Utrecht University

Master's Thesis

Abstract Model Theory for Logical Metascience

Author:
Bobby Vos

Supervisor:
Prof. Dr. F.A. Muller
Second reader:
Dr. J. van Oosten

July 8, 2017

Preface

The thesis laying before the reader in many ways represents a mile stone in a larger project which has occupied me for the entirety of my adult life. The origins of this project may be traced to my *Profielwerkstuk* for secondary school. Initially planning to devote this magnum opus of my high-school career to a topic of a purely mathematical nature, this intent was foiled by a policy of the school which required me to combine two different subjects in this project. The combination in question did not require much thought, as I had a longstanding interest in both physics and mathematics. Surely, there was no shortage of potential topics within this intersection. However, placing a perhaps excessive amount of value on the originality of my personal projects, my wish became to pioneer a new way in which to combine the two subjects. More specifically, I hoped to lay the groundwork of a 'new physics', based on a branch of mathematics not previously applied to the study of nature.

The end product of my high-school investigations did not quite live up to this rather lofty goal. Nevertheless, the ambition to formulate a physics grounded in an alternative mathematical language remained within my consciousness as I proceeded with my bachelor studies in physics and mathematics. Being exposed to more and more different subfields of mathematics and increasingly recognizing the extent to which the conceptual structure of contemporary physics was linked to its underlying mathematical formalisms, it did not take long for these dormant aspirations to reawaken.

It was not until my first encounter with the subject of logic, however, that my eyes were opened to a new dimension of my 'grand project'. The discipline of model theory, in particular, proved to be a definitive influence on the evolution of my academic interests. With the relation between formal languages and mathematical structures being at the locus of model-theoretic investigations, it did not take long for me to recognize the potential this field of study had to offer for investigating the relation between mathematical languages and physical systems (essentially moving the framework of model theory one rung down on the ladder of abstraction). Against the backdrop of my interest in 'alternative physics', it was now but a small step to formulating higher-level questions: *What is a scientific theory, construed as generally as possible? How indispensable is mathematics, and specifically calculus, to the study of nature? And how can we apply the framework of model theory to answer such questions?* These queries eventually led me to become interested in the history, logic and philosophy of science and have been at the center of spare-time readings and course papers ever since, also inspiring the topic of my bachelor's thesis.

With my interests in physics shifted to the philosophical end of the spectrum and my mathematical interests increasingly focused in the realm of logic, the direction of my future studies seemed clear. It was through these sequence of events that I would eventually find myself at Utrecht University, where I would take up a double master's program in History and Philosophy of Science on the one hand and Mathematical Sciences (with a specialization in logic) on the other. Here, I experienced an unprecedented amount of academic freedom, which I eagerly employed for the further advancement of my ever-evolving grand project. The culmination of this latest stage of my academic development may be found in the pages of the following master's thesis. With the exception of the historical dimension of my interest in 'alternative physics', which includes topics such as the Scientific Revolution, the mathematization of nature and both western and non-western schools of natural philosophy, the present text provides the reader with a reasonably comprehensive overview of the cerebrations that have occupied me for the last number of years.

I would like to thank my supervisors F.A. Muller and Jaap van Oosten for supervising such an idiosyncratic research project. Moreover, I would like to express gratitude to David Baneke for being an extremely kind and responsive study advisor during these past years and for assisting me with the administrative formalities concerning my double master's program.

It is with some sadness that I see the end of my time at Utrecht University approaching. Both academically and socially, Utrecht provided me with a homely environment of which I am glad to have been a part for the duration of my master studies. However, I am most happy to report that during the writing of this thesis, I have succeeded in obtaining a PhD position at the History and Philosophy of Science department of the University of Cambridge. I have no doubt that Cambridge will prove to be an extremely stimulating environment and I am certain I will find myself well poised for taking the next steps in my ongoing quest to understand the fundamental nature of the scientific enterprise.

Contents

Chapter 1

Prolegomena

The present chapter serves to lay out the main issues, claims and methodology that will be of concern to us throughout this thesis. In section 1.1, I argue that the field of 'metascience' is plagued by recurring methodological difficulties, which may be solved by turning to a recent stream of logical research. Subsequently, I outline the manner in which I shall unfold this argument over the course of the thesis. Next, in section 1.2, we will take a moment to reflect upon the notion of metascience and will specify which fields of study are referred to exactly by this denomination. Lastly, in section 1.3, we will discuss some preliminary notions from many-sorted first-order logic, which will be required in later chapters.

1.1 Introduction

The *meta* prefix, perhaps more so than any other of its linguistic relatives, possesses an inherent, enigmatic appeal. The act of prepending it to even the simplest of terms can open up fascinating avenues of thought. A shining example demonstrating the latter possibility is found in the study of metamathematics. It is with this example that we encounter a helpful tool in such metatheoretical investigations, viz. the discipline of logic. As is well-known, the application of logic to metamathematics has led to many results of great conceptual and technical interest, such as Gödel's incompleteness theorem.

Enamored with the early applications of logic to metamathematics during the first half of the twentieth century, the group of philosophers known as the *logical positivists* concurrently set about giving a logical formalization[1] of the

[1] Throughout this thesis, I shall employ the term *formal* and derived expressions to refer to any type of methodology or entity explicated in a rigorous, precise and (often) mathematical fashion. In particular, *formal* in general does not necessarily refer to the usage of logic and syntactic expressions as a means of explication. The only exception to this rule will be my usage of the term *formal language*, which will be used to refer to languages consisting of meaningless, syntactic symbols and statements.

natural sciences and hence **metascience** was born.[2] Despite the positivist project's eventual demise, the notion of a formally expounded metascience remained alive in the collective consciousness of the philosophy-of-science community. The twentieth century saw a host of approaches to the formalization of science employing a variety of different formalisms. Eventually, an approach was established relying only on tools from set theory, and logic was displaced to the background. Thus, while the development of *logical* metascience had all but stagnated, metascience as a whole still received a considerable amount of attention. Yet, in spite of much effort, we have yet to formulate a metascience that is on par with metamathematics in terms of fruitfulness and rigor. This has led some philosophers of science to question the tenability of such formal research programs altogether, cf. (Contessa 2006, 376).

This thesis forms part of a wider project, first described in (Vos 2015b), to reappraise metascience as a whole and logical metascience in particular. To accomplish this, I hold we must first identify a crucial flaw of traditional metascientific research.

Problem (First-Order Fixation)
The traditional approaches to metascience rely on an inappropriate conception of logic as a metatheoretical formalism, i.e. a conception which largely identifies the use of logic with the use of first-order logic.

This conflation, I hold, leads to awkward and unfruitful formal frameworks for the analysis of science. In slogan form, we might say the traditional approaches to metascience suffer from **first-order fixation**.[3,4]

Associated to this perceived problem, I of course also propagate a solution. In line with the latest developments in the field known as *universal logic*, we can adopt a more abstract view of the notion of *logic*, by identifying *a logic* with a certain type of mathematical structure. Broadening or view of logic in this manner, so I claim, holds great potential for metascience:

Solution (Logical Abstractivism)
By adopting a more abstract view of logic, in the sense of universal logic, we can circumvent many flaws of the traditional approaches and potentially arrive at much more fruitful frameworks for metascience.

[2] The usage of the term *metascience* here is a personal idiosyncrasy and does not seem to have caught on among philosophers of science. To my knowledge, the term has only seen prominent usage in the work of David Pearce and Veikko Rantala (1983a) and, to a lesser extent, Mario Bunge (1959).

[3] As has been noted by Lutz (2012, 80–3), the oft-heard claim that the logical positivists relied exclusively on first-order logic is historically false. Accordingly, the term *first-order fixation* should not be taken to denote an *exclusive* preoccupation with the formalism of first-order logic, but rather a preoccupation with formalisms *inspired* by first-order logic, such as higher-order logic or some other extensions of first-order logic.

[4] Alternatively, *predicate predilection* or *classical constipation*.

This conviction, i.e. that an abstract conception of logic may prove valuable for metascience, I have dubbed **logical abstractivism**. This, by itself, is not an entirely new idea. In fact, it has been directly inspired by the joint work of the Finnish philosopher-logician Veikko Rantala (b. 1933) and the British philosopher-logician David Pearce (b. 1952). From the late 1970s to the early 2000s, Pearce and Rantala argued in a number of papers and monographs for the application of *abstract logics* to the branch of metascience known as *structuralism.* I shall refer to this body of work as the **first wave of logical abstractivism (FWLA)**. This first wave, however, does not seem to have made a large splash in the philosophy-of-science community and remains very much at the periphery of metascientific investigations.

In this thesis, I shall not be concerned with substantiating the supposed first-order fixation of metascience and its harmfulness to the discipline. Rather, the focus here will lie with assessing the reasons for the failure of the FWLA and discussing a novel way in which we might apply methods from universal logic to metascience, laying the groundwork for what may eventually become a **second wave of logical abstractivism (SWLA)**, and showing how this can help us improve upon several important deficiencies within the first wave. More specifically, I will argue that two comparatively recent, highly abstract strands of research within universal logic can help give new impetus to the formal study of science. We can thus express the central claim of this thesis as follows.

Central Claim
The first wave of logical abstractivism has failed to, and is not well-suited to, make a significant impact on metascience. However, a second, more successful wave of logical abstractivism may be initiated by employing newer, more abstract subfields of universal logic.

Now, in expounding this claim, it is essential to first establish what the extant approaches to metascience actually are. This is not a trivial task. As noted in footnote 2, the term *metascience* is not common among philosophers of science. Hence, it is not immediately clear what the scientific counterpart of metamathematics in the philosophical literature is supposed to be. It turns out, however, that one division of the literature can be seen to provide a *de facto* account of metascience. This is the literature concerned with the question *what is a scientific theory?* There currently exist three major approaches to this question, viz. the syntactic view of theories, the structural view of theories and the categorical view of theories. Of these, the former two can be seen as 'old' approaches, having taken form in the first and second halves of the twentieth century respectively. We shall explore these old approaches in **chapter 2**. The focus, in particular, will lie with the most well-developed subapproaches within the framework of the structural view, viz. the *state-space approach* and the *set-theoretic approach.*

The set-theoretic approach, more so than the other two, is of a rather expansive nature. A discussion of this framework will thus require us to weigh completeness against concision. Therefore, I have opted for the following presentation. We will discuss the set-theoretic framework in a two-step fashion. Firstly, we simply look at the manner in which the notion of an individual scientific theory is formalized in the framework. This will already unveil some of the concepts valued by the framework's proponents. Secondly, we expand our scope to include not only the formalization of a scientific theory *by itself*, but also the explication of various kinds of *intertheory relations*, such as the *reduction* of one theory to another.

Before turning to the more recent approach to theory-structure, we will examine the first wave of logical abstractivism, as applied to the metascience of old, in **chapter 3**. This will be done by first investigating the logical formalism presupposed by the FWLA. This is the formalism widely known as *abstract model theory* but better referred to as *semi-abstract model theory* for reasons that will become apparent in due time.[5] We will look into some of the defining characteristics of the discipline and reflect on the potential it has to offer for for the study of metascience. Next, we will investigate how Pearce and Rantala have tried to actualize this potential, by discussing their application of semi-abstract model theory to the set-theoretic approach to scientific theories.

In **chapter 4**, we set foot in modern times with a discussion of the categorical approach to scientific theories. Obtaining an adequate picture of this contemporary school of metascience is a somewhat mathematically involved task, requiring notions from the discipline of category theory. Readers unfamiliar with this discipline may consult appendix A for an introduction. Typical of this approach is the existence of several different mathematical characterizations of a given notion, e.g. the concept of *theory* or *equivalence*. Hence, much of the chapter is dedicated to properly explicating the variety of these notions, following broadly the same two-step approach as used for the discussion of the set-theoretic approach.

Finally, I will argue that the first wave of logical abstractivism presents us with an ultimately misguided attempt to extend the metascientist's formal apparatus and sketch the outlines of a possible second wave of logical abstractivism in **chapter 5**. This SWLA, I argue, possesses a number of attractive features when compared to the FWLA and has a rich potential for metascientific applications. Roughly, the line of attack here will be to show that the FWLA fails in at least two different ways. That is:

(i) It fails to optimally utilize its ambient formal framework, i.e. semi-abstract model theory.

(ii) It fails to improve the metascientific framework at which it was targeted.

After this, we will explore two recent model-theoretic frameworks, viz. the study of *abstract modal logics* and the field of *institution-independent model theory*, which, each in their own way, represent marked improvements with respect to

[5] Less common names, for this discipline include *abstract logic* and *general model theory.*

semi-abstract model theory. Lastly, I will argue that these new formal framework can help give rise to a SWLA which avoids the pitfalls of the FWLA. This would establish, then, the central claim of the thesis as described above.

Now, before we commence with our undertaking, I wish to lay out some conventions to which I shall adhere throughout the thesis.

- While it is common among mathematicians to employ an omnipresent *we* in written text, I never found this practice to be of much value. Instead, I shall reserve 'we' for instances in which I am involving the reader in a certain task, e.g. 'We shall now look at...' In all other cases, I simply adhere to 'I', as in 'I will now argue...'

- New paragraphs are indicated by indentation. This means that when we have an equation, table or similar object separating two lines of the *same* paragraph, the first line following the equation/table will *not* be indented. An exception is when the first line of a new paragraph is both directly preceded and directly followed by white space in the adjacent lines. In this case, the first line of the new paragraph will be unindented.

- References and citations are given in accordance with the guidelines of the journal *Philosophy of Science*. In particular, this means that all citations are given by stating the author's last name, year of publication and, if relevant, a page number encased within parentheses in the body of the text. The author's name will be separated from the year of publication by a white space, while the year of publication is separated from the page number by a comma. An exception to this rule is when the author's name occurs as part of a sentence, in which case only the year of publication and, if relevant, a page number are given within parentheses.

- At the beginning of each chapter I briefly reiterate its purpose within the thesis as a whole. The content of the chapter will then be outlined in more detail, referencing the various sections.

- New concepts or phrases, the first time they are introduced, will be written in *italics*. For the most part, italics will also fulfill the role of quotation marks. In particular, italics rather than quotation marks will be used to refer to words, terms or symbols.[6] In addition, the use of italics will also retain its traditional meaning of placing emphasis on a certain expression. By contrast, quotation marks will generally be used to indicate 'figurative speech'. **Boldface** is used occasionally to signify an important notion or expression being introduced in running text, i.e. outside of definitions, lemmas, etc.

- For the most part, I adhere to the 'Mermin Imperative', i.e. the practice to "Number all your formulae, because although you may not refer to all

[6] If lexicographical considerations demand it, e.g. when referring to the symbol '+', I will use quotation marks instead.

of them, someone else might want to refer to some formula you do not refer to" (Muller 1998, x).

- Throughout the thesis, we will encounter several different *views*, *approaches* and *frameworks* for analyzing scientific theories and doing metascience. To allow for a healthy variation in vocabulary, I shall, for all intents and purposes, treat these three terms as synonyms.

There is nothing I value more in academic writing than a clear, well-defined structure. At no point should one be left wondering what we are doing, why we are doing it, and where we are with respect to our overall goals or argument. Accordingly, I hope the reader will find these values exemplified in the material presented below.

1.2 Metascience Defined

As already noted several times above, I identify the extant approaches to metascience with the approaches to the *structure of scientific theories* debate within the philosophy of science. Whence this identification? After all, 'metascience', in the present usage, refers to the formal study of *science*, not *scientific theories*. Thus, while we may perfectly well set a proper analysis of scientific theories as a necessary condition for a metascientific framework, it is certainly not a sufficient condition.

Let us ask, then, what requirements we *do* need a metascience to satisfy. We need not search far for these conditions. In fact, they are already present in the brief characterization given above, viz. metascience being the *formal study of science*. This suggests the following two criteria for frameworks for metascience. Such a framework should

(i) be articulated by means of some formal apparatus (e.g. logic, set theory) and

(ii) be able to explicate a reasonable number of aspects of the scientific enterprise.

The critical reader may note that these criteria still underdetermine what is to qualify as metascience and what not. In particular, it is unclear what is to count as a 'reasonable number' in criterion (ii).[7] However, while such a concern would be, in principle, justified, in practice we find that criteria (i) and (ii) are already sufficient to determine which frameworks are deserving of the name *metascience*.

To see why this the case, we must look at the evolution of the philosophy of science as a discipline. We may roughly divide the philosophy of science

[7] Note that replacing *a reasonable number of* with *all* in criterion (ii) would make it far too stringent. Indeed, not even metamathematics has managed to successfully include each and every aspect of mathematics within its scope. For example, the concept of explanation in mathematics still falls exclusively within the purview of informal philosophy of mathematics.

in two large subdisciplines. First, there is the *general philosophy of science*, which, true to its name, seeks to understand science at a general level without involving the details of a particular field of study such as physics, biology or geology.[8] Second, we have the *philosophy of the special sciences*, which serves as an umbrella term for various disciplines, such as the philosophy of physics and the philosophy of biology. Characteristic for these disciplines is the preoccupation with philosophical problems specific to the particular discipline under consideration, e.g. the measurement problem in philosophy of physics.

At first glance, we might consider the general philosophy of science to be the ideal breeding ground for metascience. And to some extent this true: every framework we will consider below originates from the general philosophy of science. Yet, the general philosophy of science traditionally is shaped in such a way that makes development of a full-fledged metascience difficult. Each general scientific concept, e.g. theory, model, explanation, accrues its own literature and research community and interconnections are rarely made. Naturally, this situation has not proved conducive to the formation of metascience. For much the same reason, we find no instances of *metabiology* or *metageology*[9] within the philosophy of the special sciences, which proceeds in a similar, piecemeal fashion as the general philosophy of science.

This overarching approach to philosophy of science has made 'grand frameworks', that analyze science or scientific discipline from a high level of generality, a rarity. Even the general philosophy of science itself has, in recent times, seen much neglect in favor of the philosophy of the special sciences. Now, the reason why the discourse on the notion of scientific theory *has* managed to produce such 'grand' approaches is because the notion of theory itself already occupies a central node in our informal conception of science. Indeed, if asked to make a list of the most important entities in the scientific enterprise, we might very well expect scientists to put the concept of theory near the top. Because of this, frameworks for analyzing scientific theories often make natural starting points for the analysis for other scientific concepts, as we will see below.

None of this, however, is to say that the shift away from generality towards more specialized problems is to be regarded as something lamentable. By focusing on specific issues in specific disciplines, the philosophy of science has seen a great increase in fruitfulness. In fact, the unsatisfactory nature of current attempts at metascience is the very starting point of this thesis. What I *do* find regrettable, though, is the complacency that has befallen the contemporary philosophy of science. Science as we know it today only took form after 2200 years of intellectual inquiry into the fundamentals of nature. Metascience, by contrast, has been abandoned after a mere century! Surely, there is more to be done. With these words of encouragement in mind, let us now proceed with our quest for a new metascience.

[8] This notwithstanding, a frequently encountered modus operandi in generalist frameworks is to explicitly address only a few paradigmatic disciplines, often from within physics, and leave different branches of science for future research.

[9] For historical reasons, the name of *metaphysics* has already been taken by a different field of philosophical study.

1.3 Logical Preliminaries

Throughout this thesis we will make use of several technical notions from mathematical logic and adjacent areas. To ensure the following text is largely self-contained, we will now go over some basic logical preliminaries which are needed below.[10] In addition, more elaborate introductions to the areas of category theory and topos theory may be found in appendices A and B respectively. As always, the most significant prerequisite required of the reader is the illustrious trait that is mathematical maturity.

The definitions that are to follow all pertain to *many-sorted first-order logic*. Whenever a new notion is introduced, this will be understood to be relative to the system of many-sorted first-order logic. For example, when we say that 'X is defined to be such and such', what is implicitly meant is that 'X is defined to be such and such with respect to many-sorted first-order logic'. We require this provision since each of the defined notions has identically named counterparts for different logical systems. In practice, however, it will usually be clear which system of logic we are working in.

Definition 1.3.1. A *signature* Σ is a set of sort symbols, relation symbols, functions symbols and constant symbols.[11] Each of these symbols are simply syntactic, meaningless objects, except that to each symbol we associate the property of *arity* as follows:

- Sort symbols are assigned no arity.
- Each relation symbol P has an arity of form $\sigma_1 \times \ldots \times \sigma_n$, where $\sigma_1, \ldots, \sigma_n$ denote sort symbols.
- Each function symbol f has an arity of form $\sigma_1 \times \ldots \times \sigma_n \to \sigma$, where $\sigma_1, \ldots, \sigma_n, \sigma$ denote sort symbols.
- Each constant symbol c has an arity of form σ, where σ denotes a sort symbol.

Intuitively, we can think of n as the number of 'open places' of a relation or function symbols.

Whenever a given signature contains only a single sort, we will adhere to the practice of omitting this sort symbol when specifying the signature in question. For example, suppose the signature Σ contains exactly one sort symbol σ and one relation symbol P then we simply denote this signature as $\Sigma = \{P\}$.

Definition 1.3.2. A *variable* x denotes a meaningless, syntactic object. The arity of a variable is given by a sort symbol σ.

[10] The following subsection borrows heavily from section 2 of (Barrett & Halvorson 2015).

[11] Some logicians refer to this notion by the names of *vocabulary* or *language*, reserving the term *signature* for another, closely related notion. In this thesis, however, we will not be concerned with this distinction.

It is assumed we have a countably infinite number of symbols and variables.

Definition 1.3.3. Let Σ be a signature and let $\sigma \in \Sigma$ be a sort symbol. Then a Σ-*term of sort* σ is recursively defined as follows:

- Every variable x of arity σ is a Σ-term of sort σ.
- Every constant symbol $c \in \Sigma$ is a Σ-term of sort σ.
- If $f \in \Sigma$ is a function symbol of arity $\sigma_1 \times \ldots \times \sigma_n \to \sigma$ and $t_1, \ldots, t_n$ are Σ-terms of sorts $\sigma_1, \ldots, \sigma_n$ respectively, then $f(t_1, \ldots, t_n)$ is a Σ-term of sort σ.

We say that t is a Σ-term if t is a Σ-term of sort σ for some $\sigma \in \Sigma$.

Definition 1.3.4. Let Σ be a signature and let t be a Σ-term. We write $t(x_1, \ldots, x_n)$ to denote the fact that the variables occurring in t are included in the set $\{x_1, \ldots, x_n\}$.[12]

Definition 1.3.5. Let Σ be a signature. Then the *atomic formulas over* Σ are defined to be all expressions either of the form

$$s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n),$$

where s and t are Σ-terms of the same sort, or of the form

$$P(t_1, \ldots, t_m),$$

where $P \in \Sigma$ is a relation symbol of arity $\sigma_1 \times \ldots \times \sigma_m$ and $t_1, \ldots, t_m$ are Σ-terms of sorts $\sigma_1, \ldots, \sigma_m$ respectively.

Definition 1.3.6. Let Σ be a signature. Then the Σ-*formulas* are recursively defined as follows:

- Every atomic formula over Σ is a Σ-formula.
- If φ is a Σ-formula, then $\neg\varphi$ is a Σ-formula as well.
- If φ, ψ are Σ-formulas, then $\varphi \vee \psi$, $\varphi \wedge \psi$ and $\varphi \to \psi$ are all Σ-formulas as well.
- If φ is a Σ-formula and x is a variable of arity σ, then $\forall_\sigma x\varphi$ and $\exists_\sigma x\varphi$ are Σ-formulas as well.

We say that φ is a *formula* if it is a Σ-formula for some signature Σ.[13]

Definition 1.3.7. Whenever a variable x in a Σ-formula φ occurs in tandem with a quantifier Qx, where $Q \in \{\forall, \exists\}$, it is said to be a *bound* occurrence of x. Alternatively, if x does not occur bound, it is referred to as a *free* occurrence.

[12] Note that this definition does not entail that all $x_1, \ldots, x_n$ necessarily occur in t.

[13] In general, whenever we define something to be an X relative to some signature Σ, we say that something is simply an X in case there exists some Σ for which it is X relative to Σ.

Definition 1.3.8. Let Σ be a signature and let φ be a Σ-formula. We write $\varphi(x_1, \ldots, x_n)$ to denote the fact that the free variables occurring in φ are included in the set $\{x_1, \ldots, x_n\}$.[14]

The goal is now to establish a relation between our meaningless, syntactic formulas and some meaningful, mathematical entities. The entities in question are given as follows:

Definition 1.3.9. Let Σ be a signature. Then a Σ-*model* M is defined to be a tuple consisting of

- for every sort symbol $\sigma \in \Sigma$, a non-empty set M_σ such that M_σ is disjoint from $M_{\sigma'}$, for any other sort symbol $\sigma' \in \Sigma$,
- for every relation symbol $P \in \Sigma$ of arity $\sigma_1 \times \ldots \times \sigma_n$ a relation $P^M \subseteq M_{\sigma_1} \times \ldots \times M_{\sigma_n}$,
- for every function symbol $f \in \Sigma$ of arity $\sigma_1 \times \ldots \times \sigma_n \to \sigma$, a function $f^M : M_{\sigma_1} \times \ldots \times M_{\sigma_n} \to M_\sigma$,
- for every constant symbol $c \in \Sigma$ of arity σ, a constant $c^M \in M_\sigma$.

In light of the above terminology, we need to distinguish carefully between, for instance, function symbols and functions. The only case in which this dichotomy is not observed for sort symbols, which, for convenience, will frequently be referred to as simply *sorts*. Accordingly, we will say that a is an element of sort σ if $a \in M_\sigma$.

Next, prior to linking these models to formulas, we need to relate them to terms.

Definition 1.3.10. Let Σ be a signature and M be a Σ-model. Let $a_1, \ldots, a_n$ be elements of sorts $\sigma_1, \ldots, \sigma_n$ respectively. For any term $t(x_1, \ldots, x_n)$ of sort σ with $x_1, \ldots, x_n$ of sorts $\sigma_1, \ldots, \sigma_n$ respectively, we recursively define a mapping to elements of M as follows:

- $x_i[a_1, \ldots, a_n] = a_i$, for any $1 \leq i \leq n$.
- $c[a_1, \ldots, a_n] = c^M$, for any constant symbol $c \in \Sigma$ with sort σ.
- For any term of the form $f(t_1, \ldots, t_n)$ with sort σ and $t_1, \ldots, t_n$ of sorts $\sigma_1, \ldots, \sigma_n$ respectively, define
$$f(t_1, \ldots, t_n)[a_1, \ldots, a_n] = f^M(t_1[a_1, \ldots, a_n], \ldots, t_n[a_1, \ldots, a_n]).$$

What the above definition tells us is that given an assignment of variables $x_1, \ldots, x_n$ to elements $a_1, \ldots, a_n$ of the appropriate sorts, we can extend this assignment to all Σ-terms in the obvious manner.

Now, the relation between models and formulas can be explicated as follows:

[14] Note that this definition does not entail that all $x_1, \ldots, x_n$ necessarily occur in φ.

Definition 1.3.11. Let Σ be a signature, $\varphi(x_1, \dots, x_n)$ be a Σ-formula, M be a Σ-model and let $a_1, \dots, a_n$ be elements of M. It is assumed that x_i and a_i are both of sort σ_i, for all $1 \leq i \leq n$. We now define the relation $\models$ between the sequence $a_1, \dots, a_n$ and formula φ, denoted $M \models \varphi[a_1, \dots, a_n]$, as follows:

- If φ is of the form $s(x_1, \dots, x_n) = t(x_1, \dots, x_n)$, then $M \models \varphi[a_1, \dots, a_n]$ if and only if $s[a_1, \dots, a_n] = t[a_1, \dots, a_n]$.
- If φ is of the form $P(t_1, \dots, t_n)$ for some relation symbol $P \in \Sigma$, then $M \models \varphi[a_1, \dots, a_n]$ if and only if $(t_1[a_1, \dots, a_n], \dots, t_n[a_1, \dots, a_n]) \in P^M$.
- $M \models \neg\varphi[a_1, \dots, a_n]$ iff not $M \models \varphi[a_1, \dots, a_n]$.
- $M \models \varphi \wedge \psi[a_1, \dots, a_n]$ iff $M \models \varphi[a_1, \dots, a_n]$ and $M \models \psi[a_1, \dots, a_n]$.
- $M \models \varphi \vee \psi[a_1, \dots, a_n]$ iff $M \models \varphi[a_1, \dots, a_n]$ or $M \models \psi[a_1, \dots, a_n]$.
- $M \models \varphi \to \psi[a_1, \dots, a_n]$ iff $M \models \varphi[a_1, \dots, a_n]$ implies $M \models \psi[a_1, \dots, a_n]$.
- $M \models \forall_\sigma y \varphi[a_1, \dots, a_n]$ iff $M \models \varphi[a_1, \dots, a_n, a]$ for all $a \in M_\sigma$.
- $M \models \exists_\sigma y \varphi[a_1, \dots, a_n]$ iff $M \models \varphi[a_1, \dots, a_n, a]$ for some $a \in M_\sigma$.

