the World*, Chicago: University of Chicago Press.

Zahar, Elie G., 2004, “Ramseyfication and Structural Realism.” *Theoria* 19, no. 1: 5-30.