This definition is often referred to as a *truth definition*, since it tells us intuitively what it means for a formula φ to be considered 'true' in a given model M and given that we interpret the variables $x_1, \dots, x_n$ as the elements $a_1, \dots, a_n$ respectively. Of particular importance is the truth definition applied to a specific type of formula.

Definition 1.3.12. Let φ be a formula. If φ contains no free variables, then it is called a *closed formula* or a *sentence.*

Definition 1.3.13. Let φ be a sentence. Then we denote the fact that $M \models \varphi[]$, i.e. φ is made true in the model M for the empty sequence, by writing simply $M \models \varphi$. We then say that M *satisfies* φ or M *makes true* φ.

Definition 1.3.14. Let Σ be a signature, Γ be a set of Σ-sentences and M be a Σ-model. Then we say that M *satisfies* Γ, symbolically $M \models \Gamma$, if we have $M \models \varphi$ for every $\varphi \in \Gamma$.

In addition, we now have the following derived notions.

Definition 1.3.15. Let Σ be a signature, Γ be a set of Σ-sentences and φ be a Σ-sentence. Then we write $\Gamma \models \varphi$ if for every Σ-model M we have $M \models \Gamma$ implies $M \models \varphi$.

Definition 1.3.16. Let Σ be a signature and Γ be a set of Σ-sentences. Then Γ is called a *theory* if Γ is deductively closed, i.e. if for every Σ-sentence φ we have $\Gamma \models \varphi$ implies $\varphi \in \Gamma$.

Definition 1.3.17. Let Σ be a signature and M be a Σ-model. Then the *theory of* M, denoted $\mathrm{Th}(M)$, is the set of all Σ-sentences φ such that $M \models \varphi$.

The above definitions lie at the heart of the discipline known as *model theory*, i.e. the branch of mathematical logic concerned with the relation between formulas and models. In particular, it is of great interest to examine to what extent truth is preserved by several different relations between models. A number of paradigmatic examples of such relations are given below.

Definition 1.3.18. Let Σ be a signature and let A, B be two Σ-models. A *homomorphism* h between A and B, denoted $h : A \to B$, is given by a family of maps $h_\sigma : A_\sigma \to B_\sigma$, with σ ranging over all sort symbols in Σ, such that

- for every relation symbol $P \in \Sigma$ of arity $\sigma_1 \times \ldots \times \sigma_n$ and elements $a_1, \ldots, a_n \in A$ of sorts $\sigma_1, \ldots, \sigma_n$ respectively, we have
$$(a_1, \ldots, a_n) \in P^A \text{ implies } (h_{\sigma_1}(a_1), \ldots, h_{\sigma_n}(a_n)) \in P^B,$$
- for every function symbol $f \in \Sigma$ of arity $\sigma_1 \times \ldots \sigma_n \to \sigma$ and all elements $a_1, \ldots, a_n$ of sorts $\sigma_1, \ldots, \sigma_n$ respectively, we have
$$h_\sigma(f^A(a_1, \ldots, a_n)) = f^B(h_{\sigma_1}(a_1), \ldots, h_{\sigma_n}(a_n)),$$
- for every constant symbol $c \in \Sigma$ of arity σ, we have
$$h_\sigma(c^A) = c^B.$$

Definition 1.3.19. Let Σ be a signature, A, B be two Σ-models and $h : A \to B$ be a homomorphism. The map h is called an *isomorphism* if every h_σ has an inverse h_σ^{-1} and the resulting family of maps h^{-1} is a homomorphism.

Definition 1.3.20. Let Σ be a signature. Two Σ-models A, B are called *isomorphic*, denoted $A \cong B$, if there exists an isomorphism $h : A \to B$.

Definition 1.3.21. Let Σ be a signature and let A, B be two Σ-models. An *elementary embedding* h between A and B, denoted $h : A \to B$, is given by an injective homomorphism $h : A \to B$ such that
$$A \models \varphi[a_1, \ldots, a_n] \text{ if and only if } B \models \varphi[h_{\sigma_1}(a_1), \ldots, h_{\sigma_n}(a_n)]$$
for all Σ-formulas $\varphi(x_1, \ldots, x_n)$ and elements $a_1, \ldots, a_n$ of sorts $\sigma_1, \ldots, \sigma_n$ respectively.

How do the notions of isomorphism and elementary embedding relate to one another? By induction on structure of formulas, a technique ubiquitous in the realm of mathematical logic, we can derive the following statement:

Proposition 1.3.22. *Let Σ be a signature and let A, B be two Σ-models. Let $h : A \to B$ be an isomorphism. Then h is also an elementary embedding.*

The converse, however, is not true in general.

Finally, we may note that the above relations are defined only for models having the same signature. We would, however, also like to have to have the ability to compare models having different signatures. To this end, the following notions prove invaluable:

Definition 1.3.23. Let Σ, Σ' be signatures such that $\Sigma \subseteq \Sigma'$ and let M be a Σ'-model. The model $M|_\Sigma$, called the *reduct* of M with respect to Σ, is the model obtained from deleting from the tuple M any set, relation, function or constant associated to a symbol in $\Sigma' \setminus \Sigma$. The model M is then referred to as an *expansion* of $M|_\Sigma$ to Σ'.

As we can see, reducts provide us with a natural way to obtain 'smaller' models from larger ones. In the case of single-sorted signatures[15], there exists another straightforward manner in which we can construct smaller models, this time without changing the underlying signature.

Definition 1.3.24. Let Σ be a single-sorted signature, M be Σ-model and $S \subseteq M$ be non-empty. Then S is called Σ-*closed in* M if, for all constant symbols $c \in \Sigma$ and function symbols $f \in \Sigma$, $c^M \in S$ and S is closed under f^M.

Definition 1.3.25. Let Σ be a single-sorted signature, M a Σ-model having relations $P_1^M, \ldots, P_k^M$, functions $f_1^M, \ldots, f_m^M$ and constants $c_1^M, \ldots, c_n^M$ and S a set Σ-closed in M. Then the Σ-model $M|S$ consisting of domain S, relations $P_1^M \cap S, \ldots, P_k^M \cap S$, functions $f_1^M|_S, \ldots, f_m^M|_S$ and constants $c_1^M, \ldots, c_n^M$ is called the *submodel of* M *generated by* S.

The notions of reduct and submodel can now be nicely brought together:

Definition 1.3.26. Let Σ, Σ' be single-sorted signatures such that $\Sigma \subseteq \Sigma'$. Let M be a Σ'-model and $P \in \Sigma$ a unary relation symbol such that P^M is Σ-closed in $M|_\Sigma$. Then the Σ-model $(M|_\Sigma)|P^M$ is called the *relativized reduct* of M with respect to Σ and P.

With these initial preliminaries in place we are now well-equipped to move on to the subsequent chapters and investigate the various ways in which logic has been applied to the study of metascience.

[15] The restriction to single-sorted signatures is not of essential import, but merely serves to make the subsequent definitions less cumbersome. For the purposes of this thesis, we shall only be requiring the single-sorted case.

Chapter 2

Metascience of Old

We will now meet what we may consider to be the 'traditional' approaches to metascience. As described in the previous chapter, the study of metascience will be identified with the philosophical investigations into the structure of scientific theories. I start by setting out a number of terminological conventions and recommendations (sec. 2.1), after which we will briefly make acquaintance with the oldest of all approaches to theory-structure, i.e. the syntactic view of theories (sec. 2.2). Since this approach largely falls outside of the scope of the FWLA, we can suffice here with only a superficial description. Next, the structural view of theories is considered. This view has given rise to a wide variety of frameworks, of which we will discuss the two most prominent, namely the state-space approach (sec. 2.3) and the set-theoretic approach (sec. 2.4). Due to its expansive nature and connection to the FWLA, the set-theoretic approach will receive the most elaborate treatment in this chapter. Having analyzed the preceding approaches, we will be well-prepared to consider the first wave of logical abstractivism in the subsequent chapter.

2.1 Notes on Terminology

Before we commence our discussion of the traditional schools of metascience, let us briefly dwell on some terminological matters pertaining to the different approaches to scientific theories.

- Two alternative designations of the syntactic view of theories are encountered in the literature, viz. the *received view*, a term first coined by Hilary Putnam, and the *statement view*. The former of these, however, has become all but obsolete, as the syntactic view has not occupied a position of dominance in the philosophical literature since the early 1950s.

- Both the set-theoretic and state-space approach are usually designated as being part of the *semantic view* or, less commonly, the *model-theoretic view*. Both names are rooted in considerations from logic, in which the structures interpreting a formal language are referred to as *models*. These models, in turn, then provide a *semantics* for the formal language in question. In this light, however, we see the standard terminology to be highly undesirable. As noted by Muller (2011, 103), it is not the case that a given class of structures which we take to represent a scientific theory is also the class of models of some set of sentences in a formal language.[1] Now, while we may or may not have good reason to assume that a given class of structures is also a class of models, *a priori* it is certainly not obvious that we ought to impose such a restriction. Thus, we see that the name *structural view* provides us with a far more fitting designation than either *semantic view* or *model-theoretic view* can hope to offer.[2]

- The denotation *set-theoretic approach* is used here as a synonym for the approach known as *structuralism*, as developed by Sneed (1971). Strictly speaking, however, *set-theoretic approach* denotes a slight more general approach to scientific theory-structure, which includes not only structuralism but also the work of Suppes (1967). The structuralist approach, however, has seen by far the most development and differs only subtly from Suppes' account. Hence, we can freely indulge in the aforementioned *totum pro parte* without significant risk of confusion.

- The term *structuralism* and derived expressions have a myriad applications within philosophy of science and the debate on theory-structure is no exception. Though not common, some authors[3] use the denotation *structuralism* to refer to a group of three similar accounts of theory-structure which, besides the Sneed's account, includes the work of Scheibe (1979) and Ludwig and Thurler (2006). Again, however, we find that the Sneedian framework eclipse in size and impact the other two approaches. In practice, therefore, we might safely use *structuralism* to refer to the body of work of Sneed et al.[4]

2.2 Syntactic Approach

The syntactic approach to scientific theories, along with much else of the philosophy of science, finds its origin in the first half of twentieth century in the work of the Viennese group of scholars known as the *logical positivists*. Inspired by the early successes of the logical formalization of mathematical theories at the hands of, among others, Peano, Frege and Russell, the positivists sought to apply the fledgling discipline of mathematical logic to the formal analysis of the

[1] We must be careful here to distinguish between the notion of a scientific theory, which is the target of our formalization efforts, and the notion of *theory* as found in logic, where it denotes a (deductive closed) set of sentences.

[2] To add insult to injury, we might note that approaches in the *semantic* or *model-theoretic view* all but never employ methods that are semantical or model-theoretic in nature.

[3] E.g. Schmidt (2014).

[4] That is, within the context of theory-structure. In the more general realm of philosophy of science, such a reference would still necessitate due amounts of elaboration.

scientific enterprise. Now, logic, at this time, was a purely symbolic affair, as the notions we would nowadays consider *semantic* would not be formulated until the 1950s. Accordingly, the positivists championed a syntactic formalization of scientific theories.

In the broadest of strokes, we may describe the syntactic view on scientific theories as the claim that a scientific theory S is specified by the following two sets of data. First, we have:

- A system of logic L, in this context taken to be first-order logic or some suitable extension thereof, equipped with a standard syntax (i.e. rules for sentence formation).

- A certain proof-theoretic structure $\vdash$ on L, allowing us to make deductions in L. The logic L together with the relation $\vdash$ may then also be called a *logical calculus.*

- A **tripartitioned** signature $\Sigma = \Sigma_M \cup \Sigma_O \cup \Sigma_T$ and Σ_M, Σ_O and Σ_T represent the theory's mathematical, observational and theoretical terms respectively.[5]

- A deductively closed set of Σ-sentences Γ.

Were we to stop at this stage, we might simply choose to identify the scientific theory S with the set Γ relative to $L, \vdash$ and Σ. Scientific theories would then simply reduce to the special case of mathematical theories with tripartitioned signatures. To characterize S as being a truly *scientific* theory, we further require:

- A set of **correspondence rules**, providing each observational term with an interpretation in terms of physical objects.

- A set of correspondence rules, providing each theoretical term with a partial interpretation in terms of observational terms.

Already, we can discern one of the problematic features associated to the syntactic view: it is very difficult to obtain a complete specification of theoretical terms in terms of observational terms. At best, we can only hope to *partially* interpret such terms. Further complicating matters is the fact that much of the positivists' work on scientific theories remains clad in informal descriptions, consequently making it harder to adequately assess the tenability of the proposed their proposed theory-concept. A more formal continuation of the syntactic view may be found in (Przełecki, 1969), although in this work too the explication of *concrete* scientific theories runs into much difficulty. I shall limit my treatment of the syntactic approach to the brief sketch provided above. For

[5] Straightforward examples include '+' for the mathematical terms, 'chair' for the observational terms and 'electron' for the theoretical terms. It should be noted, however, that the exact distinction between the latter two types of terms is a topic of much controversy within the philosophy of science.

more comprehensive descriptions of the syntactic view, see (Suppe 1977, 16–57), (Suppe 1989, 38–72).

The syntactic conception of theories reigned supreme for most of the first half of the twentieth century. In spite of, or perhaps because of, this position of dominance, the syntactic approach came to be the target of a host of criticism from philosophers of science. The untenability of the observational-theoretical distinction and the overall difficulty of formalizing actual scientific theories in formal languages were but some of the factors that eventually led to the syntactic view's demise. In its place, a new paradigm for theory-structure rose to prominence over the second half of the twentieth century that took a distinctly different approach to the formal study of theories. Characteristic of this new approach was the conviction that not statements, but *(mathematical) structures* are the most valuable tools at our disposal for the analysis of scientific theories. Accordingly, we may refer to this new view as the *structural view* of scientific theories.

Unlike its comparatively monolithic predecessor, the structural approach to theories consists of a family of different, closely related frameworks, each having a different notion of structure at its core. Of all these approaches, however, only two have seen significant uptake in the philosophical community. These are the so-called *state-space approach* to scientific theories, as pioneered by E.W. Beth (1960) and Bas C. van Fraassen (1970), and the *set-theoretic approach*, as pioneered by Patrick Suppes (1967) and Joseph Sneed (1971). The latter of these two in particular has gained a considerable following, making up a significant amount of the literature on the structural view. We will examine both approaches in the subsequent sections.

2.3 State-Space Approach

Inspired by foundational work in quantum mechanics by, among others, von Neumann, and Tarski's semantics for formal languages, the state-space approach was first developed by the Dutch logician Evert Willem Beth in the late 1940s (Dieks 2010, 278). Beth's work, however, is limited to classical and quantum mechanics and he does not present, nor claim to present, a general or complete account of theory-structure. The novel component of Beth's ideas was the manner in which physical theories could be given a formal semantics by using an appropriate state space to evaluate the truth conditions of certain statements of the theories (Beth 1960). Beth's ideas ideas were later developed by the Dutch-American philosopher Bas van Fraassen (1970).[6] American philosopher Frederick Suppe independently developed an approach similar to that of Beth and Van Fraassen (Suppe 1989, 16). For present purposes, however, I shall limit my exposition here to the version of the state-space approach as formulated by

[6] As noted in (Van Fraassen 1989, 365n), Van Fraassen later came to use a modification of the original approach, as set out in his (1970). I, however, believe Van Fraassen's early work to provide us with the more interesting explication of scientific theory-structure and shall consequently focus exclusively on the 1970 framework.

Van Fraassen.[7]

In Van Fraassen's formalization, a scientific theory[8] consists of the following components:

- A **state space** H, i.e. a mathematical space consisting of the possible states of the physical system under consideration.
- A **language** consisting of a set E of **elementary statements** of the form $P(m, r, t)$, or P for short, and **modal connectives** and '$\Box$' and '$\Diamond$'. The statement $P(m, r, t)$ should be thought to mean that the magnitude m has value r at time t. The compounded statements $\Box P$ and $\Diamond P$, signify that the statement P holds for all values of t and some value of t respectively.
- A group of operators $\mathbf{U}_t$ defined such that if s is the state of the physical system at time t', then $\mathbf{U}_t(s)$ is the system's state at time $t' + t$. In other words, the operators define a **time evolution** on the state space.
- A **satisfaction function** h, which associates to each elementary statement P a region of state space $h(P)$ and is defined as

  $$h(P(m, r, t)) = \{s \in H : \mathbf{U}_t(s) \in h(P(m, r, 0))\}. \tag{2.1}$$

  The set $h(P(m, r, 0))$ should be thought of as the region of state space in which all states satisfy the statement $P(m, r)$.

Furthermore, Van Fraassen provides us with a definition of truth for the statements of the scientific theory. Let X be the system with which the theory in question is concerned and let *loc* be a function that associates to X a state in the state space, i.e. $loc(X) \in H$. A statement P about the system X is then said to be *true* if and only if $loc(X) \in h(P)$. This truth definition may be seen to provide us with a **semantics** of the scientific theory.

The above definitions can be readily applied to Newtonian physics. Take, for instance, a one-dimensional system with a single object. The state space is then, by definition, the subset of $\mathbb{R}^2$ consisting of pairs (x, p) of all possible values of the object's position and momentum. The system's time evolution is given by Newton's second law. An instance of an elementary statement would be 'position x has value 0 at $t = 0$', which by the satisfaction function would be mapped to the p-axis of the state space. The more expressive, modal statements of the form $\Box P$ correspond here to quantified statements such as '$x(t) = 1/2gt^2$', which by h are mapped to the state-space trajectories of the system.

Overall, it seems more philosophers have taken up the set-theoretic approach than the state-space approach. Nevertheless, the state-space approach has still found several supporters. In particular, it has been applied to the analysis of the structure of biological theories, cf. (Lloyd 1994), (Thompson 1989).

[7] The subsequent material of this subsection has been taken, with minor modifications, from (Vos 2015a).

[8] It should be noted that Van Fraassen limits his analysis to non-relativistic, physical theories (1970, 328).

2.4 Set-Theoretic Approach

In this section we turn to the next major approach within the structural view of theories known as the *set-theoretic approach* or *structuralism*. Although structuralism is somewhat notorious for its elaborate formal machinery, it remains one of the most extensively developed and well-known accounts of scientific theory-structure. Heavily influenced by the work of Patrick Suppes (1967), the structuralist approach to theory-structure was first expounded by Joseph D. Sneed in his 1971 book *The Logical Structure of Mathematical Physics*. Since then, structuralism underwent a number of revisions and expansions by a number of different authors, most notably at the hands of German-Austrian philosopher Wolfgang Stegmüller (1976). I shall focus here on the views proposed in one of its most recent incarnations found in (Balzer, Moulines & Sneed 1987).

2.4.1 Theories as Structures

In the structuralist approach a scientific theory is represented by a class of models which is given by some *structure species*. Consequently, we must first examine this notion of structure species if we wish to understand the structuralist conception of theories. The definition of structure species is built on several auxiliary notions which we will have to look at first. To understand what motivates the upcoming definitions, it is useful to refer to the set-theoretic definition of a semigroup.

Definition 2.4.1. x is a *semigroup* if there exist G, $\cdot$ such that

- $x = (G, \cdot)$
- G is a non-empty set
- $\cdot$ is a binary operation
- for all $a, b, c \in G : (a \cdot b) \cdot c = a \cdot (b \cdot c)$

We can extract several general properties from the above list. First, there is the denotation of the base sets. In the case of semigroups there is only one, but we can easily imagine structures, e.g. vector spaces, for which this does not hold. Note that since we are defining the general notion of semigroup, it is irrelevant which set is used in the definition. We might just as well have used G' instead of G, without altering the definition in any salient manner. Thus, we see that the first essential component of the definition of a structure species is the **number of base sets**. Second, we find that from the base set G, a ternary relation $\cdot$ is defined. Since it is again unimportant with which symbol we refer to this relation, we only require the **typification** of the relation, i.e. the arity of the relation in the case of a single base set. Lastly, there are **set-theoretic**

sentences expressing several substantial properties of the structure species.[9] Axiom (4) obviously is such a sentence, but also the statement that $\cdot$ is a binary operation, as this does not follow from the fact that is a ternary relation.

Before proceeding let us first make the notion of typification more rigorous. We first require a number of definitions (Balzer et al. 1987, 8).

Definition 2.4.2. For each $k \in \mathbb{N}$, *k-types* σ are defined inductively as follows:

- for each $i \leq k$: i is a k-type,
- if σ is a k-type then so is $\mathcal{P}(\sigma)$,
- if σ_1 and σ_2 are k-types then $\sigma_1 \times \sigma_2$ is a k-type.

Instances of 4-types would be 2, $\mathcal{P}(1)$ and 3×4. Note that at this point such expressions are of a purely syntactic character and have yet to supplied with any meaning.[10] This constitutes our next step.

Definition 2.4.3. Suppose $k \in \mathbb{N}$, $D_1, \ldots, D_k$ are sets and σ is a k-type then the *echelon set* $\sigma(D_1, \ldots, D_k)$ is defined by induction with respect to σ as follows:

- if σ is some $i \leq k$ then $\sigma(D_1, \ldots, D_k) = D_i$,
- if σ has the form $\mathcal{P}(\sigma_1)$ where σ_1 is a k-type previously defined then $\sigma(D_1, \ldots, D_k) = \mathcal{P}(\sigma_1(D_1, \ldots, D_k))$, where $\mathcal{P}$ denotes the power set operation;
- if σ has the form $\sigma_1 \times \sigma_2$ where σ_1 and σ_2 are k-types previously defined then $\sigma(D_1, \ldots, D_k) = \sigma_1(D_1, \ldots, D_k) \times \sigma_2(D_1, \ldots, D_k)$, where $\times$ denotes the Cartesian product.

We can now give a general definition of the notion of typification.

Definition 2.4.4. A set-theoretic sentence[11] A is called a *typification* if there exists some k-type σ such that A has the form '$R \in \sigma(D_1, \ldots, D_k)$', where $R, D_1, \ldots, D_k$ denote sets.

Having defined the notion of a typification, we now also have the following:

Definition 2.4.5. A *type* τ is defined to be a tuple $(k, \sigma_1, \ldots, \sigma_n)$ such that $k \in \mathbb{N}$ and $\sigma_1, \ldots, \sigma_k$ are k-types. A *structure of type* τ is taken to be a tuple $(D_1, \ldots, D_k, R_1, \ldots, R_n)$ such that $D_1, \ldots, D_k$ are sets and for all $i \leq n$ we have $R_i \in \sigma_i(D_1, \ldots, D_k)$.

[9] The word 'substantial' is of essence here. As we will see below, typifications are defined to be a particular kind of set-theoretic sentences. Regardless, Balzer et al. still maintain a division of typifications on one hand and the set-theoretic sentences mentioned here on the another. Typifications, they say, only serve to fix the structure species's conceptual framework, e.g. the arity of its relations or operations, whereas set-theoretic sentences as used here are meant to denote more characteristic properties.

[10] Of course, the notation $\mathcal{P}$, $\times$ has been chosen so as to suggest a natural interpretation.

[11] The structuralists only use this term informally, not providing a formal definition. The intuition underlying this notion, however, seems to be clear.

We see that two of the three components of structure species mentioned above, namely the number of base sets and typifications, are fixed by the specification of a type. To complete the definition of a structure species, we thus only have to add the final part, i.e. set-theoretic sentences expressing the structures' characteristic properties.

Definition 2.4.6. A set-theoretic sentence A *applies* to some given structure $(D_1, \ldots, D_k, R_1, \ldots, R_n)$ if at most the symbols 'D_1', ..., 'R_n' occur freely in the sentence A.

With this final definition in place, we can now bring all of the preceding together.

Definition 2.4.7. If $\tau = (k, \sigma_1, \ldots, \sigma_n)$ is a type, then Σ is a *structure species of type* τ if there exist $A_1, \ldots, A_s$ such that $\Sigma = (k, \sigma_1, \ldots, \sigma_n, A_1, \ldots, A_s)$ and for all $i \leq s : A_i$ is a set-theoretic sentence[12] applying to some structure of type τ.

Definition 2.4.8. Σ is a *structure species* if there is some type τ such that Σ is a structure species of type τ.

Let us now examine how Balzer et al. link this abstract definition of structure species to scientific theories. In (Balzer et al. 1987) the discussion of theories proceeds in three stages. First, they define **theory-elements**, followed by **theory-nets** and finally **theory-holons**. The first is meant to formalize the structure of individual theories, the second of webs of interconnected theories and the third of science as a whole. Since it is my aim to give only the characteristics of the approach, I limit my discussion here to theory-elements. To this end, I shall first present the reader with the definition of a theory-element and will then proceed by explicating each concept that occurs in it in turn.

Definition 2.4.9. A *theory-element* is a pair $T = (K, I)$ where

1. $K = (M_p, M, M_{pp}, GC, GL)$ is a **theory-core** consisting of
   (a) a class of potential models M_p,
   (b) a class of models of M within M_p,
   (c) a class of partial potential models M_{pp} given by M_p and M,
   (d) the global constraint belonging to M_p,
   (e) the global link belonging to M_p.
2. a set of **intended applications** $I \subseteq M_{pp}$.

It is apparent that explicating even the fundamental notion of theory-element is a rather involved task in the structuralist framework, presupposing the definitions of a number of different ancillary notions all of which require there own set of motivations. Nevertheless, explicating these concepts will present us with

[12] Note that this includes typifications.

a characteristic insight into the set-approach approach and well thus serve well as an introduction to this framework.

First and foremost, we have the notion of **potential model**. Intuitively, a potential model of a certain structure species is a structure satisfying the basic set-theoretic sentences but not necessarily satisfying the more substantial sentences. What does it mean for a set-theoretic sentence to be 'basic' or 'substantial'? As noted before, we may view typifications as being of the former kind. However, other sentences might also fall under this denominator. The claim that $\cdot$ is a binary operation in definition 2.4.1 needs to be expressed by a certain set-theoretic sentence that is more than just a typification. Such a claim, however, can hardly be considered 'substantive' for the definition of a scientific theory.

To decide which structures are to be designated as potential models, we need a proper understanding of the 'basic' sentences which they need to satisfy. Balzer et al. note that a common feature of these basic sentences is that they contain, besides symbols for base sets, only one function or relation symbol. They refer to such sentences as *(mathematical) characterizations* (Balzer et al. 1987, 13-14). This definition is motivated by the observation that sentences expressing more substantive structural properties establish some kind of connection between the relations or functions of a given structure. For instance, while Newton's second law is a substantive statement, the prerequisite that any position function is twice-differentiable hardly seems like a substantive claim of classical mechanics and is better thought of as mathematical requirement on the position function. Therefore, characterizations (typifications included) only serve to fix the theory's conceptual framework and eliminate any structure which the theory cannot sensibly be applied to, e.g. structures with a nowhere-differentiable position function for classical mechanics. These considerations can now be formally expressed in the definition of a potential model.

Definition 2.4.10. x is said to be *potential model* with respect to some structure species $\Sigma = (k, \sigma_1, \dots, \sigma_n, A_1, \dots, A_s)$ if

- x is a structure of species Σ,
- $s = n$,
- for all $i \leq n$, A_i is a characterization.

We can now define the first major component of theory-elements.

Definition 2.4.11. A set M_p is a *class of potential models* if there is a structure species Σ such that M_p is the class of all potential models with respect to Σ.

So far the potential models. What can we now say about a theory's (actual) **models**? While the class of potential models was constrained by having to satisfy characterizations, the class of models will have to satisfy, in addition to characterizations, the 'substantive' claims mentioned earlier. These claims can intuitively be thought of as the laws of the theory. We might try to define models analogously to how we defined potential models. This would require us to specify which set-theoretic sentences qualify as laws and which do not. However, since the notion of law-likeness has proved to be hard to rigorously define in philosophy of science, Balzer et al. opt for a simpler definition of models (1987, 16).

Definition 2.4.12. A set M is a *class of models* if there exists a structure species Σ such that M is the class of all structures of species Σ and M is not a class of potential models.

In practice, this means Σ will have to contain some set-theoretic sentences other than characterizations, which we may then refer to as *laws*. If we let Σ' be the structure species equal to Σ minus the laws, it is easy to see that any structure of species Σ will also be a potential structure with respect to Σ'. Hence, any model of a theory is also a potential model of that theory.

The next variety of models we encounter in theory-elements is the **partial potential model**. As their name suggests, partial potential models are obtained by truncating potential models. The introduction of partial potential models is motivated by the great importance Balzer et al. place on explicating the role of theoretical terms in scientific theories. Accordingly, partial potential models can be seen as consisting of those and only those parts of potential models that pertain to non-theoretical terms. This is made precise below.

Definition 2.4.13. The set M_{pp} is the *class of partial potential models* given by M_p and M for a theory T if for each $x \in M_{pp}$. there exist

$$D_1, \ldots, D_k, A_1, \ldots, A_l, n_1, \ldots, n_p, t_1, \ldots, t_q$$

such that

- $x = (D_1, \ldots, D_k, A_1, \ldots, A_l, n_1, \ldots, n_p)$,
- $(D_1, \ldots, D_k, A_1, \ldots, A_l, n_1, \ldots, n_p, t_1, \ldots, t_q) \in M_p$,
- exactly $t_1, \ldots, t_q$ are T-theoretical.

As we can see, a *term* is construed here as a function or relation defined on a potential model. I will not go into detail here about the manner in which Balzer et al. define the notion of T-theoreticity for a given theory T. For the purpose of understanding the notion of theory-element, it suffices to keep in mind an intuitive division between theoretical and non-theoretical terms. A more formal discussion can be found in (Balzer et al. 1987, 61-78).

Having discussed the different varieties of models, we now turn to the last two components of theory-cores. Let us first consider the notion of a **constraint**. Informally, constraints impose certain conditions on the relations holding between potential models. To see why we ought to be interested in such relations, consider an example from classical mechanics (Balzer et al. 1987, 44–5). Suppose we want to send a rocket from the Earth to the Moon. To determine the maneuvers needed for the rocket to successfully land on the lunar surface, it is critical we know the rocket's mass. Therefore, before take-off, it is necessary to determine the mass of the rocket on Earth.

Next, the system containing the rocket, Earth and Moon is considered and the appropriate calculations are performed using, among other things, the mass of the rocket that was previously determined. In doing so, the structuralists claim, we are using an implicit assumption of classical mechanics. When we determine the mass of the rocket on Earth, we assumed its mass would be the same on its journey to the Moon (taking into account the mass of the fuel). In general, the assumption is made that in classical mechanics the mass of an object is constant in all physical systems. In structuralist terms, this translates to the assumption that if $x, x' \in M_p$ and m and m' are the respective mass functions that then $m(p) = m'(p)$ for all particles p that occur in both x and x'. This is what Balzer et al. refer to as an 'equality constraint' for classical mechanics (1987, 45). This constraint effectively forbids combinations of systems in which the same object has different mass. Formally then, we can construe a constraint as picking out certain admissible subsets of M_p. A constraint can therefore be identified with the set of all admissible subsets. This leads us to the following definition.

Definition 2.4.14. If M_p is a class of potential models, then C is a *constraint* for M_p if we have that

- $C \subseteq \mathcal{P}(M_p)$,
- $\varnothing \notin C$,
- for all $x \in M_p, \{x\} \in C$.

The first condition is a direct translation of the characterization of constraints in terms of all admissible subsets of M_p. Furthermore, Balzer et al. consider 'combinations' of zero potential models nonsensical, hence the second condition. The third requirement simply states that any potential model combined only with itself always is always admissible.

There is, of course, no reason to assume a theory has only a single constraint. Suppose $C_1, \ldots, C_n$ are all constraints belonging to some theory, we then define the **global constraint** GC of that theory to be the intersection of $C_1, \ldots, C_n$. This constitutes the fourth component of a theory-core.

To complete our survey of the components of theory-cores we now consider the notion of an inter-theoretical **link**. The introduction of this notion is motivated by the observation that theories often depend on notions from earlier theories. More explicitly, Balzer et al. note that for a theory T, it is often the case that the T-non-theoretical terms can only be determined by means of other theories that do not presuppose T (1987, 58). This is for instance the case for time in classical mechanics, since the conditions of time measurement are not a part of that theory. Even though such prerequisites are often not explicitly acknowledged in expositions of theories, it is clear they can form integral parts of these theories. Thus, Balzer et al. consider inter-theoretical links to be a necessary component of their account of theory-structure (1987, 59).

Before looking at the formal definition of links, we first consider a matter of notation.

Definition 2.4.15. For any theory T and $i_1, \ldots, i_n \in \mathbb{N}$, let $\pi(T, i_1, \ldots, i_n)$ denote the set of all tuples $(R_{i_1}, \ldots, R_{i_n})$ for which there is some $x \in M_p$ such that for all $1 \leq j \leq n$, R_{i_j} equals the i_j-th component of x.

With this convention in place, let us now define the notion of link.

Definition 2.4.16. Let M_p and M_p' be classes of potential models having m and m' relations/functions respectively. We call L a *link* between M_p and M_p' if there exist $i_1, \ldots, i_s \in \{1, \ldots, m\}$ and $j_1, \ldots, j_t \in \{1, \ldots, m'\}$ such that

- $L \subseteq M_p \times \pi(T, i_1, \ldots, i_s) \times M_p' \times \pi(T, j_1, \ldots, j_t)$,
- if $(x, (r_1, \ldots, r_s), x', (r_1', \ldots, r_t') \in L$, then for all $k \leq s$ and $l \leq t$: r_k and r_l' are the i_k-th and j_l-th components of x and x' respectively.

Suppose now that for a given theory T with class of potential models M_p, we have links $L_1, \ldots, L_n$ between M_p and the classes $M_p^1, \ldots, M_p^n$ respectively. For each $i \leq n$, let $\lambda_i = \{x \in M_p \mid \exists \bar{a} \exists x' \exists \bar{a}' : (x, \bar{a}, x', \bar{a}') \in L_i)\}$ where $\bar{a}$ and $\bar{a}'$ denote tuples of arbitrary length. Each λ_i consists of the potential models satisfying the link L_i. Finally, we define the **global link** GL of the theory T to be the intersection of all $\lambda_1, \ldots, \lambda_n$.

We have now covered the five components that make up a theory-core. While a theory-core can be thought of as representing the formal structure of a theory-element, we have yet to consider the other major component of theory-elements, i.e. the set of intended applications. From definition 2.4.9 we know that the structure the intended applications is significantly less complicated than that of the theory-core. The set of intended applications I is simply a subclass of the class of partial potential models M_{pp}. The underlying intuition here is that the partial potential models, containing only the non-theoretical part of the potential models, represent all the possible applications of the theory in question. The *intended* applications are then selected from these by specifying a subset of M_{pp}.[13] It should be noted that Balzer et al. do not claim this definition of intended applications is entirely satisfactory. Instead, they claim the present definition can only capture some necessary conditions for something to be considered an intended application. Sufficient conditions are considered much harder to identify (1987, 87–8).

How are empirical phenomena, in this case given by the intended applications of some theory, related to the theory itself? We find this question to have a

[13] An example of unintended applications, according to Balzer et al, would be a partial potential model of classical particle mechanics in which we have people instead of particles, animals instead of time and the velocity function is replaced by the function denoting how many of certain animal species each person owns (1987, 87). Even though such a structure qualifies as a partial potential model, it is clear we would not want to consider it an intended application of the theory.

rather explicit answer in the structuralist program. As we might expect, the theory-phenomena relation is explicated in terms of set-theoretic notions. In particular, we need to consider the set Cn, which denotes the **content** of a given theory.[14] The definition of content depends on several auxiliary notions that need to be considered first.

Definition 2.4.17. Let $K = (M_p, M, M_{pp}, GC, GL)$ be a theory-core. The *theoretical content* of K is the set $Cn_{th} = \mathcal{P}(M) \cap GC \ \cap \mathcal{P}(GL)$.

As we can see, the theoretical content of a theory-core K consists of those and only those sets of models of K that satisfy all constraints and inter-theoretical links. Note that the theoretical content is completely determined by only four of the five components of the theory-core. We have yet to account for the influence of the class of partial potential models. For this we require the following definitions.

Definition 2.4.18. Let $K = (M_p, M, M_{pp}, GC, GL)$ be a theory-core. Then $r : M_p \to M_{pp}$ is the map given by $r(D_1, \ldots, D_k, n_1, \ldots, n_m, t_1, \ldots, t_n) = (D_1, \ldots, D_k, n_1, \ldots, n_m)$ for any potential model in M_p.

Definition 2.4.19. Let $K = (M_p, M, M_{pp}, GC, GL)$ be a theory-core and let $r : M_p \to M_{pp}$ be as in definition 2.4.18. We then define $r' : \mathcal{P}(M_p) \to \mathcal{P}(M_{pp})$ to be the map given by $r'(X) = \{r(x) \mid x \in X\}$ for any $X \subseteq M_p$. Similarly, we let $r'' : \mathcal{P}(\mathcal{P}(M_p)) \to \mathcal{P}(\mathcal{P}(M_{pp}))$ be the map $r''(X') = \{r'(X) \mid X \in X'\}$ for any $X' \subseteq \mathcal{P}(M_p)$.

As already noted, we might think of partial potential models as representing the phenomena about which a theory speaks. What the map r does then, is map a certain part of theory (in the form of a potential model) to a certain empirical phenomenon. Since the theoretical content of a theory is given by some set of potential models Cn_{th}, the phenomena about which the theory in question speaks are given by the set $r''(Cn_{th})$. Indeed, this is precisely what we define the content of a theory to be.

Definition 2.4.20. Let $K = (M_p, M, M_{pp}, GC, GL)$ be a theory-core. The *content* of K is then given by the set $Cn := r''(Cn_{th})$.

For a theory-element $T = (K, I)$, the content of T represents all phenomena the theory-core, i.e. the formal part of the theory, speaks about. Thus, to say that a theory successfully describes the set of intended applications I, amounts to saying that I in included in theory's content. In other words (Balzer et al. 1987, 91):

Definition 2.4.21. If $T = (K, I)$ is a theory-element, then the *empirical claim* of T is that $I \in Cn$.

Thus far the structuralist construal of individual theories. Much of the preceding discussion has taken place in a rather abstract plane far removed from

[14] Or more precisely: of a given theory-core.

the level of concrete scientific theories. Therefore, let us consider an example of how the above definitions are applied to the paradigmatic example of a scientific theory, viz. **classical particle mechanics (CPM)**. To obtain the theory-element associated to CPM, we will consider each of its constituents in turn.

Again, our point of departure will be the identification of the theory's potential models. In the case of classical particle mechanics, these are defined as follows (Balzer et al. 1987, 103).

Definition 2.4.22. The class $M_p(\text{CPM})$ of potential models for CPM consists of all tuples x such that

- $x = (P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f)$,
- P is a finite, non-empty set and S, T are sets;
- $c_1 : T \to \mathbb{R}$ and $c_2 : S \to \mathbb{R}^3$ are bijections,
- $s : P \times T \to S$ is a function such that $c_2 \circ s_p \circ c_1^{-1}$ is smooth for all $p \in P$,[15]
- $m : P \to \mathbb{R}\backslash\{0\}$ and $f : P \times T \times \mathbb{N}$ are functions.

Intuitively, we should think of P as representing a set of particles, T as an interval of time and S as a region of space. In the same vein, the functions s, m and f can be taken to represent position, mass and force functions respectively, whereas the set $\mathbb{N}$ serves to label all possible component forces. With this in mind, the (actual) models of CPM are now readily defined. Let us recall that the key difference between potential and actual models of a theory was that the latter have to satisfy, in addition to the requirements placed on the potential models, certain characteristic set-theoretic sentences. Namely, they had to satisfy those sentences relating at least one function or relation to some other function or relation occurring in the models. Moreover, we noted that such sentences could informally be labeled as the *laws* of the theory under consideration. Now, in the case of CPM, it is clear which law we should consider: Newton's second law of motion. This leads us to the following definition.

Definition 2.4.23. The class $M(\text{CPM})$ of models for CPM consists of all tuples x such that

- $x = (P, T, S, \mathbb{N}, \mathbb{R}, c_1, c_2, s, m, f)$,
- $x \in M_p(\text{CPM})$,
- for all $p \in P$ and $a \in \mathbb{R}$: $m(p) \cdot D^2 r(p, a) = \sum_{i \in \mathbb{N}} f(p, c_1^{-1}(a), i)$,

where D^2 denotes the differential operator d^2/dt^2 and $r(p, a) := (c_2 \circ s_p \circ c_1^{-1})(a)$.

[15] Here, we use the notation s_p to denote the function $s(p, \cdot) : T \to S$.

Next, we consider how to obtain from our class of potential models the class of partial potential models. In effect, this is achieved by deciding which functions/relations are to be viewed as theoretical and which are to be deemed non-theoretical with respect to CPM. While this is no trivial task, especially given the controversy regarding the definability of force and mass in classical mechanics,[16] we will not enter into the philosophical details of these considerations here and will simply take as given that the force function f and mass function m are theoretical in the theory-element CPM. Thus, let r be the function truncating from any potential model in $M_p(\text{CPM})$ the force and mass functions as well as the base set $\mathbb{N}$. This then yields the following definition.

Definition 2.4.24. The class $M_{pp}(\text{CPM})$ of partial potential models for CPM is given by the class $r(M_p(\text{CPM}))$.

Lastly, we need to specify the constraints and links for the theory-element CPM. Balzer et al. identify three different constraints necessitated by the study of classical particle mechanics (1987, 106):

Definition 2.4.25. The global constraint $GC(\text{CPM})$ of CPM is given by the intersection $C_1(\text{CPM}) \cap C_2(\text{CPM}) \cap C_3(\text{CPM})$, where

(i) $C_1(\text{CPM})$ is the *equality constraint for mass*, i.e. the constraint defined by $X \in C_1(\text{CPM})$ iff X is non-empty, $X \subseteq M_p(\text{CPM})$ and for all $x, y \in X$: if $p \in P_x \cap P_y$, then $m_x(p) = m_y(p)$;

(ii) $C_2(\text{CPM})$ is the *extensivity constraint for mass* if there exists a function

$$\circ : \{P_x : x \in M_p(\text{CPM})\} \rightarrow \{P_x : x \in M_p(\text{CPM})\}$$

such that $X \in C_2(\text{CPM})$ iff X is non-empty, $X \subseteq M_p(\text{CPM})$ and for all $x \in X$: if $p \circ p' \in P_x$, then $m_x(p \circ p') = m_x(p) + m_x(p')$;

(iii) $C_3(\text{CPM})$ is the *equality constraint for force*, i.e. the constraint defined by $X \in C_3(\text{CPM})$ iff X is non-empty, $X \subseteq M_p(\text{CPM})$ and for all $x, y \in X$: if $p \in P_x \cap P_y, t \in T_x \cap T_y$ and $i \in \mathbb{N}$, then $f_x(p, t, i) = f_y(p, t, i)$.

Here, we append our sets and functions with subscripts x to denote the fact that it represents the set/function occurring in the potential model x.

Let us reflect for a moment on the above definition. The motivation for introducing the so-called equality constraint for mass has already been discussed in the example of the rocket-Moon-Earth system above. The extensivity condition, on the other hand, expresses that if we concatenate two particles p, p' to form a new particle $p \circ p'$ that the mass of this new particle equals the sum of the masses of its constituent particles. Finally, the equality constraint for force requires, akin to the equality constraint for mass, that any force f acting on some particle p at time t is independent of the ambient physical system the

[16] cf. (Balzer et al. 1987, 103-5).

particle is located in. Referring once again to the rocket example, we require that the force function denoting the gravitational force acting on the rocket in the rocket-Moon-Earth system is identical to the gravitational force function for the system consisting of the rocket and the entire solar system. Of course, we may now observe that this condition is *false* in this and many other examples. Balzer et al. (1987, 106) acknowledge this issue, but simply state that the constraint is meant to ensure that the forces in the new system do not become "too false" with respect to the old system.

Now, what are the links of the theory-element CPM? As noted above, links are intended to represent all theory upon which the theory under consideration implicitly relies in its scientific underpinning, e.g. chronometry and physical geometry for CPM. So we see that, before we could begin explicating all the different links we have for classical particle mechanics, we first require rational reconstructions of all its prerequisite theories. Moreover, it is by no means obvious which and only which theories would constitute a prerequisite theory of CPM. Thus, the full explication of the theory's links is most likely a rather arduous and daunting enterprise, one which Balzer et al. do not seek to undertake (1987, 106–7). For convenience, therefore, it assumed that each potential model of CPM in fact satisfies any given link to another theory, i.e. the links do not place any restriction on the theory-element of CPM.

Definition 2.4.26. The global link $GL(\text{CPM})$ of CPM is given by the class $M_p(\text{CPM})$.

This concludes the presentation of the theory-core of CPM and, consequently, also of the theory-element. One might wonder about the formal representation of the set I of intended applications. But as noted earlier, we cannot expect to be able to give sufficient conditions for something to constitute an *intended application.* It is a much-bemoaned observation in the philosophy of science that since we cannot meaningfully distinguish between one specific mathematical structure and another, isomorphic structure, we also cannot formulate any meaningful criteria with which to characterize the set of intended applications. Hence, any intended application will always have associated to it an isomorphic, unintended application.[17] Thus, the best we can hope to do in the structuralist approach, is specify the intended applications by means of ostention, i.e. $I = \{\text{the solar system, harmonic oscillators}, \ldots\}$. But this, of course, places no actual restrictions on the formal structure of the theory-element.

Let us conclude this subsection with a brief discussion of the role of linguistic formulations of theories. First, it should be remarked that Balzer et al. do not categorically oppose linguistic analyses of theories. Instead, they believe that reference to language and its resulting syntactic complications are simply too unpractical to allow in their account of theory-structure (1987, 306-7). Consequently, no single logical language is explicitly specified in the structuralist approach. However, while no syntax is specified with which to construct logical

[17] cf. footnote 13.

sentences, the non-logical vocabulary of possible languages *is* determined by the structuralist notion of theories.

Recall that any relation/function in a theory's potential models has an associated typification. This typification, in turn, specifies the number of open places for each relation or function symbol. Additionally, the type of the associated structure species of the potential models specifies (through specification of the number of base sets) the number of different sorts of variables that would be required for any appropriate language. So although there is no inherent logic of scientific theories that can be associated with the structuralist approach, it is not ruled out we might equipped theories' vocabularies with an appropriate syntax which might allow for an interesting linguistic analysis of scientific theories. This observation about the nature of language in the structuralist approach is of quite some import, and, as it turns out, will be vital in assessing some key aspects of the FWLA.

2.4.2 Intertheory Relations

Now that we are familiar with the structuralist analysis of scientific theories *by themselves*, we next turn to the study of **intertheory relations**, i.e. the study of the various types of relations holding between different scientific theories. This is a topic not only of crucial import to metascience generally construed, but also to the analysis of individual scientific theories. To see this, suppose we are presented by the seemingly innocuous task of formalizing classical mechanics. One way to go about this assignment would be to proceed in the same manner as the structuralists in their exposition of classical particle mechanics. In particular, this means we construct our account of classical mechanics around Newon's second law $F = ma$. Such a course of action, however, might be faced with a troubling riposte: it is not *classical mechanics* we have formalized, but only its *Newtonian* formulation. After all, classical mechanics may just as well be expounded in its Lagrangian and Hamiltonian formulations. Hence, we see that a proper understanding of the structure of the theory of classical mechanics cannot be obtained without also properly elucidating the interrelations of its various formulations.

And what to think of the relation between classical mechanics and its generalizations found in modern physics, such as relativistic dynamics? An analysis of the latter theory suggests that the type of spacetime presupposed by a physical theory is fundamental to understanding the structure underlying such a theory. This is, however, not at all evident when looking at classical mechanics in isolation. Indeed, we might wonder to what extent the structuralist construal of classical particle mechanics accurately reflects this particular feature of the theory. Again, we see that explicating a particular intertheory relation, this time between classical and relativistic dynamics, is required to fully comprehend the minutiae of the individual theories under consideration.

Recognizing the fundamental import of the study of intertheory relations to the metascientific enterprise, the structuralists devote a significant portion of their research program to the formal explication of these relations. Numerous types of relations between theories are considered including specialization, translation, equivalence and empirical equivalence. For our present purposes, however, it will suffice to consider here only a single, particularly important type of intertheory relations, i.e. the relation of **reduction**.

Traditionally, the notion of a theory T being reducible to some other theory T' is characterized along the lines of being able to deduce all laws of T from those of T' in some given system of inference. Clearly, such a description brings with it some obvious linguistic connotations. How, then, do the structuralist explicate the concept of reduction in a completely language-free manner? The answer lies in the following definition (Balzer et al. 1987, 277).

Definition 2.4.27. Let $T = (K, I)$ and $T' = (K', I')$ be two theory-elements with theory-cores $K = (M_p, M, M_{pp}, GC, GL)$ and $K' = (M'_p, M', M'_{pp}, GC', GL')$. Then a relation $\rho \subseteq M'_p \times M_p$ is said to *directly reduce*[18] T to T' if we have

(i) $rg(\rho) = M_p$,

(ii) for all $x' \in M'_p$: if $x' \in M'$ and $(x', x) \in \rho$, then $x \in M$,

(iii) for all $X' \subseteq dom(\rho)$: if $X' \in GC'$, then $\bar{\rho}(X') \in GC$,

(iv) for all $x \in M_p, x' \in M'_p$: if $x' \in GL'$ and $(x', x) \in \rho$, then $x \in GL$,

(v) for all $y \in I$: there exist $x \in M_p, x' \in M'_p, y' \in I'$ such that $(x', x) \in \rho$, $r'(x') = y', r(x) = y$,

where

- r and r' are the truncation functions for T and T' respectively, sending each potential model to its associated partial potential model;
- $dom(\rho) = \{x' \in M'_p : \exists x \in M_p : (x', x) \in \rho\}$,
- $rg(\rho) = \{x \in M_p : \exists x' \in M'_p : (x', x) \in \rho\}$,
- $\bar{\rho}(X') = \{x \in M_p : \exists x' \in X' : (x', x) \in \rho\}$.

The relation ρ itself is then called a *direct reduction* of T to T'.

Let us reflect upon the meaning of the various conditions. The first condition states that for each potential model of T we can find some ρ-related potential model of T'. Thinking of potential models as providing the conceptual bases for our theories, condition (i) can thus be paraphrased as saying that each concept of T is directly reducible to a concept in T'. Similarly, recalling that a theory's actual models can be thought to represent that theory's law statements, condition (ii) becomes the requirement that each law of T reduces to a law of T'. Conditions (iii) and (iv) merely state that the reduction relation should

[18] Balzer et al. reserves the proper term *reduces* for another, more general sense of reduction. This expanded notion, however, is not required for our current aims and is thus omitted from our discussion.

preserve constraints and links, while the final provision expresses that for each intended application y of theory-element T we can find a corresponding intended application y' of theory-element T'.

Oftentimes, the above definition proves to be of too great a generality to allow for the derivation of desirable results. Hence, Balzer et al. (1987, 278) introduce a number of ways in which we might strengthen the notion of a direct reduction. e.g.

Definition 2.4.28. Let T, T' be theory-elements and let ρ be a direct reduction between T and T'. Suppose that for all $x, y \in M_p$, $x' \in M'_p$: if $(x', x) \in \rho$ and $(x', y) \in \rho$ then $x = y$. In this case, we write $\rho : dom(\rho) \rightarrow M_p$.

Let us conclude here this exposition of the set-theoretic approach to theories. Contrasting this approach to the syntactic and state-space views we encountered earlier this chapter, one characteristic feature of the structuralist framework stands out above all else, i.e. the lack of language in nearly all of its major definitions. Even the notion of reduction, which has in previous analyses had always been inextricably linked with some underlying language,[19] is now formulated in entirely structural terms. But do such structural construals truly capture all the salient properties of the intuitive concepts they are meant to formalize? Pearce and Rantala (1983a) answer these questions in the negative, at least as far as intertheory relations are concerned. To remedy this situation, they proposes we adopt a new formal toolbox, viz. the framework of semi-abstract model theory, leading us into our next chapter.

[19] To give a short characterization: to reduce one theory to another meant the reduction of the one's *laws* to the other's. Laws, in turn, were most naturally construed as being statements expressed in some underlying language. Indeed, this type description essentially characterizes the positivist view on reduction associated to their theory-concept.

Chapter 3

First Wave of Logical Abstractivism

As evidenced by the preceding chapter, the structuralist program offers a host of formal explications of a wide variety of scientific concepts. Not all of these explications, however, were considered to be equally successful by all philosophers of science. In particular, the treatment of intertheory relations in the set-theoretic framework drew some criticism of David Pearce and Veikko Rantala, who instigated, during the late 1970s and '80s, a first wave of logical abstractivism, resulting in numerous papers and monographs at the hands of both researchers. In this chapter we will analyze more closely the contents of this FWLA. To this end, I first introduce the formalism from which Pearce and Rantala draw their most important concepts, i.e. the field of semi-abstract model theory (sec. 3.1). Following this exposition, we will look into the manner in which Pearce and Rantala apply this field of study to the metascientific enterprise (sec. 3.2).

3.1 Semi-Abstract Model Theory

We will begin our assessment of the FWLA by considering the formalism from which it draws inspiration, which we shall refer to here as *semi-abstract model theory*. This name is not standard in the literature, in which it is primarily known as the study of *abstract logic* or *abstract model theory*. The former of these will be used here for a different purpose, viz. to denote the objects of study of semi-abstract model theory, which I refer to here as *abstract logics*. As for the latter, I follow Diaconescu (2008, 3) in employing *abstract model theory* as an umbrella term for a variety of abstract approaches to model theory and referring to the framework of this chapter as being *half*- or *semi-abstract*.

The goal of this section will be to establish some familiarity with semi-abstract model theory, so that we are well-equipped to assess the usage of the discipline by Pearce and Rantala, which we consider in the subsequent section. For an impression of this field, we can turn to Barwise and Feferman (1985), in whose voluminous compendium book *Model-Theoretic Logics* we find an extensive introduction to many different facets of the discipline.[1] In the book's preface (1985, vii) , it is noted how semi-abstract model theory may be viewed as the culmination of three different strands of logical research:

- Work on **cardinality quantifiers**, mounted by Andrzej Mostowski, in the late 1950s.
- Studies of Tarski et al. into **infinitary languages** in the mid-1960s.
- Per Lindström's work on generalized quantifiers and **model-theoretic characterization theorems** in the late 1960s.

Even though semi-abstract model theory has expanded greatly since the 1960s, a cursory glance at (Barwise & Feferman 1985) reveals the three subjects still make up a significant portion of research in the field.

Since it is beyond the scope of this thesis chapter to give a detailed introduction to each of the above research topics, we will focus on one in particular, viz. the work on model-theoretic characterization theorems. As noted above, this line of research was first initiated by the Swedish logician Per Lindström, who in his (1969) proved a characterization of first-order logic that has since become known as *Lindström's theorem*. Today, the enterprise of finding Lindström-style characterization theorems, or simply *Lindström theorems*, for both systems of standard and non-standard logic remains an active area of investigation. In this section, we will look into a particularly abstract characterization result that will provide some insight into the generality of semi-abstract model theory. Now, before we can set about characterizing systems of logic, we must first agree on what exactly constitutes *a logic*. This will be the focus of the following subsection.[2]

3.1.1 Abstract Logics

Before anything else, an abstract approach to model theory will need to be grounded in an equally abstract conception of the notion of logical system.[3] That is, we will need to ask what properties a given (mathematical) structure needs to satisfy in order for it to be considered *a logic*. In semi-abstract model theory, these considerations are codified in the concept of an *abstract logic*. Before delving into its definition, we first require the following prerequisites.

[1] To my knowledge, (Barwise & Feferman 1985) remains to this day the most extensive work dedicated solely to the study of semi-abstract model theory.

[2] The following remarks are in order. Two of Lindström's original characterizations of first-order logic formed the subject of my bachelor's thesis. Thus, despite the paradigmatic status of these results for the field of semi-abstract model theory, I will not go into either of the proofs in the present text. Instead, the interested reader is referred to (Vos 2014). For sake of completeness, however, I will repeat the definition of the notion of *abstract logic* as found in (Vos 2014, 3–5), albeit in a somewhat modified form.

[3] The following subsection is based largely on (Ebbinghaus 1985, 26-45), with some minor modifications in terms of presentation.

Definition 3.1.1. Let $A = (A_s, \dots, R^A, \dots, f^A, \dots, c^A, \dots)$ be a model. We then call the signature $(s, \dots, R, \dots, f, \dots, c \dots)$ the *signature of* A and denote it as Σ_A.

Definition 3.1.2. Let Σ, Σ' be two signatures. An injection $\rho : \Sigma \to \Sigma'$ is said to be a *renaming* from Σ to Σ' if it preserves the type of each symbol (e.g. sort, relation, function) in Σ. From any Σ-model A, we can obtain a $\rho(\Sigma)$-model B by setting $B_{\rho(\sigma)} = A_\sigma$ for any sort symbol $\sigma \in \Sigma$ and $\rho(s)^B = s^A$ for any other symbol $s \in \Sigma$. We will denote such a model B by A^ρ.

We are now ready to articulate a formalized notion of logic:

Definition 3.1.3. An *abstract logic*[4] is a triple $(L, \mathit{Sent}_L, \models_L)$, with $L : \mathit{Sig} \to \mathcal{P}(\mathit{Sent}_L)$ a mapping from the set of many-sorted first-order signatures Sig to the powerset of Sent_L and $\models_L$ a relation between many-sorted first-order models and elements of Sent_L such that we have:

(i) For any model A and $\varphi \in \mathit{Sent}_L$: if $A \models_L \varphi$, then $\varphi \in L(\Sigma_A)$.

(ii) *Monotonicity Property.* For any two signatures Σ, Σ': if $\Sigma \subseteq \Sigma'$, then $L(\Sigma) \subseteq L(\Sigma')$.

(iii) *Finite Occurrence Property.* For any signature Σ and $\varphi \in L(\Sigma)$, there exists a smallest, finite signature $\Sigma_\varphi \subseteq \Sigma$ such that $\varphi \in L(\Sigma_\varphi)$.

(iv) *Isomorphism Property.* For any two models A, B and $\varphi \in \mathit{Sent}_L$: if we have $A \models_L \varphi$ and $A \cong B$, then $B \models_L \varphi$.

(v) *Reduct Property.* For any signature Σ, model A and $\varphi \in \mathit{Sent}_L$: if $\varphi \in L(\Sigma)$ and $\Sigma \subseteq \Sigma_A$, then $A \models_L \varphi$ if and only if $A|_\Sigma \models_L \varphi$.

(vi) *Renaming Property.* For any two signatures Σ, Σ', renaming $\rho : \Sigma \to \Sigma'$ and $\varphi \in \mathit{Sent}_L$: if $\varphi \in L(\Sigma)$, then there exists $\varphi^\rho \in L(\Sigma')$ such that for all Σ-models A we have $A \models_L \varphi$ if and only if $A^\rho \models_L \varphi^\rho$.

(vii) *Closure Properties.* For all signatures Σ, we have:

- for any $\varphi \in L(\Sigma)$, there exists $\psi \in L(\Sigma)$, denoted $\neg\varphi$, such that $\mathrm{Mod}_L^\Sigma(\psi) = \mathrm{Mod}(\Sigma) \backslash \mathrm{Mod}_L^\Sigma(\varphi)$;
- for any $\varphi, \psi \in L(\Sigma)$, there exists $\pi \in L(\Sigma)$, denoted $\varphi \wedge \psi$, such that $\mathrm{Mod}_L^\Sigma(\pi) = \mathrm{Mod}_L^\Sigma(\varphi) \cap \mathrm{Mod}_L^\Sigma(\psi)$;
- for any constant symbol $c \in \Sigma$ of sort σ and $\varphi \in L(\Sigma)$, there exists $\psi \in L(\Sigma \backslash \{c\})$, denoted $\exists x \varphi$, such that for all $\Sigma \backslash \{c\}$-models A we have $A \models_L \psi$ if and only if $(A, a) \models_L \varphi$ for some $a \in A_s$;

where $\mathrm{Mod}(\Sigma)$ denotes the class of all Σ-models, $\mathrm{Mod}^\Sigma(\varphi)$ denotes the class of Σ-models satisfying φ (in the sense of first-order logic) and $\mathrm{Mod}_L^\Sigma(\varphi)$ denotes the class $\{A \in \mathrm{Mod}(\Sigma) : A \models_L \varphi\}$.

[4] We may compare this terminology to that used in group theory, where one sometimes speaks of an *abstract group*.

Intuitively, we may think of the elements of $Sent_L$ as representing generalized instances of sentences and the set $L(\Sigma)$ as the set of Σ-sentences in the abstract logic L. Moreover, the relation $\models_L$ is readily seen to be a generalization of the satisfaction relation $\models$ of first-order logic.

In the remainder of this thesis, we will follow the usual abuse of notation in denoting an entire mathematical structure by one its constituents. In this case, we will consistently write L to denote the abstract logic $(L, Sent_L, \models_L)$. Elements of $L(\Sigma)$, for some signature Σ will, on occasion, be informally referred to as *sentences in* L. Similarly, the relation $A \models_L \varphi$ will be casually expressed by saying that A *makes true* φ *in* L.

Already, we can clearly discern the model-theoretic nature of the framework in definition 3.1.3. At no point in this definition, have we been concerned with anything resembling rules of deduction or proof trees. This is why in some sources, e.g. (Barwise & Feferman 1985), abstract logics are referred to as *model-theoretic languages*. Accordingly, the framework of semi-abstract model theory is well-suited to formulate and study abstract versions of various model-theoretic properties of first-order logic:

Definition 3.1.4. Let L be an abstract logic. Then we say that

- L is κ-*compact* for some infinite cardinal κ if for all signatures Σ and $\Gamma \subseteq L(\Sigma)$ of cardinality at most κ, we have that if every finite subset of Γ has a model then Γ itself has a model;

- L is *countably compact* or has the *countable compactness property* if it is $\aleph_0$-compact;

- L is *compact* or has the *compactness property* if it is κ-compact for all infinite cardinals κ;

- L has the *Löwenheim-Skolem property down to* κ, for some infinite cardinal κ, if for all signatures Σ and $\varphi \in L(\Sigma)$ we have that if φ has a model then φ has a model of cardinality at most κ;

- L has the *downward Löwenheim-Skolem property* if it has the Löwenheim-Skolem property for $\aleph_0$.

Let us now consider some examples of abstract logics. Clearly first-order logic, denoted by $L_{\omega\omega}$, satisfies the conditions of an abstract logic under the obvious interpretations of L and $\models_L$. Other instances of abstract logics, as we will see, can be found by considering extensions[5] of first-order logic. Indeed, many of the results we encounter in semi-abstract model theory hold true only under the assumption that the abstract logics under consideration are extensions of first-order logic.[6] Examples of such logics include:

[5] The notion of one logic *extending* another will be made precise below.

[6] This already gives us some idea of why we would want to refer to this formalism as *semi*-abstract model theory.

Second-Order Logic. Denoted L_2. This is the abstract logic obtained by

- letting $L(\Sigma)$ be the class of first-order Σ-sentences augmented, for each sort symbol $\sigma \in \Sigma$ and $n, m \in \mathbb{N}$, by variables and quantifiers ranging over relations of arity n and functions of arity m on the domain of sort σ.
- letting $\models_{L_2}$ be identical to the usual satisfaction relation $\models$, extended by the provision that sentences of the form $\forall_\sigma X \varphi(X)$ and $\exists_\sigma X \varphi(X)$, where X denotes a relation variable of sort σ and arity n, are considered true under $\models_{L_2}$ if $\varphi(X)[S]$ holds true[7] for respectively all and some relations $S \subseteq A_s^n$ and an analogous provision for quantification over function variables.[8]

Weak Second-Order Logic. Denoted L_{w2}. This is the abstract logic that is identical to L_2 except in that we have only relation variables and quantification over these variables is restricted over *finite* predicates.

Logics of form $L_{\omega\omega}(Q_\alpha)$. For any ordinal number α, we can define an abstract logic $L_{\omega\omega}(Q_\alpha)$, which denotes the abstract logic obtained by

- letting $L_{\omega\omega}(Q_\alpha)(\Sigma)$ be the class of first-order Σ-sentences augmented by the additional quantifier Q_α^σ for every sort symbol $\sigma \in \Sigma$,
- letting $\models_{L_{\omega\omega}(Q_\alpha)}$ be identical to the usual satisfaction relation $\models$, extended by the provision that sentences of the form $Q_\alpha^\sigma x \varphi(x)$ are considered true under $\models_{L_{\omega\omega}(Q_\alpha)}$ if there exist at least $\aleph_\alpha$ many elements a in domain A_s such that $\varphi(x)[a]$ is true in the model.

As we can see, abstract logics are abstract mainly in the sense that their *syntax* is left open. By contrast, the allowed class of models and the nature of signatures necessarily remain unaltered from their first-order counterparts. This, then, provides us with some ready examples of logics *not* qualifying as abstract logics:

Topological Logic. The 'logic of topological structures' fails to qualify as an abstract logic by sheer virtue of the fact it deals with *topological structures* as opposed to first-order models.

ω-logic. The logic known as ω-logic is obtained by fixing a single-sorted signature $\{\sqsubset\}$, with binary relation symbol $\sqsubset$, and considering only signatures Σ such that $\Sigma \supseteq \{\sqsubset\}$. Furthermore, for any such Σ-model A, we require that $A|_{\{\sqsubset\}} \cong (\mathbb{N}, <)$. As in the above case, this logic cannot be accommodated by the notion of abstract logic, since it does not allow us to restrict the kind of

[7] The formal definition of the expression $\varphi(X)[S]$ proceeds analogously to that of its first-order counterpart $\varphi(x)[a]$.

[8] This type of semantics for second-order logic is usually referred to as the *standard semantics* for second-order logic. Alternatively, one might also equip second-order logic with the more restrictive *Henkin semantics*, which we will not explore here.

models under consideration.

These counterexamples are somewhat disheartening, as they seem to greatly restrict the scope and applicability of semi-abstract model theory, especially if we are interested in a formalism that can help elucidate the logical structure of science. Fortunately, there are ways around this restriction which will allow us to make some meaningful statements about such non-standard logics. The case of topological logic will be considered in the next subsection. For ω-logic, we can at this stage introduce a straightforward generalization of the abstract logic concept. Just as we generalized the notion of sentence and the satisfaction relation to give shape to abstract logics in the sense of definition 3.1.3, so too can we allow for a generalization of the class of models.

Definition 3.1.5. A *generalized abstract logic* is a quadruple

$$(L, Sent_L, Mod_L, \models_L) \tag{3.1}$$

with $L = (L^1, L^2), L^1 : Sig \to \mathcal{P}(Sent_L)$ and $L^2 : Sig \to \mathcal{P}(Mod_L)$, such that it satisfies the conditions set out in definition 3.1.3 under the appropriate modifications.[9]

Henceforth, generalized abstract logics shall be referred to as simply *abstract logics*, subsuming the above usage of the term.

With this generalization in place, we see that the only place where abstract logics are still yoked to the first-order paradigm is in the set Sig of first-order signatures. Indeed, it is exactly in this regard that the formalisms which we will encounter in chapter 5 can be said to provide us with an improvement with respect to semi-abstract model theory. For the moment, however, we may note that definition 3.1.5 already aids us greatly in expanding the scope of the abstract logic concept.

Definition 3.1.6. Let L be an abstract logic and Σ_0 a single-sorted signature.[10] Let P be a unary relation symbol such that $P \notin \Sigma_0$ and let $\mathfrak{R}$ be a class of Σ_0-models closed under isomorphism. The logic $L(\mathfrak{R})$ is then defined as follows:[11]

For any signature Σ, let

$$L^1(\mathfrak{R})(\Sigma) = \begin{cases} L^1(\Sigma) & \text{if } \Sigma_0 \cup \{P\} \subseteq \Sigma, \\ \varnothing & \text{otherwise.} \end{cases} \tag{3.2}$$

and let

$$L^2(\mathfrak{R})(\Sigma) = \begin{cases} \mathcal{C}(\Sigma) & \text{if } \Sigma_0 \cup \{P\} \subseteq \Sigma, \\ \varnothing & \text{otherwise.} \end{cases} \tag{3.3}$$

[9] That is, by incorporating in the conditions the class $L_2(\Sigma)$ of Σ-models in L.

[10] For convenience, we consider here only the definition for single-sorted signatures.

[11] Note that the dependency on Σ_0 and P is suppressed in this notation. In practice, it will be clear what the intended interpretations of the symbols in $\Sigma_0 \cup \{P\}$ are supposed to be.

where $\mathcal{C}(\Sigma)$ denotes the class of all Σ-models A such that P^A is Σ_0-closed in A and $(A|_{\Sigma_0})|P^A \in \mathfrak{K}$.

Furthermore, for any signature Σ, Σ-model A and $\varphi \in L(\mathfrak{K})(\Sigma)$, let

$$A \models_{L(\mathfrak{K})} \varphi \text{ iff } A \in L^2(\mathfrak{K})(\Sigma_A), \varphi \in L^1(\mathfrak{K})(\Sigma_A) \text{ and } A \models_L \varphi. \tag{3.4}$$

We can now formally represent ω-logic as the abstract logic $L_{\omega\omega}(\Omega)$, where Ω is the class of all $\{\sqsubset\}$-models $(A, \sqsubset^A)$ such that $(A, \sqsubset^A) \cong (\mathbb{N}, <)$ for $\sqsubset$ a binary relation symbol.[12] In the same vein, we can now associate to any isomorphically closed class of first-order models $\mathfrak{K}_0$ a corresponding abstract logic $L(\mathfrak{K}_0)$ relative to some underlying abstract logic L. In particular, consider the following.

Definition 3.1.7. Let $\Sigma = \{\sqsubset\}$ be a signature with binary relation symbol $\sqsubset$ and let $\varphi_0 \in L_{\omega\omega}(Q_1)(\Sigma)$ be the sentence expressing that, for any Σ-model A, the relation $\sqsubset^A$ is a total order. Then we define an $\aleph_1$-*like ordering* to be any Σ-model A such that

$$A \models_{L_{\omega\omega}(Q_1)} \{\varphi_0, Q_1 x(x = x) \wedge \forall y \neg Q_1 x(x \sqsubset y)\}. \tag{3.5}$$

Now, let $\mathfrak{K}$ denote the class of all $\aleph_1$-like orderings. By isomorphism property of abstract logics, we may observe that $\mathfrak{K}$ is closed under isomorphism. Hence, from the construction in definition 3.1.6, we obtain an abstract logic $L_{\omega\omega}(\mathfrak{K})$.[13] We will return to this particular abstract logic in the discussion below.

Having seen various examples of abstract logics, let us next consider how different abstract logics may be relate to one another. One of the most natural questions to ask concerning two logics is whether one is **stronger** than the other. In other words: is the one logic an **extension** of the other? Ebbinghaus (1985, 43) discuss several different ways in which this type of relation can be made precise. First, let us introduce some preliminary notions.

Definition 3.1.8. Let L be an abstract logic, Σ be some signature and $\mathfrak{K}$ a class of Σ-models. Then $\mathfrak{K}$ is called an *elementary class* in L, or simply *EC* in L, if there exists some $\varphi \in L^1(\Sigma)$ such that $\mathfrak{K} = \mathrm{Mod}_L^{\Sigma}(\varphi)$.

Elementary classes are useful tools in determining the definability of properties of models. In some cases, however, it can be seen to be far too strict. In particular, any property involving two or more models of different signatures can never be specified by an elementary class, since all the models in such a class must necessarily be of the same signature. In light of this deficiency, we might desire a liberalization of definition 3.1.8. To this end, introduce

Definition 3.1.9. Let L be an abstract logic, Σ be some signature and $\mathfrak{K}$ a class of Σ-models. Then $\mathfrak{K}$ is called a *projective class* in L, or simply *PC* in L, if there exists some $\Sigma' \supseteq \Sigma$ with the same sort symbols as Σ and an elementary class $\mathfrak{K}'$ of Σ'-models such that $\mathfrak{K} = \mathfrak{K}'|_{\Sigma} := \{A|_{\Sigma} : A \in \mathfrak{K}'\}$.

[12] Here, we take $P^A = dom(\sqsubset^A) \cup rg(\sqsubset^A)$.

[13] The symbol P is interpreted in the same manner as for ω-logic.

In addition, it will prove useful to go one step further and allow a class of models to be definable as being a 'subclass' of some projective class. This is accomplished as follows.

Definition 3.1.10. Let L be an abstract logic, Σ be some signature and $\mathfrak{K}$ a class of Σ-models. Suppose Σ is single-sorted. Then $\mathfrak{K}$ is called a *relativized projective class* in L, or simply *RPC*, in L if there exists some $\Sigma' \supseteq \Sigma$, unary relation symbol $P \in \Sigma' \setminus \Sigma$ and elementary class $\mathfrak{K}'$ such that

$$\mathfrak{K} = \{(A|_{\Sigma})|P^A : A \in \mathfrak{K}' \text{ and } P^A \text{ is } \Sigma\text{-closed in } A\}. \tag{3.6}$$

Lastly, let us note that the above explications of definability all presuppose that we may only use a single sentence in the specification of model classes. This observation leads us to some straightforward liberalizations:

Definition 3.1.11. Let L be an abstract logic, Σ be some signature and $\mathfrak{K}$ a class of Σ-models. Then $\mathfrak{K}$ is called a Δ-*elementary class* in L, or simply EC_Δ in L, if there exists some $\Gamma \subseteq L^1(\Sigma)$ such that $\mathfrak{K} = \mathrm{Mod}_L^{\Sigma}(\Gamma)$.

Definition 3.1.12. Let L be an abstract logic, Σ be some signature and $\mathfrak{K}$ a class of Σ-models. Then $\mathfrak{K}$ is called a Δ-*projective class* in L, or simply PC_Δ in L, if there exists some $\Sigma' \supseteq \Sigma$ with the same sort symbols as Σ and a Δ-elementary class $\mathfrak{K}'$ of Σ'-models such that $\mathfrak{K} = \{A|_{\Sigma} : A \in \mathfrak{K}'\}$.

The above definition only applies to signature having a single sort, which suffices for present purposes. In the many-sorted case, note that we can weaken the condition on the signatures in definition 3.1.9 to state that Σ and Σ' can have possible different sort symbols to achieve a many-sorted counterpart of relativized projective classes.

Now, a natural first candidate for the notion of an extension is the following:

Definition 3.1.13. Let L, L' be abstract logics. Then L' is an *extension* of L, symbolically $L \leq L'$, if we have that every elementary class in L is also an elementary class in L'. We say L' is *equivalent* to L, symbolically $L \equiv L'$, if $L \leq L'$ and $L' \leq L$.

Definition 3.1.14. Let L, L' be abstract logics. Then L' is said to be *stronger* than L, symbolically $L < L'$, if $L \leq L'$ and not $L \equiv L'$.

Indeed, the above notion of extension conforms well with many of our intuitive judgments about logics' relative strength: L_2 is stronger than L_{w2}, L_{w2} is stronger than $L_{\omega\omega}$, etc. In some instances, however, we may find this particular notion of extension to be less than satisfactory. Consider, for example, the abstract logics $L_{\omega\omega}(Q_1)$ and $L_{\omega\omega}(\mathfrak{K})$, where $\mathfrak{K}$ once again denotes the class of all $\aleph_1$-like orderings. Now, observe (Ebbinghaus 1985, 43):

Proposition 3.1.15. *Let $\mathfrak{K}$ denote the class of $\aleph_1$-like orderings. Then we have that $L_{\omega\omega}(\mathfrak{K}) \leq L_{\omega\omega}(Q_1)$.*

Proof. For the first statement, let $\mathfrak{R}$' be an elementary class in $L_{\omega\omega}(\mathfrak{R})$ for some signature Σ. This means that there exists some $\varphi \in L^1_{\omega\omega}(\mathfrak{R})(\Sigma)$ such that $\mathfrak{R}' = \{A \in \text{Mod}(\Sigma) : A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi\}$. Looking at the construction in definition 3.1.6, we see that, necessarily, φ must be a first-order sentence over a signature Σ containing at least one binary relation symbol $\sqsubset$. Moreover, by condition (3.4), we know that $A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi$ is equivalent to

$$A \in L^2_{\omega\omega}(\mathfrak{R})(\Sigma_A), \varphi \in L^1_{\omega\omega}(\mathfrak{R})(\Sigma_A) \text{ and } A \models_{L_{\omega\omega}} \varphi. \tag{3.7}$$

which, in turn, reduces to

$$A \in \mathcal{C}(\Sigma_A), \varphi \in L^1_{\omega\omega}(\Sigma_A) \text{ and } A \models_{L_{\omega\omega}} \varphi, \tag{3.8}$$

with $\mathcal{C}(\Sigma_A)$ the class of all Σ_A-models A' such that $(A'|_{\{\sqsubset\}})|P^{A'} \in \mathfrak{R}$. Then, by definition of $\mathfrak{R}$ and condition (3.8), we can thus express the conditions placed on A to be as follows:

- A makes true φ in $L_{\omega\omega}$,
- $(A|_{\{\sqsubset\}})|P^A$ is an $\aleph_1$-like ordering.

Now, let ψ denote the following Σ-sentence in $L_{\omega\omega}(Q_1)$:

$$\varphi \wedge \forall x \forall y \forall z (Px \wedge Py \wedge Pz \to \varphi_0(x,y,z) \wedge Q_1 x(x = x) \wedge \forall y \neg Q_1 x(x \sqsubset y)), \tag{3.9}$$

in which $\varphi_0(x, y, z)$ denotes the formula

$$(x \sqsubseteq y \wedge y \leq x \to x = y) \wedge (x \sqsubseteq y \wedge y \sqsubseteq z \to x \sqsubseteq z) \wedge (x \sqsubseteq y \vee y \sqsubseteq x). \tag{3.10}$$

where the symbol $\sqsubseteq$ is defined in terms of $\sqsubset$ in the obvious way. Clearly, we have that $A \models_{L_{\omega\omega}(Q_1)} \psi$. Thus, we see that for any Σ-model A we have that $A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi$ implies $A \models_{L_{\omega\omega}(Q_1)} \psi$.

Conversely, let A be a Σ-model such that $A \models_{L_{\omega\omega}(Q_1)} \psi$. It is then readily verified that A makes true φ in $L_{\omega\omega}$ and A is an $\aleph_1$-like ordering. Hence, condition 3.7 is satisfied and we infer $A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi$. Thus, we have established:

$$A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi \text{ if and only if } A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi, \tag{3.11}$$

i.e.

$$\mathfrak{R}' = \{A \in \text{Mod}(\Sigma) : A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi\} = \{A \in \text{Mod}(\Sigma) : A \models_{L_{\omega\omega}(Q_1)} \psi\}. \tag{3.12}$$

We conclude $\mathfrak{R}$' is an elementary class in $L_{\omega\omega}(Q_1)$ as well and hence that $L_{\omega\omega}(\mathfrak{R}) \leq L_{\omega\omega}(Q_1)$.

■

Of course, the question whether the converse statement of proposition 3.1.15 holds true now immediately suggests itself. In other words, do we have that $L_{\omega\omega}(Q_1) \leq L_{\omega\omega}(\mathfrak{R})$? We answer this question in the negative:

Proposition 3.1.16. *Let $\mathfrak{R}$ denote the class of $\aleph_1$-like orderings. Then we have $L_{\omega\omega}(Q_1) \not\leq L_{\omega\omega}(\mathfrak{R})$.*

Proof. Suppose $L_{\omega\omega}(Q_1) \leq L_{\omega\omega}(\mathfrak{R})$. Consider the elementary class

$$\mathfrak{R}' := \{A \in \mathrm{Mod}(\Sigma) : A \models_{L_{\omega\omega}(Q_1)} Q_1 x(x = x)\}. \tag{3.13}$$

By assumption, then, $\mathfrak{R}'$ is also *EC* in $L_{\omega\omega}(\mathfrak{R})$, i.e. there exits a signature Σ and $\varphi \in L_{\omega\omega}(\mathfrak{R})(\Sigma)$ such that

$$\mathfrak{R}' = \{A \in \mathrm{Mod}(\Sigma) : A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi\}. \tag{3.14}$$

Examining (3.13), we see that $\mathfrak{R}$' is exactly the class of all Σ-models of cardinality at least $\aleph_1$. Applying this knowledge to (3.14), we see that

$$A \models_{L_{\omega\omega}(\mathfrak{R})} \varphi \text{ if and only if } A \text{ has cardinality at least } \aleph_1. \tag{3.15}$$

Now, by the construction given in definition 3.1.6 and (3.15), we obtain the statement that A has cardinality at least $\aleph_1$ if and only if

$$A \in L^2_{\omega\omega}(\mathfrak{R})(\Sigma_A),\ \varphi \in L^1_{\omega\omega}(\mathfrak{R})(\Sigma_A) \text{ and } A \models_{L_{\omega\omega}} \varphi. \tag{3.16}$$

In other words, we have that A has at least $\aleph_1$ elements if and only if $A \models_{L_{\omega\omega}} \varphi$ and $(A|_{\{\sqsubset\}})|P^A$ is an $\aleph_1$-like ordering. So, in particular, we have that if A has at least $\aleph_1$ elements then $(A|_{\{\sqsubset\}})|P^A$ is an $\aleph_1$-like ordering. But it is clear that this is too strong an implication. Hence, we have arrived at a contradiction and conclude that $\mathfrak{R}'$ cannot be *EC* in $L_{\omega\omega}(\mathfrak{R})$.

■

A brief comparison of propositions 3.1.15 and 3.1.16 might now lead us to think that $L_{\omega\omega}(\mathfrak{R})$ and $L_{\omega\omega}(Q_1)$, in all likelihood, will have some rather different model-theoretic properties. By virtue of the above propositions, we can of course conclude this to be true to some extent. In another sense, however, we see that the two abstract logics have, in fact, quite comparable expressive capabilities. This is captured by the following statement:

Proposition 3.1.17. *Let $\mathfrak{R}$ denote the class of $\aleph_1$-like orderings. Suppose that. Then a class is RPC in $L_{\omega\omega}(Q_1)$ if and only if it is RPC in $L_{\omega\omega}(\mathfrak{R})$, provided we do not use the symbol $\sqsubset$ used to denote the $\aleph_1$-like orderings in the signatures for $L_{\omega\omega}(Q_1)$.*

Proposition 3.1.17 now suggests to us another way for comparing the expressive power of abstract logics, one not grounded in elementary classes, but in (relativized) projective classes instead:

Definition 3.1.18. Let L, L' be two abstract logics. We write $L \leq_{(R)PC} L'$ if every (relativized) projective class in L is also a (relativized) projective class in L'. If both $L \leq_{(R)PC} L'$ and $L' \leq_{(R)PC} L$, then we write $L \equiv_{(R)PC} L'$.

By proposition 3.1.17, we immediately see that $L_{\omega\omega}(\mathfrak{R}) \equiv_{RPC} L_{\omega\omega}(Q_1)$.

Obviously, the ordering specified in definition 3.1.17 is of a quite more liberal nature than the one given in definition 3.1.13. In what sense, then, can we consider $\leq_{RPC}$ to represent an adequate ordering of abstract logics? As it turns out, $\leq_{RPC}$ does, in fact, possess a number of desirable properties we might expect from an ordering relation on abstract logics. More specifically, we find that several model-theoretic properties 'descend down' the ordering $\leq_{RPC}$. For example, consider compactness. We have the following theorem for $L_{\omega\omega}(\mathfrak{R})$.

Theorem 3.1.19. *Let $\mathfrak{R}$ be the class of $\aleph_1$-like orderings. Then $L_{\omega\omega}(\mathfrak{R})$ has the countable compactness property.*

Now, suppose we are interested in the compactness properties of the logic $L_{\omega\omega}(Q_1)$. Rather than expending much effort in the attempting to establish such properties from scratch, we can use the fact that $L_{\omega\omega}(Q_1) \leq_{RPC} L_{\omega\omega}(\mathfrak{R})$, in tandem with theorem 3.1.19, to establish swiftly:

Theorem 3.1.20. *$L_{\omega\omega}(Q_1)$ has the countable compactness property.*

With the above considerations in mind, it is clear that there does not exist a single uniform way of defining the notion of a logical extension. In particular, we have seen that a case can be made for defining extensions in terms of (relativized) projective classes as opposed to elementary classes despite the former being inherently a more crude ordering. For the remainder of this thesis, we will conform to the notion of extension as presented in definition 3.1.13. Nevertheless, it is my hope that the preceding discussion will have already given the reader some feeling for the framework of semi-abstract model theory and an appreciation for the level of generality that comes along with it.

3.1.2 Characterization Results

In this subsection, we will take up some of the most characteristic results of semi-abstract model theory, viz. **Lindström-style characterization theorems** for a number of different abstract logics.[14] The paradigmatic result here is *Lindström's theorem*:

Theorem 3.1.21. *Let L be an abstract logic with $L^2(\Sigma)$ the class of first-order models for any signature Σ. Suppose L extends $L_{\omega\omega}$ and has the countable compactness property and the downward Löwenheim-Skolem property. Then L is equivalent to $L_{\omega\omega}$.*

In slogan form, this theorem tells us that first-order logic is the strongest logic with the countable compactness and downward Löwenheim-Skolem properties. That is, this theorem provides us with a *characterization* of first-order logic with respect to the class of its extensions.

Lindström's theorem has given rise to a host of model-theoretic investigations hoping to establish similar characterization results, both for $L_{\omega\omega}$ and other,

[14] This subsection is based primarily on (1985, 91–120).

non-standard systems of logic. Let us start here by considering an alternative to theorem 3.1.21, which seeks to characterize first-order logic in terms of the *Tarski union property.*

Definition 3.1.22. Let L be an abstract logic, Σ be a signature, A and B be Σ-models. Then we say A is an *elementary submodel* of B relative to L, symbolically: $A \preceq_L B$, if A is a submodel of B and it holds for every sentence φ in L and $a \in A$ that $A \models_L \varphi$ iff $B \models_L \varphi$. The model B is then called an *elementary extension* of A relative to L.

Definition 3.1.23. Let L be an abstract logic. Then an *elementary chain* relative to L is a set of models $\{A_n\}_{n\in\mathbb{N}}$ such that $A_n \preceq_L A_{n+1}$ for all $n \in \mathbb{N}$.

Definition 3.1.24. Let L be an abstract logic. Then L is said to have the *Tarski union property* if for every elementary chain relative to L it holds that

$$A_k \preceq_L \bigcup_{n\in\mathbb{N}} A_n \tag{3.17}$$

for every $k \in \mathbb{N}$, where the model

$$\bigcup_{n\in\mathbb{N}} A_n = (A, R_1^A, \ldots, R_p^A, f_1^A, \ldots, f_q^A, c_1^A, \ldots, c_r^A) \tag{3.18}$$

is defined by setting

- $A = \cup_{n\in\mathbb{N}} A_n$,
- $R_i^A = \cup_{n\in\mathbb{N}} R_i^{A_n}$ for each $1 \leq i \leq p$,
- $f_i^A = \cup_{n\in\mathbb{N}} f_i^{A_n}$ for each $1 \leq i \leq q$,
- $c_i^A = c_i^{A_n}$ for each $1 \leq i \leq r$.

Informally, an abstract modal logic L has the Tarski union property if the union of any elementary chain is an elementary extension of each model in the chain relative to L.

To obtain a characterization for $L_{\omega\omega}$, we require the following preliminaries:

Definition 3.1.25. Let L be an abstract logic and A some model for L. Then the *L-theory* of A is defined to be the set

$$\mathrm{Th}_L(A) := \{\varphi \in \mathit{Sent}_L : A \models_L \varphi\}. \tag{3.19}$$

Furthermore, the *L-diagram* of A is taken to be the set $D_L(A) := \mathrm{Th}_L(A, (a)_{a\in A})$.

Definition 3.1.26. Let L be an abstract logic and A and B some models for L. Then A and B are said to be *L-elementarily equivalent*, symbolically $A \equiv_L B$, if $\mathrm{Th}_L(A) = \mathrm{Th}_L(B)$.

Lemma 3.1.27. *Let L, L' be abstract logics such that L' extends L. Suppose L' has the compactness property and, for any models A, B, we have:*

$$A \equiv_L B \textit{ implies } A \equiv_{L'} B \tag{3.20}$$

Then L' is equivalent to L.

Proof. Consider an arbitrary elementary class in L', given by some L'-sentence φ. First, note that for any model A, we have:

$$A \models_{L'} \varphi \text{ iff } A \models_L \bigvee_{B \models_{L'} \varphi} \psi_B \tag{3.21}$$

with

$$\psi_B := \bigwedge \{\psi \in \mathit{Sent}_L : B \models_L \psi\}. \tag{3.22}$$

For the if-direction, it suffices to note that if $A \models_{L'} \varphi$ then $a \models_L \psi_A$. Conversely, if we have $A \models_L \psi_B$ for some model B such that $B \models_{L'} \varphi$, then it readily follow that $A \equiv_L B$. By (3.20), we then also have $A \equiv_{L'} B$ and therefore $A \models_{L'} \varphi$.

Next, we need to verify that the conjunction and disjunction in (3.21) can be replaced by finite ones, since this would establish that the elementary class in L' given by φ is then an elementary class in L as well. We consider here only the case for disjunction. The conjunction case proceeds analogously.

Suppose there exists no finite sequence of L-sentences $\psi_{B_1}, \ldots, \psi_{B_n}$ such that we have

$$A \models_{L'} \varphi \text{ iff } A \models_L \psi_{B_1} \vee \ldots \vee \psi_{B_n} \tag{3.23}$$

and $B_i \models_{L'} \varphi$ for every $1 \leq i \leq n$. Now, consider the set

$$\{\varphi\} \cup \{\neg \psi_B : B \models_{L'} \varphi\}. \tag{3.24}$$

By our preceding assumption, every finite subset of this set will have at least one model. It then follows from the compactness property, however, that the entire set has a model. But this would contradict (3.21). We conclude that the elementary class of φ in L' is indeed also an elementary class in L and, consequently, that L' is equivalent to L. ■

Lemma 3.1.28. *Let L be an abstract logic and A some model for L. Then the elementary extensions of A relative to L are exactly the reducts of the models of $D_L(A)$ to Σ_A.*

Lemma 3.1.29. *Let L be an abstract logic with $L^2(\Sigma)$ the class of first-order models for any signature Σ. Suppose L extends $L_{\omega\omega}$ and has the compactness property. Then, for any model A and set of L-sentences Γ, we have: if $\mathit{Th}_{L_{\omega\omega}}(A) \cup \Gamma$ has a model then there exists a model B such that*

$$A \preceq_{L_{\omega\omega}} B \textit{ and } B \models_L \Gamma. \tag{3.25}$$

Now, we can state:

Theorem 3.1.30. *Let L be an abstract logic with $L^2(\Sigma)$ the class of first-order models for any signature Σ. Suppose L extends $L_{\omega\omega}$, has the compactness property and has the Tarski union property. Then L is equivalent to $L_{\omega\omega}$.*

Proof. Suppose L is not equivalent to $L_{\omega\omega}$. Then, by lemma 3.1.27, there exist some models A, B and sentence φ in L such that

$$A \equiv_{L_{\omega\omega}} B, \;\; A \models_L \varphi \text{ and } B \models_L \neg\varphi. \tag{3.26}$$

Next, we inductively construct a sequence $(A_n)_{n\in\mathbb{N}}$ as follows. Set $A_0 = A$. Furthermore, let A_1 be an arbitrary model such that $A \preceq_{L_{\omega\omega}} A_1$ and $A_1 \models_L \neg\varphi$. The existence of such an A_1 is guaranteed by lemma 3.1.29 and the observation that $\mathrm{Th}_{L_{\omega\omega}}(A) \cup \{\neg\varphi\} = \mathrm{Th}_{L_{\omega\omega}}(B) \cup \{\neg\varphi\}$ has at least one model by (3.26). Now, for any $n > 1$, define the model A_{n+1} to be an arbitrary model such that $(A_n, (a)_{A_{n-1}}) \preceq_{L_{\omega\omega}} (A_{n+1}, (a)_{a\in A_{n-1}})$ and $(A_{n+1}, (a)_{a\in A_{n-1}}) \models_L D_L(A_{n-1})$. The existence of such an A_{n+1} is guaranteed, as before, by lemma 3.1.29 as well as the observation that there exists a model satisfying

$$\mathrm{Th}_{L_{\omega\omega}}(A_n, (a)_{a\in A_{n-1}}) \cup D_L(A_{n-1}) = D_L(A_{n-1}). \tag{3.27}$$

From this construction, it is evident that $A_n \preceq_{L_{\omega\omega}} A_{n+1}$ for any $n \in \mathbb{N}$. Furthermore, it follows from lemma 3.1.28 that $A_{n-1} \preceq_L A_{n+1}$ for any $n \geq 1$. That is:

$$A_0 \preceq_{L_{\omega\omega}} A_1 \preceq_{L_{\omega\omega}} A_2 \preceq_{L_{\omega\omega}} \dots \tag{3.28}$$

$$A_0 \preceq_L A_2 \preceq_L A_4 \preceq_L \dots \tag{3.29}$$

$$A_1 \preceq_L A_3 \preceq_L A_5 \preceq_L \dots \tag{3.30}$$

Consider now the union $C := \bigcup_n A_{2n} = \bigcup_{2n+1} A_{2n+1}$. Since L has the Tarski union property, we know that $A_k \preceq_L C$ for all $k \in \mathbb{N}$. In particular, we have $A_0 = A \preceq_L C$ and $A_1 \preceq_L C$. However, recall that also have $A \models_L \varphi$ and $A_1 \models_L \neg\varphi$. But since C is an L-elementary extension of both A and A_1, we infer $C \models_L \varphi$ as well as $C \models_L \neg\varphi$. Thusly, we have arrived at a contradiction. ■

Again, note that this is but an example of how first-order logic can be characterized by some of its well-known model-theoretic properties. Different properties, such as the *Robinson property* and the *omitting types property*, might similarly be employed for the characterization of $L_{\omega\omega}$. Let us consider a final example for the case of $L_{\omega\omega}$, which will prove to be significant for our considerations below. Once again, some preliminary notions are required:

Definition 3.1.31. Let A and B be Σ-models for some signature Σ. A *partial isomorphism* between A and B is a relation I on pairs of finite sequences $(a_1, \dots, a_n)$, $(b_1, \dots, b_n)$ of elements of A and B of the same length such that the following hold:

- $\varnothing I \varnothing$, where $\varnothing$ denotes the empty sequence.

- If $(a_1, \dots, a_n)I(b_1, \dots, b_n)$ then $(A, a_1, \dots, a_n)$ and $(B, b_1, \dots, b_n)$ satisfy the same atomic Σ'-sentences in $L_{\omega\omega}$, where $\Sigma' = \Sigma \cup \{c_1, \dots, c_n\}$ and $c_1, \dots, c_n$ denote fresh constant symbols not occurring in Σ.
- *Back-and-Forth Property.* If $(a_1, \dots, a_n)I(b_1, \dots, b_n)$ then for all $a \in A$ there exists $b \in B$ such that $(a_1, \dots, a_n, a)I(b_1, \dots, b_n, b)$ and vice versa.

We call A and B *partially isomorphic*, symbolically $A \cong_p B$, if there exists a partial isomorphism between A and B.[15]

Definition 3.1.32. Let L be an abstract logic. L is said to be *invariant under partial isomorphisms* or $\cong_p$*-invariant* if for any signature Σ, $\varphi \in L^1(\Sigma)$ and $A, B \in L^2(\Sigma)$, we have

$$A \cong_p B \text{ and } A \models_L \varphi \text{ implies } B \models_L \varphi. \tag{3.31}$$

Alternatively, abstract logics satisfying (3.31) are also said to have the *Karp property*, though in this thesis we will adhere to the above terminology.

We are now ready to formulate:

Theorem 3.1.33. *Let L be an abstract logic with $L^2(\Sigma)$ the class of first-order models for any signature Σ. Suppose L extends $L_{\omega\omega}$, has the countable compactness property and is invariant under partial isomorphisms. Then L is equivalent to $L_{\omega\omega}$.*

The proof of theorem 3.1.33 overlaps almost completely with the proof of theorem 3.1.21 as expounded in (Vos 2014). In fact, once a proof has been provided for theorem 3.1.33, we can then suffice by proving the downward Löwenheim-Skolem property implies invariance under partial isomorphisms in abstract logics satisfying the appropriate conditions to establish theorem 3.1.21.

Now, the observant reader may note that the above characterization theorems do *not* exhaust the level of generality offered by the notion of an abstract logic. More specifically, we have the following, interconnected points:

(i) The preceding theorems were all aimed at situating *first-order logic* within a certain class of abstract logics. Thus, a natural continuation of our model-theoretic investigations would now be to look for characterization theorems for **different abstract logics**.

(ii) All the above characterization results have explicitly required the class of models $L^2(\Sigma)$ to be identical to the class of first-order Σ-models for any signature Σ. This can be problematic if we want to transcend the realms of standard predicate logic. Therefore, it is of great interest to examine whether we can formulate Lindström-style characterization results for **different classes of structures**.

[15] This definition of partial isomorphisms has been adapted from (Vos 2014, 6–7).

Let us, for the remainder of this subsection, explore how we can extend the formal machinery of semi-abstract model theory along the two lines mentioned above. As it turns out, we can formulate an abstract characterization result resolving both of the points (i) and (ii).

Of all the characterization theorems considered so far, it turns out that theorem 3.1.33 lends itself the best for the desired generalizations. This is quite fortunate: it might be argued on independent grounds[16] that characterizations in terms of invariance properties are to be preferred over other types of characterization results.[17] The generalization of theorem 3.1.33 will naturally require suitably generalized ancillary concepts. Hence, we have:

Definition 3.1.34. Let L be an abstract logic, φ a sentence in L and R a binary relation on the class of all models of L. Then φ is called *invariant under* R or R*-invariant* if for any models A, B in L we have

$$ARB \text{ and } A \models_L \varphi \text{ implies } B \models_L \varphi. \tag{3.32}$$

Definition 3.1.35. Let L be an abstract logic and R a binary relation on the class of all models of L. Then we say that L is *invariant under* R or R*-invariant* if every sentence φ in L is R-invariant.

Finally, we have:

Definition 3.1.36. Let R be a binary relation on the class of all first-order models. Then we say that R is *invariantly definable with definable finite approximations* if for any signature Σ, there exists a signature $\Sigma' \supseteq \Sigma$ and first-order Σ'-sentences $\varphi_0, \varphi_1, \ldots$ such that for any two Σ-models A, B we have

ARB iff there is an expansion C of (A, B) to Σ' such that $C \models \{\varphi_i : i \in \mathbb{N}\}$,[18]

and, for any $n \in \mathbb{N}$ and binary relation R_n defined by

AR_nB iff there is an expansion C' of (A, B) to Σ' such that $C' \models \{\varphi_i : i \leq n\}$,

the following properties hold true:

- R_n is an equivalence relation on $\mathfrak{N}^\Sigma$.
- For any $D \in \mathfrak{N}^\Sigma$, there exists a first-order Σ-sentence ψ^n_D such that for any $E \in \mathfrak{N}^\Sigma$ we have

$$DR_nE \text{ if and only if } E \models \psi^n_D.$$

We can now establish the following generalization of theorem 3.1.33.

[16] That is, independent from the mathematical considerations of semi-abstract model theory.

[17] We will return to this argument briefly in chapter 5.

[18] The notation (A, B) denotes the model obtained from concatenating the tuples A and B. In case A and B have overlapping sorts, we simply replace in (A, B) one of the corresponding domains and the associated relations, functions and constants by isomorphic copies.

Theorem 3.1.37. *For any signature Σ, let $\mathfrak{R}^{\Sigma}$ some Δ-elementary class in $L_{\omega\omega}$ and let R be binary relation on the class of all first-order models such that the following hold:*

(i) $R \cap \mathrm{Mod}(\Sigma)^2$ *is an equivalence relation on* $\mathfrak{R}^{\Sigma}$ *for any signature* Σ.

(ii) For any renaming $\rho : \Sigma \to \Sigma'$ *and* Σ'*-models* A, B*, we have that* ARB *implies* $A|^{\rho}_{\Sigma} R B|^{\rho}_{\Sigma}$.

(iii) For any two models A *and* B *we have* ARB *implies* $A, B \in \mathfrak{R}^{\Sigma}$ *for some signature* Σ.

(iv) R *is definably invariant with definable finite approximations.*

Furthermore, let $L^{R}_{\omega\omega}$ *denote the abstract logic such that, for any signature* Σ, $L^{R1}_{\omega\omega}(\Sigma)$ *consists of all first-order* Σ*-sentences invariant under* R *and* $L^{R2}_{\omega\omega} = \mathfrak{R}^{\Sigma}$.

Now, let L *be an abstract logic with* $L^2(\Sigma) = \mathfrak{R}^{\Sigma}$*, for any signature* Σ*, and suppose that* L *extends* $L^{R}_{\omega\omega}$*, has the compactness property and is invariant under* R*. Then* L *is equivalent to* $L^{R}_{\omega\omega}$.

Proof. The proof of this theorem proceeds trough a proof of a stronger statements, viz. that any L-elementary classes can be separated by an $L^{R}_{\omega\omega}$-elementary class.[19] The theorem then follows from applying this more general result to any L-elementary class along with its complement.

Fist, note that $L^{R}_{\omega\omega}$ indeed qualifies to be an abstract logic by conditions (i)–(ii) above. Moreover, it is readily verified that $L^{R}_{\omega\omega}$ inherits the compactness property from $L_{\omega\omega}$.

Now, let φ, ψ be sentences over some signature Σ and consider the L-elementary classes $\mathrm{Mod}_L(\varphi), \mathrm{Mod}(\psi)$. Then, note that we have for any $n \in \mathbb{N}$ and model A:

$$\text{if } A \models_L \varphi \text{ then } A \models_L \bigvee_{B \models_L \varphi} \psi^n_B, \tag{3.33}$$

where ψ^n_B is the sentence as specified in definition 3.1.36. To see this, note that R_n is an equivalence relation on $\mathfrak{R}^{\Sigma}$ for any $n \in \mathbb{N}$ and thus at least we will always $A \models_L \psi^n_A$ for any A. Now, employing a similar compactness argument as used in the proof of lemma 3.21, we can replace the disjunction in (3.33) by a finite disjunction. Thus, (3.33) can be reformulated, for any $n \in \mathbb{N}$ model A, as:

$$\text{if } A \models_L \varphi \text{ then } A \models_L \chi_n, \tag{3.34}$$

where $\chi_n := \psi^n_{A^n_1} \vee \ldots \vee \psi^n_{A^n_m}$.

We shall now show the set $\{\chi_i : i \in \mathbb{N}\} \cup \{\psi\}$ has no model in $\mathfrak{R}^{\Sigma}$. Once this is established, it follows from the compactness property that there exists some finite subset $\{\chi_0 \wedge \ldots \wedge \chi_n\} \cup \{\psi\}$ having no model. That is:

$$\mathrm{Mod}_L(\chi_0 \wedge \ldots \wedge \chi_n) \cap \mathrm{Mod}_L(\psi) = \varnothing. \tag{3.35}$$

[19] Two classes A, B are said to be *separated* by a third class C if $A \subseteq C$ and $B \cap C = \varnothing$

In addition, we know from (3.34) that

$$\mathrm{Mod}_L(\varphi) \subseteq \mathrm{Mod}_L(\chi_0 \wedge \ldots \wedge \chi_n). \tag{3.36}$$

Hence, $\mathrm{Mod}_L(\chi_0 \wedge \ldots \wedge \chi_n)$ would then be the desired class separating $\mathrm{Mod}_L(\varphi)$ and $\mathrm{Mod}_L(\psi)$.

Let us thus suppose there exists some model B of $\{\chi_i : i \in \mathbb{N}\} \cup \{\psi\}$ in $\mathfrak{R}^\Sigma$. By definition of the χ_i, we then have, for every $i \in \mathbb{N}$, a model A_n in $\mathfrak{R}^\Sigma$ such that $A_n \models_L \varphi$ and $B \models_L \psi^n_{A_n}$. Since R is definably invariant with definable approximations, the latter of these facts implies that $A_n R_n B$ for every $n \in \mathbb{N}$. This, in turn, implies there exists an expansion of C of the model (A_n, B) satisfying $\{\varphi_i : i \leq n\} \cup \{\varphi, \psi\}$ for every $n \in \mathbb{N}$. By compactness, we then also have, for appropriate choice of A', B', an expansion C' of (A', B') satisfying $\{\varphi_i : i \in \mathbb{N}\} \cup \{\varphi, \psi\}$ such that $A' \models_L \varphi$ and $B' \models_L \psi$. But this also means that $A'RB'$. Consequently, we obtain from the R-invariance of L (and hence of φ) that $B' \models_L \varphi$ as well and hence $B \in \mathrm{Mod}_L(\varphi) \cap \mathrm{Mod}_L(\psi)$. This, however, contradicts our original assumption that $\mathrm{Mod}_L(\varphi)$ and $\mathrm{Mod}_L(\psi)$ are disjoint.

■

Theorem 3.1.37, also known as the *abstract maximality theorem*, provides us with a characterization result for any abstract logic whose classes of models are given by first-order Δ-elementary classes and thus represents a significant improvement on theorem 3.1.33, which we now obtain as a special case by setting $R = \cong_p$ and letting $\mathfrak{R}^\Sigma$ denote simply the class of all first-order Σ-models for every signature Σ. This is, however, but one of myriad possibilities. In particular, we can specialize the theorem in such a manner that it provides us with a characterization of the fragment of first-order logic for topological structures as the strongest abstract logic satisfying compactness and invariance under *partial homeomorphisms*.

The abstract maximality theorem is of great interest for the metascientist in search of model-theoretic formalization tools. It shows us that semi-abstract model theory, as represented by field of study aimed at finding Lindström-style characterization theorems, is by no means limited to the investigation of (extensions of) first-order logic. What makes this observation particularly salient is that past formalization efforts for metascience have been regarded as being too unwieldy, exactly because of the difficulties involved in formalizing theories of science in the basic language of first-order logic.

What the abstract maximality theorem now shows us is that, once we have broken free of the chains of first-order fixation, that there still exist ample possibilities for finding powerful formal tools in the field of logic. The application of theorem 3.1.37 to the case of topological logic, for instance, might be of great value to the formalization of theories from physics. Moreover, if we expand the scope of our inquires into semi-abstract model theory beyond the domain of characterization theorems, we also find much work being done in the logic of probability, topology and Borel structures.[20]

[20] cf. part E of (Barwise & Feferman 1985). While of arguably of greater immediate rele-

3.2 The Pearce-Rantala Approach

The goal of this section is to give an exposition of the approach to metascience as expounded by David Pearce and Veikko Rantala in numerous papers and monographs over the course of the 1970s and '80s which I collectively refer to as the *first wave of logical abstractivism* (FWLA)[21] due to the account's reliance on semi-abstract model theory. To my knowledge, the earliest published account of the FWLA can be found in (Rantala 1978) after which Pearce and Rantala set out a general program for applying semi-abstract model theory to the study of metascience in their (1983a). We find the most formally elaborate account in (Pearce 1985), which will serve as the basis of the presentation below.

Of particular concern to Pearce (1985) is elucidation of intertheory relations, most notably the relation of **translation**. It is the notion of translation, Pearce claims, that holds the key to understanding intertheory relations more generally. In particular, the claim is that by understanding translation we open the door to understanding **reduction** relations between scientific theories. Now, we already encountered one explication of the notion of reduction in subsection 2.4.2. But as already noted there, this encounter was rather atypical in the sense that the reduction relation considered was one between *structures* rather than a relation between linguistic entities. Attuned to this discrepancy, the aim of Pearce (1985) is now twofold:

- Using concepts from semi-abstract model theory, to develop a notion formalizing the concept of translation between languages.[22]
- Using this notion of translation, to show how we may rework the set-theoretic explication of reduction so as to incorporate language into its account, mending the structuralist and linguistic views on reduction.

We will take up each of these points in turn, with subsection 3.2.1 focused on the explication of a model-theoretic concept of translation and subsection 3.2.2 showcasing the reworking of the structuralist framework. Finally, in subsection 3.2.3, I will draw upon the preceding exposition in order to distill the primary tenets of the Pearce-Rantala approach as well as offer a preliminary appraisal.

vance to issues pertaining to metascience, I have omitted a discussion of these logics from my presentation here because such an exposition would make, in my view, less apparent the level of generality afforded to us by semi-abstract model theory.

[21] On occasion, I shall employ the term *Pearce-Rantala approach* as a synonym.

[22] Note that we have tacitly made a switch here from translation construed as a relation between *theories* to translation as a relation between *languages*. For our present purposes, this makes little difference, since translation merely serves as a gateway to reformulating the intertheory relation of reduction. If one wished, however, we could just as well define translation between theories in terms of translation between languages. What exactly is meant here by *language* will be made clear in the next subsection.

3.2.1 Semantic Systems

Central to the framework of (Pearce 1985) is the notion of an **abstract semantic system**.[23] Let us start by examining this concept.

Definition 3.2.1. Let L be an abstract logic. Then an *abstract semantic system* for L is a pair $S = (\Sigma, \mathfrak{R})$, where Σ is a many-sorted first-order signature and $\mathfrak{R} \subseteq L^2(\Sigma)$ closed under isomorphism.

Immediately, then, we see a first, tentative link between the Pearce-Rantala approach to metascience and the framework of semi-abstract model theory. Some caveats are in order, however, which unveil already some ways in which the FWLA makes only limited use of the generality offered by semi-abstract model theory. Any abstract logic L in the Pearce-Rantala approach is subject to the following restrictions:

- The class Mod_L of models of L is a subclass of all first-order models.
- L is an extension of first-order logic, i.e. $L_{\omega\omega} \leq L$.

Intuitively, we may view an abstract semantic system as a signature plus some interpretation provided by the class $\mathfrak{R}$. That is, we might consider it to be an *interpreted language*. Our objective is now to expound a notion of translation between abstract semantic systems. This is accomplished as follows:

Definition 3.2.2. Let $S = (\Sigma, \mathfrak{R})$ and $S' = (\Sigma', \mathfrak{R}')$ be abstract semantic systems for some abstract logic L. Then S' is called a *conservative extension* of S if we have $\Sigma \subseteq \Sigma'$ as well as $\mathfrak{R}' = \mathfrak{R}|_{\Sigma}$.

Definition 3.2.3. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma', \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L. Then S is called a *common conservative extension* of S_1 and S_2 if it holds that S is conservative extension of both S_1 and S_2.

Pearce (1985, 109–10) then defines translation as follows:

Definition 3.2.4. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L and suppose that S is a common conservative extension of S_1 and S_2. Then two sentence $\varphi \in L^1(\Sigma_1), \psi \in L^1(\Sigma_2)$ are said to be *intertranslatable*[24] relative to S if it holds for every $A \in \mathfrak{R}$ that

$$A \models_L \varphi \text{ if and only if } A \models_L \psi. \tag{3.37}$$

Definition 3.2.5. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L and suppose that S is a common conservative extension of S_1 and S_2. Then a sentence $\varphi \in L^1(\Sigma)$ is called *translatable* into S_2 relative to S if there exists a sentence $\psi \in L^1(\Sigma_2)$ such that φ and ψ are intertranslatable relative to S.

[23] Adapted from the original terminology of *general semantical system* (Pearce 1985 109).

[24] Here, I employ the term *intertranslatable* as opposed to Pearce's phrasing of φ *is translated by* ψ because I believe it to better accentuate the symmetrical natural of the notion.

Definition 3.2.6. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L and suppose that S is a common conservative extension of S_1 and S_2. Then S is said to be a *translation*[25] of S_1 into S_2 if every sentence $\varphi \in L^1(\Sigma_1)$ is translatable into S_2.

Having established the notion of translation as a relation between abstract semantic systems, Pearce (1985, 114–24) next seeks to identify which sufficient conditions on the systems' respective classes of models need to hold true in order for this relation to hold as well. The motivation for such an exercise may be fathomed by looking ahead to subsection 3.2.2: if we wish to combine the present notion of translation with the structuralist notion of reduction, it is of interest to examine how translation can be expressed in terms of structures, i.e. models. Central to this undertaking are the following notions, dating back to Feferman (1974):

Definition 3.2.7. Let L be an abstract logic. Then a *compactness property* for L is a pair $\delta = (F, I)$ of sets F, I of L-sentences such that

(i) I contains F,

(ii) $\varphi \in Sent_L$ implies $\{\varphi\} \in F$,

(iii) $\Gamma \in F \cap L^1(\Sigma)$ implies $\mathrm{Mod}^L(\Gamma)$ is elementary,

(iv) $\Gamma \in F$ implies $\Gamma \cap L^1(\Sigma) \in F$ for all signatures Σ,

(v) I is closed under unions and renaming,[26]

(vi) if $\Gamma \in I, \Gamma \subseteq L^1(\Sigma)$ and for every $\Gamma' \subseteq \Gamma$ we have that $\Gamma' \in F$ implies $\mathrm{Mod}^L(\Gamma) \neq \varnothing$ then $\mathrm{Mod}^L(\Gamma) \neq \varnothing$,

where $\mathrm{Mod}^L(\Gamma)$ denotes the class of all models satisfying Γ.

Definition 3.2.8. Let L be an abstract logic, $\mathfrak{R} \subseteq L^2(\Sigma)$ for some signature Σ and δ a compactness property for L. Then $\mathfrak{R}$ is called a δ-elementary class, or simply EC_δ, in L if we have $\mathfrak{R} = \mathrm{Mod}^L_\Sigma(\Gamma)$ for some $\Gamma \in I \cap L^1(\Sigma)$. Similarly, $\mathfrak{R}$ is called a δ-projective class in L, or simply PC_δ in L if there exists some $\Sigma' \supseteq \Sigma$ and $\mathfrak{R}'$ such that $\mathfrak{R}'|_\Sigma = \mathfrak{R}$ and $\mathfrak{R}'$ is EC_δ in L.

Intuitively, conditions (i)–(v) of definition 3.2.7 express that F and I 'behave' as finite and infinite sets of sentences. Condition (vi) may then be viewed as generalization of the the standard compactness property for abstract logics. Indeed, taking in this definition F to consist of all finite sets of first-order sentences and and I the set of all possible infinite first-order sentences, we recover the compactness property for $L_{\omega\omega}$. However, in cases where the usual compactness fails

[25] In Pearce's terminology: *full translation*.

[26] For any abstract logic L, a renaming $\rho : \Sigma \to \Sigma$ will induce in the obvious manner a map $\rho' : L^1(\Sigma) \to L^1(\Sigma)$. Closure under renaming can then be defined as closure under the induced map ρ'.

for an abstract logic L, the presence of some (weaker) compactness property δ for L might still enable us to prove some desirable results.

Next, we require a further property of abstract logics. This property may be compared to the closely related Robinson property encountered in subsection 3.1.2. We have:

Definition 3.2.9. Let L be an abstract logic. Then we say L has the *interpolation property* if, for any signature Σ, any two disjoint projective classes $\mathfrak{K}_1, \mathfrak{K}_2 \subseteq L^2(\Sigma)$ are separated by an elementary class $\mathfrak{K} \subseteq L^2(\Sigma)$.

To further demonstrate the manner in which compactness properties can be of help in find useful features of abstract logic which might otherwise lack nice properties as compactness and interpolation, consider:

Definition 3.2.10. Let L be an abstract logic and δ a compactness property for L. Then we say that L has the *δ-interpolation property* if the conditions of definition 3.2.9 hold true with projective classes replaced by δ-projective classes in the hypothesis.

That δ-interpolation indeed represents a weaker property than usual interpolation is made explicit by the proposition below (Feferman 1974, 159):

Proposition 3.2.11. *Let L be an abstract logic, $\delta = (F, I)$ a compactness property for L and suppose that L has the interpolation property. Then L has the δ-interpolation property.*

Proof. Let $\mathfrak{K}_1, \mathfrak{K}_2$ be two disjoint δ-projective classes of signatures Σ_1, Σ_2 respectively and let $\Sigma = \Sigma_1 \cap \Sigma_2$ and $\Sigma' = \Sigma_1 \cup \Sigma_2$. For $i \in \{1, 2\}$, let $\Gamma_i \in I$ be a set of L-sentences over Σ' such that $\mathfrak{K}_i = \text{Mod}^L_{\Sigma'}(\Gamma_i)|_{\Sigma_i}$. Since I is closed under unions, we know that $\Gamma = \Gamma_1 \cup \Gamma_2 \in I$ as well. Now, we know Γ cannot have a Σ'-model A. For suppose there exists such a model A. Then $A|_{\Sigma_i} \in \text{Mod}^{\Sigma'}_L(\Gamma_i)$. But it would then hold that the Σ-model $A_0 := (A|_{\Sigma_1})|_\Sigma = (A|_{\Sigma_2})|_\Sigma \in \mathfrak{K}_1 \cap \mathfrak{K}_2$, resulting in a contradiction.

So we indeed have $\text{Mod}^L_{\Sigma'}(\Gamma) = \varnothing$. By property (vi) of definition 3.2.7, we then know that there must also exist some $\Gamma^0 \subseteq \Gamma$ such that $\Gamma^0 \in F$ and $\text{Mod}^L_{\Sigma'}(\Gamma^0) = \varnothing$. Now, let $X_i = \Gamma^0 \cap L^1(\Sigma_i)$. By closure of F under intersections, we then also have $X_i \in F$. It then follows, by property (iii) of definition 3.2.7, that $\text{Mod}^L_{\Sigma_i}(X_i)$ is EC in L. Consequently, we have that $\mathfrak{K}^*_i := \text{Mod}^L_{\Sigma_i}(X_i)|_\Sigma$ is PC in L. Moreover, since $X_i \subseteq \Gamma_i$ and there exists no Σ'-model for Γ^0, we see that $\mathfrak{K}_i \subseteq \mathfrak{K}^*_i$ and $\mathfrak{K}^*_1 \cap \mathfrak{K}^*_2 = \varnothing$. That is, $\mathfrak{K}^*_1, \mathfrak{K}^*_2$ are two disjoint projective classes in L. Hence, since L has the interpolation property, we infer there exists some elementary class in L separating $\mathfrak{K}^*_1$ and $\mathfrak{K}^*_2$. A fortiori, this elementary class then also separates $\mathfrak{K}_1$ and $\mathfrak{K}_2$. ■

Let us now continue to the next order of business. Recall that our aim is to understand how translations between semantic systems can be represented by relations between models. To this end, let us consider:

Definition 3.2.12. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L with $\Sigma_1 \cap \Sigma_2 = \varnothing$[27] and $\Sigma = \Sigma_1 \cup \Sigma_2$. Then let $\mathcal{R} \subseteq L^2(\Sigma_1) \times L^2(\Sigma_2)$ denote the relation given by $\mathcal{R}(A, B)$ if and only if $(A, B) \in \mathfrak{R}$.

Definition 3.2.13. Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L, with $\Sigma_1 \cap \Sigma_2 = \varnothing$ and $\Sigma = \Sigma_1 \cup \Sigma_2$. Then the relation $\mathcal{R}$ is called *elementary*, *projective*, *δ-elementary* and *δ-projective* if $\mathfrak{R}$ is EC, PC, EC_δ, PC_δ in L respectively.

We now have:

Proposition 3.2.14. *let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for some abstract logic L, with $\Sigma_1 \cap \Sigma_2 = \varnothing$ and $\Sigma = \Sigma_1 \cup \Sigma_2$. Then S is a common conservative extension of S_1 and S_2 if and only if $dom(\mathcal{R}) = \mathfrak{R}_1$ and $rg(\mathcal{R}) = \mathfrak{R}_2$.*

Proof. First, suppose S is a common conservative extension of S_1 and S_2. We then know $\mathfrak{R}|_{\Sigma_i} = \mathfrak{R}_i$ for $i \in \{1, 2\}$. Now, if $A \in dom(\mathcal{R})$, then $(A, B) \in \mathfrak{R}$ for some B. Clearly, then, we have $A \in \mathfrak{R}|_{\Sigma_1}$ and hence $A \in \mathfrak{R}_1$. Conversely, if $A \in \mathfrak{R}_1$ then $A \in \mathfrak{R}|_{\Sigma_1}$ and hence there exists some B such that $(A, B) \in \mathfrak{R}$, i.e. $\mathcal{R}(A, B)$. Thus, we have $A \in dom(\mathcal{R})$. We conclude $dom(\mathcal{R}) = \mathfrak{R}_1$. The identity $rg(\mathcal{R}) = \mathfrak{R}_2$ is obtained in an analogous fashion.

Next, suppose $dom(\mathcal{R}) = \mathfrak{R}_1$ and $rg(\mathcal{R}) = \mathfrak{R}_2$. Clearly, it suffices to establish $dom(\mathcal{R}) = \mathfrak{R}|_{\Sigma_1}$ and $rg(\mathcal{R}) = \mathfrak{R}|_{\Sigma_2}$. Now, $A \in dom(\mathcal{R})$ is equivalent to the existence of some B such that $\mathcal{R}(A, B)$ which, in turn, is equivalent to the existence of some B such that $(A, B) \in \mathfrak{R}$, that is $A \in \mathfrak{R}|_{\Sigma_1}$. Hence, $dom(\mathcal{R}) = \mathfrak{R}|_{\Sigma_1}$. In as similar vein, we obtain the second identity as well. Thus, we conclude S is a common conservative extension of S_1 and S_2. ■

At last, we are now able to express the so-called *uniform reduction theorem*, first formulated by Feferman (1974, 161–2), which reads:

Theorem 3.2.15. *Let L be an abstract logic with the interpolation property and compactness property δ. Furthermore, let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for L, with $\Sigma_1 \cap \Sigma_2 = \varnothing$. Suppose that $\mathcal{R}$ is a δ-projective relation. Then, for any $\psi \in L^1(\Sigma_2)$ such that for all A, B, $B' \in L^2(\Sigma)$ we have*

$$\mathcal{R}(A, B) \text{ and } \mathcal{R}(A, B') \text{ implies } (B \models_L \psi \iff B' \models_L \psi) \tag{3.38}$$

there exists $\varphi \in L^1(\Sigma_1)$ such that for all A, $B \in L^2(\Sigma)$

$$\mathcal{R}(A, B) \text{ implies } (A \models_L \varphi \iff B \models_L \psi). \tag{3.39}$$

[27] In practice, this condition imposes on us no real restrictions: if two signatures Σ_1, Σ_2 were to overlap, we can simply rename the symbols in, say, Σ_2, by means of a suitable renaming $\rho : \Sigma_2 \to \Sigma_3$ in such a way that $\Sigma_1 \cap \Sigma_3 = \varnothing$.

Informally, condition (3.38) states that a certain property, expressed by ψ, is invariant on the 'range' of $\mathcal{R}$. Condition (3.39), on the other hand, states that for the sentence ψ we have a corresponding, semantically equivalent sentence φ on the 'domain' of $\mathcal{R}$. Theorem 3.2.15 then express the fact that to every property, expressed by some ψ, invariant on the range of $\mathcal{R}$, there exists a corresponding, equivalent property, expressed by φ, on the domain of $\mathcal{R}$. Such a sentence φ is then called a *uniform reduction* of the property ψ of B to A.

Pearce (1985, 120–1) now invokes the uniform reduction theorem to prove the following two statements concerning translations:

Proposition 3.2.16. *Let L be an abstract logic with the interpolation property and compactness property δ. Furthermore, Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for L, with $\Sigma_1 \cap \Sigma_2 = \varnothing$ and $\Sigma = \Sigma_1 \cup \Sigma_2$, and suppose that S is a common conservative extension of S_1 and S_2. Finally, suppose that $\mathcal{R}$ is a δ-projective relation. Then, for any $\psi \in L^1(\Sigma_2)$ such that for all A, B, $B' \in L^2(\Sigma)$ we have*

$$\mathcal{R}(A, B) \text{ and } \mathcal{R}(A, B') \text{ implies } (B \models_L \psi \iff B' \models_L \psi) \tag{3.40}$$

there exists a sentence $\varphi \in L^1(\Sigma_1)$ such that φ and ψ are intertranslatable relative to S.

Proof. Let φ be the sentence associated to ψ by the uniform reduction theorem. By definition 3.2.4, this φ will be intertranslatable with ψ relative to S if the consequent of (3.39) holds for all Σ-models C in $\mathfrak{R}$. Hence, let $C \in \mathfrak{R}$. Then we can write $C = (A, B)$ for some A, B. By definition of $\mathcal{R}$, it follows that $\mathcal{R}(A, B)$. By implication (3.39), we can then infer $A \models_L \varphi$ iff $B \models_L \psi$. Now, note that since φ and ψ are sentences over the signatures Σ_1 and Σ_2 respectively, we can use the reduct property of abstract logics to conclude that $C \models_L \varphi$ iff $C \models_L \psi$. Hence, φ and ψ are intertranslatable relative to S. ■

Proposition 3.2.17. *Let L be an abstract logic with the interpolation property and compactness property δ. Furthermore, Let $S = (\Sigma, \mathfrak{R}), S_1 = (\Sigma_1, \mathfrak{R}_1)$ and $S_2 = (\Sigma_2, \mathfrak{R}_2)$ be abstract semantic systems for L, with $\Sigma_1 \cap \Sigma_2 = \varnothing$ and $\Sigma = \Sigma_1 \cup \Sigma_2$, and suppose that S is a common conservative extension of S_1 and S_2. Finally, suppose that $\mathcal{R}$ is a δ-projective relation. Then, if we have for A, B, $B' \in L^2(\Sigma)$ that*

$$\mathcal{R}(A, B) \text{ and } \mathcal{R}(A, B') \text{ implies } B \equiv_L B' \tag{3.41}$$

then it holds that S is a translation of S_2 into S_1.

Proof. Let $\psi \in L^1(\Sigma_2)$. Applying (3.41) to ψ, we obtain condition (3.40) from proposition 3.2.16. It then follows that there exists some $\varphi \in L^1(\Sigma)$ such that φ and ψ are intertranslatable relative to S. We conclude S is indeed a translation of S_2 into S_1. ■

Proposition 3.2.17 now provides us with the desired characterization of the notion of translation in terms of relations between models. In the next subsection, we will be putting this characterization to work, viz. be applying it to obtain a model-theoretic reformulation of the structuralist account of reduction.

3.2.2 Structuralism Revisited

In the preceding subsection we have seen how Pearce (1985) defines a notion of translation between abstract semantic systems. To show how this notion carries over to reduction between theories, we first need a way to articulate a notion of translation for theories in the structuralist sense. Now, recall that theories in the structuralist framework are given by particular classes of structures referred to as *theory-elements.* Thinking back to subsection 2.4.1, remember that a theory-element was given by a pair $T = (K, I)$, with theory-core $K = (M_p, M, M_{pp}, GC, GL)$ and intended applications $I \subseteq M_{pp}$. Central to this definition was the class of potential models M_p, in terms of which all other components were defined. The class M_p, in turn, was specified by means of a structure species, consisting of a type τ and a number of set-theoretic sentences expressing additional structural properties.

How are we now to introduce the linguistic concept of translation into the seemingly language-free framework of the structuralists? The key here is to note that, while structuralism is indeed free of *syntactic* notions, *language* still factors into the account through the type τ of any given structure species. By a trivial procedure, we can associate to any type τ a signature Σ_τ containing the same information as τ and vice versa. Now, drawing on our experience in dealing with translations between abstract semantic systems in the preceding subsection, we are well-poised to introduce corresponding relations between theory-elements or, more precisely, theory-cores:

Definition 3.2.18. Let L be an abstract logic and K, K' be theory-cores with structure species τ, τ' respectively. Furthermore, let $\mathcal{R} \subseteq M'_p \times M_p$ be a binary relation. Then $\mathcal{R}$ is said to be a *translation* of K into K' relative to L if

- $dom(\mathcal{R}) \subseteq M'_p$ and $rg(\mathcal{R}) = M_p$,
- for every $\psi \in L^1(\Sigma_\tau)$ there exists $\varphi \in L^1(\Sigma_{\tau'})$ such that for, for all models A, B, we have:

$$\mathcal{R}(A, B) \text{ implies } (A \models_L \varphi \iff B \models_L \psi). \tag{3.42}$$

At last, then, let us consider how to relate this notion of translation for theory-cores to the structuralist conception of reduction. A straightforward application of the uniform reduction theorem yields

Proposition 3.2.19. *Let L be an abstract logic with compactness property δ and K, K' be theory-cores with structure species τ, τ' respectively. Furthermore, let $\mathcal{R} \subseteq M'_p \times M_p$ be a binary relation. Suppose that*

- *L has the interpolation property,*
- *$\mathcal{R}$ is a δ-projective relation,*
- *$dom(\mathcal{R}) \subseteq M'_p$ and $rg(\mathcal{R}) = M_p$,*

- *for all models A, B, B', we have: if $\mathcal{R}(A, B)$ and $\mathcal{R}(A, B')$ then $B \equiv_L B'$.*

Then $\mathcal{R}$ is a translation of K into K' relative to L.

Recalling definitions 2.4.27 and 2.4.28 from the previous chapter, we thus obtain:

Proposition 3.2.20. *Let L be an abstract logic with compactness property δ and $T = (K, I), T' = (K', I')$ be theory-elements with structure species τ, τ' respectively. Furthermore, let $\mathcal{R} \subseteq M_p' \times M_p$ be a binary relation. Suppose that*

- *L has the interpolation property,*
- *$\mathcal{R}$ is a δ-projective relation,*
- *$\mathcal{R}$ is a unique direct reduction from T to T'.*

Then $\mathcal{R}$ is a translation from K into K' relative to L.

Proof. First, note that $dom(\mathcal{R}) \subseteq M_p'$ and $rg(\mathcal{R}) = M_p$ holds true for any direct reduction. Next, observe that the uniqueness of $\mathcal{R}$ guarantees that $\mathcal{R}(A, B)$ and $\mathcal{R}(A, B)$ implies $B = B'$ and hence $B \equiv_L B'$. The statement then follows trivially from proposition 3.2.19. ■

As we can see, proposition 3.2.20 provides us with some sufficient conditions for which the structuralist notion of reduction coincides with Pearce's notion of translation as defined on theory-cores. This characterization, however, may be noted to of questionable value. Indeed, the number of conditions that need to be imposed on the relation $\mathcal{R}$ in order for propositions 3.2.19 and 3.2.20 to hold seems to make theses results rather trivial in nature. The implications of this observation for our valuation of the FWLA are considered in more detail in chapter 5. In the final subsection of the current chapter, we will take a bird's-eye view of our findings up until this point and use this to contemplate the general characteristics of the Pearce-Rantala approach.

3.2.3 FWLA: An Appraisal

Having examined the formal machinery behind the FWLA as well as some modest applications to the structural concept of reduction, let us take a moment now to evaluate the approach to metascience offered to us by Pearce and Rantala. As one may recall from section 1.1, I defined logical abstractivism to be the conviction that the field of metascience may benefit from an abstract conception of logic. Note, however, that this conviction by itself still leaves open to a great extent how exactly we ought to apply an abstract conception of logic to the metascientific enterprise or even what an *abstract conception of logic* ought to entail in the first place. In answering these questions, we will be led to the defining features of the FWLA.

Starting from the central tenet of logical abstractivism, the approach of Pearce and Rantala is further pinned down by the following additional characteristics. In order of increasing specificity:

(i) The choice of **semi-abstract model theory** as the formalism with which to explicate an abstract conception of logic.

(ii) The choice to include only **extensions of first-order logic** within the scope of the abstract logics.

(iii) The focus on **logical liberalization**, i.e. the focus on liberalizing formal definitions of metascientific concepts from a specific underlying logic, allowing such definitions to involve any logical system meeting some specific set of desiderata.

(iv) The choice to use the aforementioned formal machinery to try and modify the **structuralist approach** to scientific theories by incorporating language into its framework.

The choices (i)-(iv) can be thought to characterize the FWLA within the wider class of frameworks we might label *logically abstractivist.*[28]

Let us now dwell briefly on each of the above points and their ramifications for the FWLA. Point (i), I believe, may be deemed as fairly inevitable. Looking at (Beziau 2007, 3–19), we see that much of the early work involving abstract conceptions of logic emphasized the syntactic and proof-theoretic rather than the semantic and model-theoretic nature of logic. At the same time, such proof-theoretic vantage points were focused heavily on *propositional logics*. Consequently, the choice for a model-theoretic framework may have simply been predetermined by the need to consider the inner structure of logical sentences. This, in turn, would then have largely fixed the choice for semi-abstract model theory: to my knowledge, it remained the only substantive model-theoretic approach to universal logic until at least the 1990s.[29]

The motivation for choice (ii), on the other hand, may already be less obvious. There is no a priori reason to limit our logical abstractivism to only those logics extending first-order logic. As seen in subsection 3.1, there is, even within the confines of semi-abstract model theory, much to be gained by letting our abstract logics range over a wider class of logics than merely the extensions of first-order logic. Nevertheless, such a restriction may be understood, by noting that, also in section 3.1, we frequently encountered just this restriction as a prerequisite of most of the established results.

Point (iii), by contrast, is more contingent in nature. This choice, i.e. the choice to utilize semi-abstract model theory so as to free metascience from the limitations of any single logical system, may be appreciated by reflecting on the interplay between logic and metascience during the first half of the twentieth century. Indeed, the expressive limitations of first-order logic was oft-lamented by those seeking to apply logical methods to the formalization of science. In this light, a tendency towards logical liberalization becomes more understandable. After all, we might reason, the alternative to logical liberalization would be

[28] Of course, the existence of such a 'wider class' is at this stage still very much hypothetical.

[29] Recall from section 1.1 that the term *universal logic* is used here as an umbrella term for frameworks concerned with the abstract study of logics.

to choose simply a different, but fixed underlying logic for our metascientific investigations.

One of two eventualities might then occur. If, on the one hand, this logic turns out to be inadequate for this assignment, e.g. by lacking sufficient expressive power or by being too ill-behaved from a model-theoretic perspective, our account of metascience falls with it. If, on the other hand, such a logic would prove to be well-suited to its task, we can simply conduct our metascience within this single logical system, dispelling the need for semi-abstract model theory and logical abstractivism altogether. Such a dilemma, however, would be duplicitous. I will argue in chapter 5 that there is a natural way out of this predicament, that is at present concealed by the adherence to point (ii) of the Pearce-Rantala system.

Lastly, let us note that choice (iv) also may be deemed as entirely contingent. Throughout their body of work, Pearce and Rantala show a clear preference to the structuralist school of metascience and accordingly base much of their own account on the structuralist framework. Admittedly, structuralism was and still is one of most extensive accounts of metascience. However, at the time of the FWLA, there was still ample selection from a variety of alternative frameworks within the structural view of theories, most notably the state-space approach. In chapter 5, I will argue that the state-space approach, in fact, provides us with an attractive alternative to structuralism and lends itself naturally to serve as the basis for a logically abstractivist approach to metascience.

In sum, we see that the FWLA can be characterized by a sequence of non-trivial choices and that a commitment to the ideology of logical abstractivism should be no means be equated with the specific brand of abstractivism advocated by Pearce and Rantala. How effective is their approach when it comes to the elucidation of the structure of science? I will postpone such value judgments to chapter 5. Let me suffice here by articulating my strong conviction that logical abstractivism has much more to offer to the field of metascience than we have witnessed so far.

Chapter 4

Contemporary Metascience

Any new approach to the study of metasciene would do well to take note of not only preceding schools of metascience, but the contemporary activities in the field as well. This is why, before entering into the critical analysis of traditional metascience and the first wave of logical abstractivism, the present chapter will first introduce us to the most prominent present-day account of scientific theory-structure. This approach, as we will see, relies heavily on concepts and methods from the highly abstract mathematical discipline of category theory. As before, we start our discussion by first considering the formalization of theories in isolation (sec. 4.1), after which we turn to the problem of intertheory relations (sec. 4.2). In each of case, we are presented with a number of options for formalization, each of which we will take up and examine extensively.

4.1 Theories as Categories

So far, our discussion of metascience has been concerned mainly with the program of structuralism, as formulated during the 1970s and '80s. This program, while enjoying substantial levels of development, has failed to consolidate into a widely accepted and fruitful metascientific framework: while rigorous in nature, the account nevertheless is sorely lacking in terms of mathematical depth. This is why, during the 2010s, a new approach to metascience started to emerge at the hands of Hans Halvorson and Dimitris Tsementzis (2016) and James Weatherall (2016).[1] The driving force behind this new account was to investigate how we could take notions from the abstract field of category theory and give them metascientific application. As before, *metascience* here refers primarily to the study of scientific theory-structure. Hence, we might dub this account the *categorical* approach to the structure of scientific theories. This adjective will also be employed more generally to refer to concepts of or relating to category theory.

[1] One may note a similarity here between these efforts and earlier work by Albert Visser (2004) on intertheory relations in the context of mathematical logic.

In this more general context, I will on occasion also employ *category-theoretic* as a synonym for *categorical.*

Unlike the previous two approaches, the categorical framework is not committed to any particular entity, e.g. sentences or mathematical structures, with which to formalize scientific theories. Instead, this view brings with it a new dimension in the methodological debate in metascience. More precisely, if we opt to formalize the concept of theory with sentences, then, the claim goes, we should not formalize the theory as simply a *set* of sentences, but as a *category* of sentences instead. Similarly, instead of using a *class* of structures to represent theories, it has been claimed that we had better consider a *category* of structures. This being said, Halvorson and Tsementzis (2016, 11–3) have acknowledged the advantages a syntactic approach might have over a structural approach when the categorical viewpoint is taken, although this conviction seems to hinge on a particular view of the category of structures that is not shared by all authors working in the categorical framework.[2]

In this section, we will investigate the categorical framework for theories in more detail, focusing on categories of sentences and as well as categories of structures. Throughout this section, a basic knowledge of category theory is presupposed. An overview of the most important notions of this discipline may be found in appendix A. As before, we will start of our discussion by examining how individual theories are formalized in the framework under consideration. The categorical approach offers us two options in this regard: formalize a theory as a category of sentences (*syntactic category*) or formalize it as a category of structures (*semantic category*). Each of these possibilities will be considered in turn, being taken up in subsections 4.1.1 and 4.1.2 respectively. Much of the present section is based on (Halvorson & Tsementzis 2016). It should be noted at the outset, however, that there exists a closely related strand of literature within the categorical approach, exemplified by (Weatherall 2016), which differs from (Halvorson & Tsementzis 2016) in a number of significant ways. This is particularly pertinent for categories of structures. Accordingly, I shall address this divergence in the relevant subsection below.

4.1.1 Syntactic Categories

We start by considering what the categorical framework has to say about theories construed as syntactic entities. In the field of categorical logic, there exists a well-known procedure for constructing from a theory T,[3] a more elaborate structure, i.e. *the syntactic category* C_T of T. Before delving into this new notion, we must make an important caveat: the constructions that will follow below are applicable only to theories formulated in a particular *fragment* of first-order logic, known as the *coherent fragment.* Before proceeding, let us thus reflect on this system of logic.

[2] We shall return to this point in our discussion of *semantic categories* below.

[3] Recall that a *theory* in this sense refers to a deductively closed set of sentences.

Definition 4.1.1. The *coherent fragment* of first-order logic, or simply: *coherent logic*, is the fragment of first-order logic obtained by restricting its connectives to $\wedge, \vee, \top, \bot$ and its quantifiers to $\exists$. We denote this logic by $L^g_{\omega\omega}$.[4] A *coherent theory* is a theory in coherent logic.

Whence this restriction? Halvorson and Tsementzis note as a particular advantage that within the coherent fragment the 'distinction between intuitionistic and classical logic essentially disappears' (2016, 4). Furthermore, we do not loose any real expressive power when restricting ourselves to the coherent fragment of first-order logic, as may be seen through the process known as *Morleyization* (2016, 3), although we will not go into the details of this procedure here.

With this in mind, let us return to the notion of syntactic category. In the below definition, I follow Van Oosten (2002, 39-40), except in that I have modified the definition such that it avoids the notion of sequent. This modification requires us to allow for a larger class of sentences for $L^g_{\omega\omega}$ than we would expect from definition 4.1.1. More specifically, we hereby extend the set of sentences for $L^g_{\omega\omega}$ so as to include any sentence of the form

$$\forall\vec{x}[\varphi(\vec{x}) \to \psi(\vec{x})], \tag{4.1}$$

with $\vec{x} = (x_1, \ldots, x_n)$ a finite tuple, $\forall\vec{x}$ an abbreviation for $\forall x_1 \ldots \forall x_n$ and φ, ψ coherent formulas in the sense of definition 4.1.1. Throughout this subsection, all formulas will be assumed to be formulas in coherent logic.

With these provisions in place, let us now consider:

Definition 4.1.2. Let T be coherent theory. We then define the *syntactic category* C_T of T as follows:

- *Objects.* The objects of C_T are classes of formulas $\varphi(\vec{x})$ of $L^g_{\omega\omega}$ such that any two formulas $\varphi(\vec{x})$, $\psi(\vec{y})$ are in the same class if $\varphi(\vec{x})$ and $\psi(\vec{y})$ are the same formulas modulo renaming of variables.[5] We will denote an object of C_T as $[\varphi(\vec{x})]$, where $\varphi(\vec{x})$ is a formula belonging to its associated equivalence class. It is easily verified that the subsequent definitions do not depend on which formula from within an equivalence class we choose as its representative.

- *Morphisms.* The morphisms of C_T are defined to be equivalence classes of so-called *(T-provably) functional relations.* We thus need to understand both the functional relations themselves as well as the equivalence relation between them.

  - *Functional relations.* A *functional relation* from one formula $\varphi(\vec{x})$ to another formula $\psi(\vec{y})$ is a formula $\chi(\vec{x}, \vec{y})$ such that the following formulas are in T:

[4] This notation originates from the fact that coherent logic is occasionally also referred to as *geometric logic*.

[5] Note that the objects of C_T are independent from the theory T.

(i) $\forall\vec{x}\ \forall\vec{y}\ [\chi(\vec{x},\vec{y}) \to \varphi(\vec{x}) \wedge \psi(\vec{y})]$,

(ii) $\forall\vec{x}\ \forall\vec{y}\ \forall\vec{y'}[\chi(\vec{x},\vec{y}) \wedge \chi(\vec{x},\vec{y'}) \to \vec{y} = \vec{y'}]$,

(iii) $\forall\vec{x}[\varphi(\vec{x}) \to \exists\vec{y}\chi(\vec{x},\vec{y})]$.

That is, χ is a formula such that for any $\vec{x}$ for which $\varphi(\vec{x})$ holds there exists a unique $\vec{y}$ for which both $\chi(\vec{x},\vec{y})$ and $\psi(\vec{y})$ hold.[6]

– *Equivalence relation.* Let $[\varphi(\vec{x})]$, $[\psi(\vec{y})]$ be equivalence classes of formulas and let $\chi_1(\vec{x},\vec{y})$, $\chi_2(\vec{x},\vec{y})$ be two functional relations between the representatives $\varphi(\vec{x})$ and $\psi(\vec{y})$. We then define $\chi_1(\vec{x},\vec{y})$ and $\chi_2(\vec{x},\vec{y})$ to be *T-equivalent* if it holds that

$$\forall\vec{x}\ \forall\vec{y}\ [\chi_1(\vec{x},\vec{y}) \leftrightarrow \chi_2(\vec{x},\vec{y})] \in T. \tag{4.2}$$

A morphism between $\varphi(\vec{x})$ and $\psi(\vec{y})$ is then a T-equivalence class of functional relations. For any functional relation $\chi(\vec{x},\vec{y})$, we denote its T-equivalence class as $[\chi(\vec{x},\vec{y})]_T$.

- *Composition.* Given two morphisms

$$[\chi_1(\vec{x},\vec{y})]_T : [\varphi(\vec{x})] \to [\psi(\vec{y})],$$
$$[\chi_2(\vec{y},\vec{z})]_T : [\psi(\vec{y})] \to [\omega(\vec{z})],$$

we define the composition $[\chi_2(\vec{y},\vec{z})]_T \circ [\chi_1(\vec{x},\vec{y})]_T : [\varphi(\vec{x})] \to [\omega(\vec{z})]$ of the two morphisms to be given by $[\chi_{21}(\vec{x},\vec{z})]_T$, where

$$\chi_{21}(\vec{x},\vec{z}) = \exists\vec{y}[\chi_1(\vec{x},\vec{y}) \wedge \chi_2(\vec{y},\vec{z})]. \tag{4.3}$$

Let us verify that this construction is internally consistent. We have already noted how the definitions above do not depend on the choice of representative for the [.]-classes. But what of the $[.]_T$-classes?

Proposition 4.1.3. *The equivalence class $[\chi_{21}(\vec{x},\vec{z})]_T$ does not depend on the choice of representatives for $[\chi_1(\vec{x},\vec{y})]_T$ and $[\chi_2(\vec{y},\vec{z})]_T$.*

Proof. Let $\chi'_1(\vec{x},\vec{y}) \in [\chi_1(\vec{x},\vec{y})]_T$ and $\chi'_2(\vec{y},\vec{z}) \in [\chi_2(\vec{y},\vec{z})]_T$. We want to show that

$$\chi'_{21}(\vec{x},\vec{z}) \in [\chi_{21}(\vec{x},\vec{z})]_T. \tag{4.4}$$

By definition of T-equivalence classes, this is the case when

$$\forall\vec{x}\ \forall\vec{z}\ [\chi_{21}(\vec{x},\vec{z}) \leftrightarrow \chi'_{21}(\vec{x},\vec{z})] \in T. \tag{4.5}$$

To see that this is indeed the case, let us note that, by the definition of morphism composition for syntactic categories (4.3), we have:

$$\chi_{21}(\vec{x},\vec{z}) = \exists\vec{y}[\chi_1(\vec{x},\vec{y}) \wedge \chi_2(\vec{y},\vec{z})], \tag{4.6}$$
$$\chi'_{21}(\vec{x},\vec{z}) = \exists\vec{y}[\chi'_1(\vec{x},\vec{y}) \wedge \chi'_2(\vec{y},\vec{z})], \tag{4.7}$$

[6] Thus, we can view χ as a specifying a function from the set of all $\vec{x}$ such that $\varphi(\vec{x})$ holds to the set of all $\vec{y}$ such that $\psi(\vec{y})$ holds. Whence the terminology *functional relation.*

Moreover, by (4.2), the fact that $\chi_1'(\vec{x},\vec{y}) \in [\chi_1(\vec{x},\vec{y})]_T$ and $\chi_2'(\vec{x},\vec{y}) \in [\chi_2(\vec{x},\vec{y})]_T$ immediately yields:

$$\forall\vec{x}\ \forall\vec{y}\ [\chi_1(\vec{x},\vec{y}) \leftrightarrow \chi_1'(\vec{x},\vec{y})] \in T, \tag{4.8}$$

$$\forall\vec{x}\ \forall\vec{y}\ [\chi_2(\vec{y},\vec{z}) \leftrightarrow \chi_2'(\vec{y},\vec{z})] \in T. \tag{4.9}$$

Combining observations (4.6), (4.7) with (4.8) and (4.9), it readily follows that (4.5) holds. Hence, the equivalence class $[\chi_{21}(\vec{x},\vec{z})]_T$ is well-defined.

■

Next, we would do well to check that the syntactic category, as defined above, indeed satisfies the requirements for categories. Let us observe the following.

Proposition 4.1.4. *For any coherent theory T, the syntactic category C_T satisfies the axioms of category theory.*

Proof. First, we must check whether the composition operation is well-defined, i.e. whether the composition of two morphisms again results in a morphism. This is the case when for any two morphisms $[\chi_1(\vec{x},\vec{y})]_T, [\chi_2(\vec{y},\vec{z})]_T$ the composition $[\chi_{21}(\vec{x},\vec{z})]_T$ is also a morphism. This, in turn, is the case if

$$\chi_{21}(\vec{x},\vec{z}) = \exists\vec{y}[\chi_1(\vec{x},\vec{y}) \wedge \chi_2(\vec{y},\vec{z})] \tag{4.10}$$

is a functional relation from $[\varphi(\vec{x})]$ to $[\omega(\vec{z})]$, where $[\varphi(\vec{x})]$ is the source object of $[\chi_1(\vec{x},\vec{y})]_T$, and $[\omega(\vec{z})]$ is the target object of $[\chi_2(\vec{y},\vec{z})]_T$. For this to hold, $\chi_{21}(\vec{x},\vec{z})$ has to satisfy the conditions (i)-(iii) in definition 4.1.2. Using the fact that $[\chi_1(\vec{x},\vec{y})]_T$ and $[\chi_2(\vec{y},\vec{z})]_T$ are functional relations and hence satisfy conditions (i)-(iii) themselves, we can infer in a straightforward manner that $[\chi_{21}(\vec{x},\vec{z})]_T$ also satisfies each of the three requirements.

Next, we want to verify that morphism composition is associative. Thus, let

$$\begin{aligned}
[\chi_1(\vec{x},\vec{y})]_T &: [\varphi(\vec{x})] \to [\psi(\vec{y})],\\
[\chi_2(\vec{y},\vec{z})]_T &: [\psi(\vec{y})] \to [\omega(\vec{z})],\\
[\chi_3(\vec{z},\vec{w})]_T &: [\omega(\vec{z})] \to [\upsilon(\vec{w})],
\end{aligned}$$

be three morphisms and consider

$$[\chi_3(\vec{z},\vec{w})]_T \circ ([\chi_2(\vec{y},\vec{z})]_T \circ [\chi_1(\vec{x},\vec{y})]_T). \tag{4.11}$$

Twice applying the definition of morphism composition (4.3) to (4.11) shows the latter to be equal to

$$[\chi_3(\vec{z},\vec{w})]_T \circ [\exists\vec{y}[\chi_1(\vec{x},\vec{y}) \wedge \chi_2(\vec{y},\vec{z})]] = \tag{4.12}$$

$$[\exists\vec{z}[\chi_3(\vec{z},\vec{w}) \wedge \exists\vec{y}[\chi_1(\vec{x},\vec{y}) \wedge \chi_2(\vec{y},\vec{z})]]]_T := \tag{4.13}$$

$$[\chi_{321}(\vec{x},\vec{w})]_T. \tag{4.14}$$

In a similar vein, the morphism

$$([\chi_3(\vec{z},\vec{w})]_T \circ [\chi_2(\vec{y},\vec{z})]_T) \circ [\chi_1(\vec{x},\vec{y})]_T \tag{4.15}$$

is seen to be equal to

$$[\exists\vec{z}[\chi_2(\vec{y},\vec{z}) \wedge \chi_3(\vec{z},\vec{w})]] \circ [\chi_1(\vec{x},\vec{y})]_T = \tag{4.16}$$

$$[\exists\vec{y}[\exists\vec{z}[\chi_2(\vec{y},\vec{z}) \wedge \chi_3(\vec{z},\vec{w})] \wedge \chi_1(\vec{x},\vec{y})]]_T := \tag{4.17}$$

$$[\chi'_{321}(\vec{x},\vec{w})]_T. \tag{4.18}$$

Now, in order to conclude that (4.14) and (4.18) are equal, it suffices to show that $\chi_{321}(\vec{x},\vec{w}) \in [\chi'_{321}(\vec{x},\vec{w})]_T$, which is the case if

$$\forall\vec{x}\forall\vec{w}[\chi_{321}(\vec{x},\vec{w}) \leftrightarrow \chi'_{321}(\vec{x},\vec{w})] \in T. \tag{4.19}$$

That is is indeed the case, can be readily verified by eliminating and reintroducing existential quantifiers in an appropriate way in $\chi_{321}(\vec{x},\vec{w})$ and $\chi'_{321}(\vec{x},\vec{w})$. We can thus conclude that composition of morphisms satisfies associativity.

Lastly, we need to check the existence of the identity morphism. Let $[\varphi(\vec{x})]$ be an object. What is the identity morphism belonging to this object? First, note that, we can equivalently write the object in question as $[\varphi(\vec{x'})]$. The claim is that the identity morphism from $[\varphi(\vec{x})]$ to $[\varphi(\vec{x'})]$ is given by

$$[\varphi(\vec{x}) \wedge \vec{x} = \vec{x'}]_T. \tag{4.20}$$

We readily verify that this a functional relation. Furthermore, for any two objects $[\varphi(\vec{x})], [\psi(\vec{y})]$ and morphism $[\chi(\vec{x},\vec{y})]_T$, we have

$$[\chi(\vec{x},\vec{y})]_T \circ [\varphi(\vec{x}) \wedge \vec{x} = \vec{x'}]_T = \tag{4.21}$$

$$[\chi(\vec{x'},\vec{y})]_T \circ [\varphi(\vec{x}) \wedge \vec{x} = \vec{x'}]_T = \tag{4.22}$$

$$[\exists\vec{x'}[\chi(\vec{x'},\vec{y}) \wedge \varphi(\vec{x}) \wedge \vec{x} = \vec{x'}]]_T. \tag{4.23}$$

Note that in the step from (4.21) to (4.22) we can substitute the expression $[\chi(\vec{x'},\vec{y})]_T$ for $[\chi(\vec{x},\vec{y})]_T$, since the formulas $\chi(\vec{x},\vec{y})$ and $\chi(\vec{x'},\vec{y})$ are T-provably equivalent. Moreover, we find

$$[\psi(\vec{y}) \wedge \vec{y} = \vec{y'}]_T \circ [\chi(\vec{x},\vec{y})]_T = \tag{4.24}$$

$$[\exists\vec{y}[\psi(\vec{y}) \wedge \vec{y} = \vec{y'} \wedge \chi(\vec{x},\vec{y})]]_T. \tag{4.25}$$

We now need to check the equality of (4.23), (4.25) and $[\chi(\vec{x},\vec{y})]_T$. To this end, let us verify the T-provable equivalence of $\chi(\vec{x},\vec{y})$ and

$$\exists\vec{x'}[\chi(\vec{x'},\vec{y}) \wedge \varphi(\vec{x}) \wedge \vec{x} = \vec{x'}]. \tag{4.26}$$

Clearly, we can deduce $\chi(\vec{x},\vec{y})$ from (4.26). Conversely, we can readily deduce the formula

$$\exists\vec{x'}[\chi(\vec{x'},\vec{y}) \wedge \vec{x} = \vec{x'}] \tag{4.27}$$

from $\chi(\vec{x},\vec{y})$. Now, we simply have to note that criterion (i) in the definition of functional relations guarantees that $\chi(\vec{x},\vec{y})$ implies $\varphi(\vec{x})$ and we are done. The argument for the equivalence of $\chi(\vec{x},\vec{y})$ and (4.25) proceeds in a completely analogous manner. Hence, we conclude that the syntactic category C_T indeed satisfies the axiom for identity morphisms. ■

Of what use are these syntactic categories to the metascientific enterprise? Halvorson and Tsementzis (2016, 4) note how syntactic categories help us to 'partially eliminate' the much bemoaned yoke of syntax: if T and T' are two theories formulated in different signatures Σ and Σ', then they might still have *equivalent* syntactic categories. There is, however, no real novelty to be had in addressing this particular problem, as it is has been tackled well before the genesis of the categorical view.[7] A second advantage of working with syntactic categories cited by Halvorson and Tsementzis (*ibid.*) is that the resulting categories possess the structure of what is called a *coherent category*, which is noted to have "just the right amount of structure to express models of coherent theories." Of course, the desirability of this feature is hinged entirely on our commitment to coherent logic as valuable instrument for formalizing scientific theories: a presupposition that is far from trivial. Indeed, we find the primary justification, as given by Halvorson and Tsementzis (2016, 14), for the potential fruitfulness of the categorical approach to metascience is based on the observed effectiveness of its methods in the discourse of metamathematics. The validity of such an analogy, however, is entirely contingent on the exact manner in which we choose to explicate our metascientific framework: mutual transferability of successful methods between the discourses of metascience and metamathematics is by no means guaranteed.

As of yet, the question of how to apply syntactic categories to the analysis of concrete scientific theories in a non-trivial fashion remains without answer. Until such applications come about, it will remain difficult to gauge the potential syntactic categories hold for the metascientific enterprise. We shall return to this issue in chapter 5, where the categorical approach will be evaluated in its entirety while contemplating its compatibility with the ideology of logical abstractivism. For the moment, however, let us turn our attention towards different facets of the categorical school.

4.1.2 Semantic Categories

As has already been noted above, the categorical view per se is not committed to a specific kind of entity, e.g. statements or structures, with which to formalize scientific theories. Instead, it only seeks to advise us on *how* a given type of entity can aid the formalization of theories, viz. by not relying on the traditional methods of metamathematics, but rather by turning to the new framework of category theory. Accordingly, the syntactic view of theories is not the only view to receive a categorical transformation: just as we can associate a category to a class of sentences in a formal language, so too can we construct a category for a given class of *structures.*

Looking at the literature, there appears to exist a dichotomy in the manner structures are incorporated into the categorical framework. On the one hand, we have (Halvorson & Tsementzis 2016), which considers a class of structures to be a class of structures *for some coherent theory.* In this case, the category

[7] cf. (Muller 2011, 15–23).

of structures reduces to a *category of models*, also referred to as a *semantic category*. On the other hand, (Barrett, Rosenstock & Weatherall 2015) and (Weatherall 2016) do not impose any such restriction on their structures. Instead, a class of structures is defined directly and the corresponding category simply consists of the structures of this class along with a suitable notion of morphism. For example, Barrett, Rosenstock and Weatherall (2016, 311) in their analysis of general relativity represent this scientific theory directly by a category having *relativistic spacetimes*[8] as objects and isometries between these spacetimes as morphisms.[9]

Which of these, if any, is the 'right' approach to structures? For essentially the same reasons as expressed in section 2.1, viewing structures as being models of some class of sentences in a formal language imposes unnecessary restrictions on the structural view. Moreover, the papers in the second strand of literature have as an added bonus the fact that they analyze actual scientific theories, such as general relativity and Newtonian gravity. By contrast, we have yet to see any applications to scientific theories of the notion of structure employed in the first strand.[10] On the flip side, however, we might note that the approach taken by Halvorson and Tsementzis (2016) is of a far more general nature, attempting to formulate methods for the analysis of arbitrary theories (even if these are toy theories in coherent logic). Such a general vantage point, by contrast, is not taken up by Weatherall and his collaborators, who instead seem to operate on a case-by-case basis.

In summary, we can distinguish two varieties of the categorical approach: one in the style of James Weatherall, focusing on concrete scientific theories and working with categories of structures not yoked to a particular formal language, the other in the style of Hans Halvorson, focusing on the general concept of a scientific theory and working with syntactic categories or, alternatively, categories of models of coherent theories. Based on the geographical distribution of the principal exponents of each approach, California for the former and New Jersey for the latter, we might refer to these two different approaches within the categorical school as the *west-coast style* on the one hand and the *east-coast style* on the other. Out of practical considerations, I shall limit myself in this thesis to the east-coast style framework of (Halvorson & Tsementzis 2016) and, hence, will limit the present discussion on categories of structures to semantic categories. However, if one were to formulate a proper translation of the structural view in category-theoretic terms, the above dichotomy would require serious attention.

[8] A *relativistic spacetime* is defined to be a Lorenztian manifold (M, g), consisting of a smooth four-dimensional manifold M and a Lorentzian metric g (Barett et al. 2015, 310).

[9] Of course, this category may still be construed of as a category of models, in the sense that relativistic spacetimes may be viewed as 'models' of general relativity. This construal of the term *model*, however, should be carefully distinguished from the *models* of model theory, where they serve to interpret not *scientific theories* but theories formulated within a particular system of logic. It is this latter type of model, as we will see, that makes up the *category of models* considered by Halvorson and Tsementzis (2016).

[10] This is, of course, closely connected to the lack of any significant applications for the concept of syntactic category in the philosophy of science, as discussed above.

Let us now consider semantic categories in more detail. Halvorson and Tsementzis (2016, 9–11) present us with two possible definitions. First, consider:

Definition 4.1.5. Let T be a coherent theory over a signature Σ. The *semantic category* $\mathrm{Mod}_1(T)$ of T is the category having as objects models of T and homomorphisms between Σ-models as morphisms.

We may note that, in general, there exist many homomorphisms between models of a given theory. This observation leads us to the following, alternative definition of semantic category.

Definition 4.1.6. Let T be a coherent theory over a signature Σ. The *semantic category* $\mathrm{Mod}_2(T)$ of T is the category having as objects models of T and elementary embeddings between Σ-models as morphisms.

Clearly, every elementary embedding between Σ-structures is also a homomorphism,[11] while the converse is certainly not the case. Thus, the category $\mathrm{Mod}_2(T)$ will be a subcategory of $\mathrm{Mod}_1(T)$, having fewer morphisms between each pair of objects.

Which of these two definitions provides us with the 'best' notion of semantic category? Ironically enough, Halvorson and Tsementzis (2016, 12) note that both notions are, in fact, woefully inadequate.[12] More specifically, they argue that $\mathrm{Mod}_1(T)$ and, a fortiori, $\mathrm{Mod}_2(T)$, do not contain certain 'topological' information about the collections of models that is implicit in the theory T. To see this, note that we can define a topology on the collection of all models of T by letting any sequence $(M_i)_{i\in\mathbb{N}}$ of models of T converge to another model M of T if and only if the truth value of φ in M_i converges to the truth value of φ in M for every sentence φ in T. Let us denote the resulting topological space by $\underline{\mathrm{Mod}}(T)$. We can now, for the idealized case of theories in propositional logic, invoke a well-known result known as the *Stone duality theorem* to establish:

Theorem 4.1.7. *The collection of compact open subsets of* $\underline{\mathrm{Mod}}(T)$ *forms a Boolean lattice that is equivalent, in the categorical sense, to* $C(T)$.

Without entering into the details of this result, we can already see that in a definite sense the syntactic category $C(T)$ contains more information than either $\mathrm{Mod}_1(T)$ or $\mathrm{Mod}_2(T)$ as far as propositional theories T are concerned. Whether we can establish similar results for stronger logics, such as coherent or first-order logic, is still an area of active investigation, with much progress being made at the hands of Awodey and Forssell (2013).

Halvorson and Tsementzis (2016, 13) infer from the above observation a more general metascientific morale, which states that the content of a certain *scientific* theory T is not exhausted by either its category of models or, a fortiori, its class of models. It is in this sense that both authors seem to favor a syntactic approach to scientific theory-structure over a structural one. Such a valuation of the

[11] Cf. section 1.3 for the definitions.

[12] This lack of usefulness might also explain why a standard definition of semantic category does not seem to exist in the first place.

structural view would, however, rely entirely on our willingness to accept results obtained for classical systems of logic as indicative for logical metascience, as acknowledged by Halvorson and Tsementzis (2016, 13–4). Accordingly, any such recommendations must be approached with due levels of reservation. In particular, we might note that theorem 4.1.7 loses its metascientific import if we forego the identification of categories of structures with categories of models, as is done in the west coast-style of categorical metascience. We shall return to this observation in the next chapter. For the moment, let us note that *given* a methodological commitment to coherent/first-order logic as the basis of our metascientific frameworks, we can indeed view semantic categories as less versatile than their syntactic counterparts.

4.2 Intertheory Relations

Having discussed the categorical view of theories, we can now turn our attention to the explication of intertheory relations. The motivation for such an exercise has already been discussed at the start of subsection 2.4.2 and will not be reiterated here. As noted above, the categorical view is, in principle, neutral as to whether we need to represent theories by means of statements or structures. This being said, when it comes to explicating intertheory relations in full generality, exponents of the categorical view seem to favor the syntactic viewpoint, i.e. the usage of syntactic categories.

As in the preceding frameworks, the relation of equivalence has received the most attention in the literature and will accordingly be at the center of our attention in this section. It is here that we encounter a notion that is ubiquitous in the literature on the categorical approach, viz. the notion of *Morita equivalence*. The claim, as expressed by Halvorson and Tsementzis (2016, 9), is that this is exactly the right type of equivalence with which to capture the equivalence between theories.

Let us thus consider what this notion of Morita equivalence is supposed to be. In the literature, we can identify two different types of Morita equivalence, referred to as *J-Morita equivalence* and *T-Morita equivalence* by Tsementzis (2015). In brief, the difference between the two notions may be characterized as follows. On the one hand, T-Morita equivalence is notion of a syntactic nature, in the sense that it is defined as a relation between two different first-order theories T_1 and T_2. By contrast, J-Morita equivalence is fundamentally a category-theoretic notion and is accordingly defined in terms of syntactic categories $C(T_1)$ and $C(T_2)$. In the case of coherent logic, both notions of Morita equivalence turn out to be equivalent, as demonstrated by Tsementzis (2015). Through which avenue we choose to familiarize ourselves with the concept of Morita equivalence is thus largely of matter of taste. For the purposes of this section, our main focus will be with the notion of T-Morita equivalence, as it is by far the more intuitive of the two.

4.2.1 Morita Equivalence

The concept of T-Morita equivalence finds its origin in (Barrett & Halvorson 2015), in which the authors set out to find appropriate criteria for the equivalence of theories in many-sorted first-order logic. Now, T-Morita equivalence certainly is not the fist notion of equivalence that has been formulated for first-order theories. So what do we stand to gain from introducing a new type of equivalence? Consider, for instance, the following definition.

Definition 4.2.1. Let T_1, T_2 be two theories.[13] Then T_1 and T_2 are said to be *logically equivalent* if they have the same class of models.

Is logical equivalence a satisfactory criterion for equivalence between theories? A brief moment of reflection reveals a negative answer to this question. From any theory T, we can obtain a new theory T' by simply renaming all the non-logical symbols occurring in the original theory. By any measure, we would like to consider T and T' equivalent theories. Yet, they fail the criterion for logical equivalence, since T and T' have different underlying signatures, the theories will also have two different classes of models. Thus, we can easily observe that logical equivalence is far too strict to adequately capture equivalence between theories.

Clearly, we are in need of a more nuanced criterion for theoretical equivalence. To this end, Barrett and Halvorson (2015, 1) consider three possible candidates:

(i) definitional equivalence,

(ii) T-Morita equivalence,

(iii) categorical equivalence.

The argument for T-Morita equivalence, which we will take up in more detail below, now runs as follows. Definitional equivalence is too strict: just as logical equivalence, it distinguishes theories that ought to be equivalent. Categorical equivalence, on the other hand, is too liberal: it equates theories we would like to consider different. T-Morita equivalence, however, is just right: it sits in between definitional and categorical equivalence and it equates just those and only those theories that we would intuitively like to think of as equivalent. It is this process of justification, along with the different notions of equivalence (i)–(iii), that shall be at the center of our attention for the remainder of this subsection.

At the very least, a criterion for equivalence should allow for theories formulated in different signatures to be equivalent. This is exactly what motivates the concept of definitional equivalence. Before we can consider this type of equivalence, we require several ancillary notions.

[13] From now until the end of this section, we will assume all signatures, theories and models to be defined with respect to many-sorted first-order logic.

Definition 4.2.2. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $P \in \Sigma' \setminus \Sigma$ be a relation symbol of arity $\sigma_1 \times \ldots \times \sigma_n$. Then an *explicit definition of P in terms of* Σ is a Σ'-sentence

$$\forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n [P(x_1, \ldots, x_n) \leftrightarrow \varphi(x_1, \ldots, x_n)], \tag{4.28}$$

where φ is a Σ-formula.

Definition 4.2.3. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $f \in \Sigma' \setminus \Sigma$ be a function symbol of arity $\sigma_1 \times \ldots \times \sigma_n \to \sigma$. Then an *explicit definition of f in terms of* Σ is a Σ'-sentence

$$\forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n \forall_\sigma y [f(x_1, \ldots, x_n) = y \leftrightarrow \psi(x_1, \ldots, x_n, y)], \tag{4.29}$$

where ψ is a Σ-formula. To this explicit definition we associate the sentence

$$\forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n \exists!_\sigma y [\psi(x_1, \ldots, x_n, y)], \tag{4.30}$$

called the *admissibility condition* for (4.29), which expresses a necessary condition in the signature Σ for ψ defining a function symbol.

Definition 4.2.4. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $c \in \Sigma' \setminus \Sigma$ be a constant symbol of arity σ. Then an *explicit definition of c in terms of* Σ is a Σ'-sentence

$$\forall_\sigma x [x = c \leftrightarrow \rho(x)], \tag{4.31}$$

where ρ is a Σ-formula. To this explicit definition we associate the sentence

$$\exists!_\sigma x [\rho(x)], \tag{4.32}$$

called the *admissibility condition* for (4.31), which expresses a necessary condition in the signature Σ for ρ defining a constant symbol.

Definition 4.2.5. Let T be a theory over the signature Σ and let Σ' be a signature extending Σ such that Σ and Σ' have the same sort symbols. A *definitional extension* of T to Σ' is a theory logically equivalent to

$$T \cup \{\delta_s \mid s \in \Sigma' \setminus \Sigma\}, \tag{4.33}$$

where δ_s is an explicit definition for the symbol s. In case s is a function or constant symbol, we also require that $T \models \alpha_s$, where α_s is the admissibility condition for δ_s.

The final clause in definition 4.2.5 is required to eliminate possibility of nonsensical 'definitional extensions'. For example, consider:

Example 4.2.6. Let Σ be the signature consisting of only a binary relation symbol R and let T be the deductive closure of the sentence $\forall x[xRx]$. That is, T is the theory of reflexive relations. Next, let $\Sigma' = \Sigma \cup \{\Box\}$, where $\Box$ is a constant symbol and consider the explicit definition $\forall x[x = \Box \leftrightarrow \forall y[xRy]]$. Clearly, it does not hold that $T \models \alpha_\Box$. Ignoring the final clause of definition 4.2.5, we could now call $T \cup \{\delta_\Box\}$ a 'definitional extension' of T to Σ'. Such a claim would vacuous, however, since we can readily identify models of T for which the explicit definition $\delta_\Box$ fails to determine a unique expansion to Σ'.

We are now ready to formulate the concept of definitional equivalence.

Definition 4.2.7. Let T_1 and T_2 be theories over the signatures Σ_1 and Σ_2 respectively. T_1 and T_2 are *definitionally equivalent* if there exist theories T_1', T_2' such that

(i) T_1' is a definitional extension of T_1 to $\Sigma_1 \cup \Sigma_2$,

(ii) T_2' is a definitional extension of T_2 to $\Sigma_1 \cup \Sigma_2$,

(iii) T_1' and T_2' are logically equivalent $\Sigma_1 \cup \Sigma_2$-theories.

To paraphrase, two theories are definitionally equivalent if they have a 'common definitional extension'.

We can immediately verify that definitional equivalence is indeed much more liberal than logical equivalence. In what follows, we shall write Γ^{cl} to denote the deductive closure for any set of sentences Γ.

Proposition 4.2.8. *If two theories are logically equivalent, then they are definitionally equivalent. The converse, however, does not generally hold true.*

Proof. For the first statement, observe that if two theories T_1 and T_2 are logically equivalent, then they have identical signatures $\Sigma_1 = \Sigma_2 = \Sigma$. Consequently, we can simply set $T_1' := T_1$, $T_2' := T_2$ and the conditions for definitional equivalence are satisfied.

To see that the converse implication does not hold, we can construct an easy counterexample. Take, for instance, the theory of *non-strict* total orders, formulated in the signature $\{\sqsubseteq\}$ consisting of a single binary relation symbol:

$$\begin{aligned} T_1 = \{&\forall x \forall y[x \sqsubseteq y \wedge y \sqsubseteq x \to x = y], \\ &\forall x \forall y \forall z[x \sqsubseteq y \wedge y \sqsubseteq z \to x \sqsubseteq z], \\ &\forall x \forall y[x \sqsubseteq y \vee y \sqsubseteq x]\}^{\text{cl}}. \end{aligned}$$

Now, compare this to the theory of *strict* total orders, formulated in the signature $\{\sqsubset\}$ similarly consisting of a single binary relation symbol:

$$\begin{aligned} T_2 = \{&\forall x \forall y[x \sqsubset y \to \neg(y \sqsubset x)], \\ &\forall x \forall y \forall z[x \sqsubset y \wedge y \sqsubset z \to x \sqsubset z], \\ &\forall x \forall y[x \neq y \to x \sqsubset y \vee y \sqsubset x]\}^{\text{cl}}. \end{aligned}$$

Clearly, T_1 and T_2 fail to be logically equivalent since they have non-identical classes of models: non-strict total orders for the former and strict total order for the latter. We can, however, define a common definitional extension. Let

$$T_1' = T_1 \cup \{\forall x \forall y[x \sqsubset y \leftrightarrow x \sqsubseteq y \wedge x \neq y]\}^{\text{cl}}$$

and

$$T_2' = T_2 \cup \{\forall x \forall y[x \sqsubseteq y \leftrightarrow x \sqsubset y \vee x = y]\}^{\text{cl}}.$$

Then, clearly, T_1' and T_2' are logically equivalent theories over the extended signature $\{\sqsubseteq, \sqsubset\}$.

We conclude that T_1 and T_2, while not logically equivalent, do meet the requirements for being definitionally equivalent.

■

The notion of definitional equivalence is well established in the logic literature and has received considerable attention from logicians. We might thus hope that with definitional equivalence, we have come across the appropriate notion with which to capture theoretical equivalence. Yet, as noted Barrett and Halvorson (2015, 8–9), we still find several instances in which definitional equivalence proves to be too harsh of a criterion. For instance, consider the theories $T_1 := \mathrm{Th}(\mathbb{Z}, \leq)$ and $T_2 := \mathrm{Th}(Z^+, Z^-, \leq^+, \leq^-)$, with Z^+, Z^- the non-negative and negative integers and $\leq^+, \leq^-$ the corresponding orders, formulated over the signatures $\{\sigma_0, \sqsubseteq\}$ and $\{\sigma_1, \sigma_2, \sqsubseteq^+, \sqsubseteq^-\}$ respectively. There is clearly ample reason to consider both theories equivalent. However, T_1 and T_2 fail to meet the criteria of definitional equivalence. To see this, note that while definitional extensions allow us to define new relation, function and constant symbols in terms of other such symbols, they do *not* allow us to define new *sort symbols* in an analogous manner. Hence, any two theories formulated in signatures with different sort symbols will automatically be disqualified from being definitionally equivalent, regardless of any conceptual similarity the two might share.

To remedy this situation, we need to generalize the notion of definitional equivalence in such a manner that we can define now sort symbols from old ones. This will lead us to the much-anticipated notion of T-Morita equivalence. It turns out that, for all intents and purposes, it will suffice to define four particular ways in which we can construct new sort symbols. Let us consider each of these four in turn.

Definition 4.2.9. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $\sigma \in \Sigma' \setminus \Sigma$. An *explicit definition* of σ as a *product sort* in terms of Σ is a Σ'-sentence

$$\forall_{\sigma_1} x \forall_{\sigma_2} y \exists !_{\sigma} z[\pi_1(z) = x \wedge \pi_2(z) = y], \tag{4.34}$$

where $\sigma_1, \sigma_2 \in \Sigma$ are sort symbols and $\pi_1, \pi_2 \in \Sigma' \setminus \Sigma$ are function symbols with arities $\sigma \to \sigma_1$ and $\sigma \to \sigma_2$ respectively.

Some words of clarification might be in order. As its name suggests, the product sort allows us to syntactically codify what it means for one domain in a model to be given by the product, in the set-theoretic sense, of two other domains. The function symbols π_1, π_2 can then be seen as syntactic representation of the projection functions from the new domain to the original ones.

Moving on, we have

Definition 4.2.10. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $\sigma \in \Sigma' \setminus \Sigma$. An *explicit definition* of σ as a *coproduct sort* in terms of Σ is a Σ'-sentence

$$\forall_{\sigma} z[\exists !_{\sigma_1} x[\rho_1(x) = z] \vee \exists !_{\sigma_2} y[\rho_2(y) = z]] \wedge \forall_{\sigma_1} x \forall_{\sigma_2} y[\rho_1(x) \neq \rho_2(y)], \tag{4.35}$$

where $\sigma_1, \sigma_2 \in \Sigma$ are sort symbols and $\rho_1, \rho_2 \in \Sigma' \setminus \Sigma$ are function symbols with arities $\sigma_1 \to \sigma$ and $\sigma_2 \to \sigma$ respectively.

The motivation here is virtually identical to that for the product case. We now define what it means for a domain to be the coproduct, in the set-theoretic sense, of two other domains. The function symbols i_1, i_2 represent the inclusion maps from the original domains to the new one.

We see that the idea underlying these definitions is to codify certain set-theoretic constructions in the language of many-sorted first-order logic, thereby increasing our language's expressive power. Thus, in a similar vein, we find

Definition 4.2.11. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $\sigma \in \Sigma' \setminus \Sigma$. An *explicit definition* of σ as a *subsort* in terms of Σ is a Σ'-sentence

$$\forall_{\sigma_1} x[\varphi(x) \leftrightarrow \exists_\sigma z[i(z) = x]] \wedge \forall_\sigma z_1 \forall_\sigma z_2[i(z_1) = i(z_2) \to z_1 = z_2], \tag{4.36}$$

where φ is a Σ-formula, $\sigma_1 \in \Sigma$ is a sort symbol and $i \in \Sigma' \setminus \Sigma$ is a function symbol with arity $\sigma \to \sigma_1$. As before, we associate to this explicit definition an admissibility condition, viz. the sentence $\exists_{\sigma_1} x[\varphi(x)]$.

This time, it is the subset relation we are seeking to formalize. To this end, we employ a function symbol i which is interpreted as the injection sending each element z to the same element $i(z)$ in the extended domain.

Our final definition now reads:

Definition 4.2.12. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. Let $\sigma \in \Sigma' \setminus \Sigma$. An *explicit definition* of σ as a *quotient sort* in terms of Σ is a Σ'-sentence

$$\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2[\varepsilon(x_1) = \varepsilon(x_2) \to \varphi(x_1, x_2)] \wedge \forall_\sigma z \exists_{\sigma_1} x[\varepsilon(x) = z], \tag{4.37}$$

where φ is a Σ-formula, $\sigma_1 \in \Sigma$ is a sort symbol and $\varepsilon \in \Sigma' \setminus \Sigma$ is a function symbol with arity $\sigma_1 \to \sigma$. To this explicit definition we associate the following three admissibility conditions:

(i) $\forall_{\sigma_1} x[\varphi(x, x)]$,

(ii) $\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2[\varphi(x_1, x_2) \to \varphi(x_2, x_1)]$,

(iii) $\forall_{\sigma_1} x_1 \forall_{\sigma_1} x_2 \forall_{\sigma_1} x_3[\varphi(x_1, x_2) \wedge \varphi(x_2, x_3) \to \varphi(x_1, x_3)]$.

In this case, we are defining a new domain to consist of equivalence classes of elements of our original domain. The equivalence relation in question is encoded by the formula $\varphi(x_1, x_2)$ and the function mapping each element of the original domain to its equivalence class in the new domain is represented by ε.

Having defined four ways in which we might construct new sort symbols from a given signature, we are now almost ready to consider the definition of T-Morita equivalence.

Definition 4.2.13. Let Σ, Σ' be signatures such that $\Sigma \subseteq \Sigma'$ and let T be a theory, formulated in the signature Σ. A *Morita extension* of T to the signature Σ' is a Σ'-theory that is logically equivalent to

$$T \cup \{\delta_s \mid s \in \Sigma' \setminus \Sigma\}, \tag{4.38}$$

where δ_s is an explicit definition for each symbol s, such that (i) for any sort symbol σ and any function symbol f occurring in the explicit definition of σ we have $\delta_f = \delta_\sigma$, and (ii) for any admissibility condition α_s associated to an explicit definition δ_s, we have $T \models \alpha_s$.

Similarly to the definition of definitional equivalence, we now define T-Morita equivalence as there existing a 'common Morita extension' for two theories.

Definition 4.2.14. Let T_1 and T_2 be theories over signatures Σ_1 and Σ_2 respectively. Then T_1 and T_2 are *T-Morita equivalent* if there exist sequences of theories $T_1^1, \ldots, T_1^n$ and $T_2^1, \ldots, T_2^m$ such that

(i) Each theory T_1^{i+1} is a Morita extension of T_1^i,

(ii) Each theory T_2^{j+1} is a Morita extension of T_2^j,

(iii) The theories T_1^n and T_2^m are logically equivalent theories in some signature Σ containing both Σ_1 and Σ_2.

Comparing the definitions of definitional and T-Morita equivalence, we might be struck by a discrepancy. Why is it that we require sequences of theories for T-Morita equivalence, while not doing so for definitional equivalence? To answer this question, we must look at the following theorem concerning definitional equivalence.

Theorem 4.2.15. *Let Σ be a signature, T be a Σ-theory and Σ' be a signature extending Σ. If T' is a definitional extension of T to Σ', then for every Σ'-formula $\varphi(x_1, \ldots, x_n)$ there exists a Σ-formula $\varphi^*(x_1, \ldots, x_n)$ such that*

$$T' \models \forall_{\sigma_1} x_1 \ldots \forall_{\sigma_n} x_n[\varphi(x_1, \ldots, x_n) \leftrightarrow \varphi^*(x_1, \ldots, x_n)]. \tag{4.39}$$

From this theorem, we can infer the following helpful corollary.

Corollary 4.2.16. *If T' is a definitional extension of T from Σ to Σ', and T'' is a definitional extension of T' from Σ' to Σ'', then T'' is also a definitional extension of T from Σ to Σ''.*

Thus, by corollary 4.2.16, including sequences of definitional extensions in definition 4.2.7 would be superfluous. The same, however, does not hold for Morita extensions, for which we have the following weaker analogue of theorem 4.2.15:

Theorem 4.2.17. *Let Σ be a signature, T be a Σ-theory and Σ' be a signature extending Σ. If T' is a Morita extension of T to Σ', then for every Σ'-sentence φ there exists a Σ-formula φ^* such that $T' \models \varphi \leftrightarrow \varphi^*$.*

Consequently, we have to explicitly allow for sequences of theories in definition of T-Morita equivalence.

Again, we would like to verify that the notion of equivalence we have expounded is more liberal than the one preceding it.

Theorem 4.2.18. *If two theories are definitionally equivalent, then they are T-Morita equivalent. The converse, however, does not generally hold true.*

Proof. Clearly, any definitional extension of a theory is also a Morita extension. Hence, if we have two definitionally equivalent theories T_1, T_2, we can simply select the definitional extensions T_1', T_2' to serve as the required Morita extensions and we are done.

To see that the converse implication does not hold, consider the signatures $\Sigma_1 = \{\sigma_1, P, Q\}$ and $\Sigma_2 = \{\sigma_2, \sigma_3\}$ and consider the Σ_1-theory

$$\begin{aligned} T_1 = \{&\exists_{\sigma_1} x[P(x)], \exists_{\sigma_1} x[Q(x)], \\ &\forall_{\sigma_1} x[P(x) \vee Q(x)], \\ &\forall_{\sigma_1} x[\neg(P(x) \wedge Q(x))]\}^{\mathrm{cl}}, \end{aligned}$$

and the Σ_2-theory $T_2 = \varnothing^{\mathrm{cl}}$. Both theories express the fact that we can partition our domain of discourse into two, disjoint regions, T_1 doing so by means of two relation symbols and T_2 doing so by means of two sort symbols. Thus, intuitively, we might reasonably deem T_1 and T_2 to be equivalent theories.

Since these theories have different sort symbols, it immediately follows that they are not definitionally equivalent. They do, however, meet the criteria for T-Morita equivalence. To see this, consider the signature $\Sigma = \Sigma_1 \cup \Sigma_2 \cup \{i_2, i_3\}$, with i_2 and i_3 function symbols of arity $\sigma_2 \to \sigma_1$ and $\sigma_3 \to \sigma_1$ respectively. We are going to construct three Morita extensions: one for T_1, one for T_2 and one for the Morita extension of T_2. This will then suffice to establish T-Morita equivalence. Let us consider each extension in turn.

- Let the Σ-sentences δ_{σ_2}, δ_{σ_3} be the explicit definitions of σ_2 and σ_3 as the subsort symbols of σ_1 with respect to the function symbols i_2 and i_3 respectively. We then consider the Morita extension $T_1^1 = T_1 \cup \{\delta_{\sigma_2}, \delta_{\sigma_3}\}$ of T_1 to Σ.

- Let δ_{σ_1} be the explicit definition of σ_1 as the coproduct sort of σ_2 and σ_3 with respect to the function symbols i_2 and i_3. Let us now define the Morita extension $T_2^1 = T_2 \cup \{\delta_{\sigma_1}\}$ of T_2 to $\Sigma_2 \cup \{\sigma_1, i_2, i_3\}$.

- Let δ_P and δ_Q be explicit definitions for P and Q given by

  $$\forall_{\sigma_1} x[P(x) \leftrightarrow \exists_{\sigma_2} y[i_2(y) = x]], \quad \forall_{\sigma_1} x[Q(x) \leftrightarrow \exists_{\sigma_3} y[i_3(y) = x]]. \quad (4.40)$$

  and consider the Morita extension $T_2^2 = T_2^1 \cup \{\delta_P, \delta_Q\}$ of T_2^1 to Σ.

It is now straightforwardly verified that T_1^1 and T_2^2 are logically equivalent Σ-theories. We can conclude that T_1 and T_2 are indeed T-Morita equivalent.

■

Let us take a moment to reflect on our progress so far. We have climbed up from logical equivalence, to definitional equivalence, to T-Morita equivalence. In every step, it was argued that the preceding notion of equivalence provided us with too strict a criterion for theoretical equivalence, by noting how it forced us to differentiate between theories we would intuitively like to consider equivalent. Now, however, we will need to proceed in the opposite direction: we need to show (i) that we did not 'overshoot' when going from definitional equivalence to T-Morita equivalence, i.e. that T-Morita equivalence still meets the basic desiderata that we expect for theoretical equivalence; and (ii) that we cannot 'go farther' than T-Morita equivalence.

To establish point (i), Barrett and Halvorson note (2015, 11–5) how T-Morita equivalence preserves, to a certain extent, a number of nice properties of definitional equivalence. More specifically, consider the following theorem for definitional extensions.

Theorem 4.2.19. *Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. If T' is a definitional extension of T to Σ' then every model M of T has, up to isomorphism, a unique expansion M' that is a model of T'.*

In the terms of Barrett and Halvorson (2015, 7), this result shows us how a definitional extension 'says no more' than the original theory from a semantic point of view. Naturally, we would then like for this result to carry over to the case of Morita extensions. Fortunately, we can indeed prove a result analogous to theorem 4.2.19.

Theorem 4.2.20. *Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$. If T' is a Morita extension of T to Σ' then every model M of T has, up to isomorphism, a unique expansion M' that is a model of T'.*

The similarity between theorems 4.2.19 and 4.2.20 is already a good indication that T-Morita equivalence is not too liberal a criterion for theoretical equivalence. But we can do more. Consider the following:

Definition 4.2.21. Let Σ, Σ' be signatures such that $\Sigma \subset \Sigma'$ and let T and T' be a Σ-theory and Σ'-theory respectively. We say that T' is a *conservative extension* of T if for every Σ-sentence φ it holds that

$$T \models \varphi \text{ if and only if } T' \models \varphi. \tag{4.41}$$

Theorem 4.2.22. *If T' is a definitional extension of T, then T' is a conservative extension of T.*

This provides us with another sense in which a definitional extension T' 'says no more' than the original theory T, i.e. T' says no more than T with respect to the original signature Σ. Once again, we might wonder whether this 'nice' property of definitional extensions transfers cleanly to Morita extensions. Barrett and Halvorson (2015, 14) answer in the affirmative:

Theorem 4.2.23. *If T' is a Morita extension of T, then T' is a conservative extension of T.*

Proof. For sake of contradiction, suppose T' is not a conservative extension of T. Since T is contained in T', it is clear that $T \models \varphi$ implies $T' \models \varphi$ for every Σ-sentence φ. Hence, it is the converse implication that must fail for some Σ-sentence φ, i.e. we have $T' \models \varphi$ but $T \not\models \varphi$. This, in turn, means there exists some model M of T such that $M \models \neg\varphi$. By theorem 4.2.20, we know that this model M must have a unique expansion M' that is a model of T'. Then, since $T' \models \varphi$, it holds that $M' \models \varphi$. However, since M' is an expansion of M, we also have that $M' \models \neg\varphi$. We have arrived at the desired contradiction. ■

We see that T-Morita equivalence preserves several nice features of definitional equivalence. Thus, it seems reasonable to conclude that we have not generalized excessively in moving from definitional to T-Morita equivalence.

Finally, we consider whether T-Morita equivalence truly is general enough. To this end, we examine one final notion of equivalence, viz. *categorical equivalence*, that is even more general than T-Morita equivalence and see why we may deem it *too general.* As its name suggests, categorical equivalence is a notion rooted in category theory. In what follows, we will rely heavily on material covered in appendix A. The reader without a reasonable background in category theory is advised to consult this appendix before proceeding with the material below. In particular, we require:

Proposition 4.2.24. *A functor is an equivalence of categories if and only if it is full, faithful and essentially surjective.*

Definition 4.2.25. Two categories C and D are *equivalent* if there exists an equivalence of categories between them.

Now, how are we to relate the above notion of equivalence to equivalence between *theories*? From the previous section, we might recall there exist at least two different ways of relating theories and categories: syntactic categories on the one hand and semantic categories on the other. Barrett and Halvorson (2015, 18) opt for the latter choice. In what follows, we shall take the strongest of the two proposals for the notion of semantic category as our definition. That is, given a first-order theory T, we let $\mathrm{Mod}(T) := \mathrm{Mod}_2(T)$ be its semantic category.[14] In particular, this means that the morphisms in our semantic categories will be given by elementary embeddings between models.

With the above provisions in place, we now have:

Definition 4.2.26. Two theories T_1 and T_2 are *categorically equivalent* if their semantic categories $\mathrm{Mod}(T_1)$, $\mathrm{Mod}(T_2)$ are equivalent.

As before, we would like to show that this notion of equivalence is more liberal than the one directly preceding it, i.e. T-Morita equivalence. This will, however, require us to do some work. Fist, we need the following concept.

[14] The attentive reader may note that we had actually defined $\mathrm{Mod}_2(T)$ for *coherent* theories only. However, it is clear that nothing in the given definition is contingent on the theory being formulated in $L^g_{\omega\omega}$ and so we can readily generalize to the full first-order case.

Definition 4.2.27. Let T' be a Morita extension of T. We then define the *projection functor* $\Pi : \mathrm{Mod}(T') \to \mathrm{Mod}(T)$ to be the functor such that

- $\Pi(M) = M|_\Sigma$ for every object M in $\mathrm{Mod}(T')$;
- $\Pi(h) = h|_\Sigma$ for every morphism $h : M \to N$ in $\mathrm{Mod}(T')$.

Note that this definition does not require T' to be a Morita extension of T per se. We might just as well have defined the projection functor for any other type of extension. For our present purposes, however, it suffices to consider this functor in relation to Morita extensions.

Proposition 4.2.28. *If T' is a Moria extension of T, then the projection functor Π : Mod(T') $\to$ Mod(T) is essentially surjective.*

Proof. Recall theorem 4.2.20, which states that for every model M of T there exists a unique expansion of M to a model M' of T'. Hence, for any model M in Mod(T), we have a model M' in $\mathrm{Mod}(T')$ such that $\Pi(M') = M'|_\Sigma = M$. We conclude Π is essentially surjective. ■

Proposition 4.2.29. *If T' is a Moria extension of T, then the projection functor Π : Mod(T') $\to$ Mod(T) is faithful.*

Proof. Let $h : M \to N$ and $g : M \to N$ be morphisms in $\mathrm{Mod}(T')$ and suppose that $\Pi(h) = \Pi(g)$. This means that $h_\sigma = g_\sigma$ for every $\sigma \in \Sigma$. To demonstrate faithfulness, we now have to show that this identity also holds for the sort symbols from $\Sigma' \setminus \Sigma$.

Thus, let $\sigma \in \Sigma' \setminus \Sigma$. Since T' is a Morita extension of T, we can obtain σ from the sorts in Σ by a combination of the four operations on sort symbols defined above.

Suppose T' defines σ as a coproduct sort with function symbols ρ_1 and ρ_2 with arities $\sigma_1 \to \sigma$ and $\sigma_2 \to \sigma$ respectively. First, note that, for $i \in \{1, 2\}$,

$$h_\sigma \circ \rho_i^M = \rho_i^N \circ h_{\sigma_i}, \tag{4.42}$$

which we can see by applying the fact h is an elementary embedding to the formula $\rho_i(x) = y$. Next, observe that

$$\rho_i^N \circ h_{\sigma_i} = \rho_i^N \circ g_{\sigma_i}, \tag{4.43}$$

since we know by assumption that $h_i = g_i$. Now, again using the fact that h_σ is an elementary embedding, we can infer

$$\rho_i^N \circ g_{\sigma_i} = g_\sigma \circ \rho_i^M. \tag{4.44}$$

Successive application of equations (4.42)–(4.44) thus yields:

$$h_\sigma \circ \rho_1^M = g_\sigma \circ \rho_1^M \text{ and } h_\sigma \circ \rho_2^M = g_\sigma \circ \rho_2^M. \tag{4.45}$$

Finally, recall that since T' defines σ as the coproduct sort of σ_1 and σ_2, it holds that

$$M \models \forall_\sigma z[\exists!_{\sigma_1} x[\rho_1(x) = z] \vee \exists!_{\sigma_2} y[\rho_2(y) = z]]. \tag{4.46}$$

So for any $c \in M_\sigma$, we can either find $a \in M_1$ or $b \in M_2$ such that $\rho_1(a) = c$ or $\rho_2(b) = c$. Combining this observation with (4.45), we conclude that $g_\sigma = h_\sigma$, as desired.

Next, consider the case where σ is defined as a quotient sort with function symbol ε with arity $\sigma_1 \to \sigma$. Using the same reasoning as in the coproduct case, we see that

$$h_\sigma \circ \varepsilon^M = g_\sigma \circ \varepsilon^M. \tag{4.47}$$

Moreover, since T' defines σ as a quotient sort with respect to σ_1, we know

$$M \models \forall_\sigma z \exists_{\sigma_1} x[\varepsilon(x) = z]. \tag{4.48}$$

Combining (4.47) and (4.48), we may again conclude $g_\sigma = h_\sigma$.

The cases for product sorts and subsorts proceed in a similar fashion and may be found in (Barrett & Halvorson 2015, 19–20). ■

Finally, we have:

Proposition 4.2.30. *If T' is a Moria extension of T, then the projection functor $\Pi : Mod(T') \to Mod(T)$ is full.*

We are now ready to state the relation between categorical equivalence and T-Morita equivalence:

Theorem 4.2.31. *If two theories are T-Morita equivalent, then they are categorically equivalent. The converse, however, does not generally hold true.*

Proof. Suppose T_1 and T_2 are T-Morita equivalent. By definition, there then exist sequences $T_1^1, \ldots, T_1^n$ and $T_2^1, \ldots, T_2^m$ of theories such that each successor is a Morita extension of its predecessor and T_1^n and T_2^m are logically equivalent. The latter, by definition, implies that the semantic categories $\mathrm{Mod}(T_1^n)$ and $\mathrm{Mod}(T_2^m)$ are equivalent. The former, by propositions 4.2.28, 4.2.29 and 4.2.30, implies that for each T_1^i, T_1^{i+1} and T_2^j, T_2^{j+1}, the corresponding projection functors are equivalences between categories. Using the transitivity of equivalences of categories, we now obtain an equivalence of categories between T_1 and T_1^n, and another one between T_2 and T_2^m. Now, by the equivalence of $\mathrm{Mod}(T_1^n)$ and $\mathrm{Mod}(T_2^m)$, we conclude that also $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are equivalent. Hence, T_1 and T_2 are categorically equivalent.

To see that the converse implication fails, consider the following example. Let $\Sigma_1 = \{\sigma_1, P_0, P_1, \ldots\}$ and $\Sigma_2 = \{\sigma_2, Q_0, Q_1, \ldots\}$ be signatures with each one sort symbol and countably many unary relation symbols. Next, let

$$T_1 = \{\exists!_{\sigma_1} x[x = x]\}^{\mathrm{cl}}, \tag{4.49}$$

$$T_2 = \{\exists!_{\sigma_2} y[y = y], \forall_{\sigma_2} y[Q_0(y) \to Q_1(y)], \forall_{\sigma_2} y[Q_0(y) \to Q_2(y)], \ldots\}^{\mathrm{cl}}, \tag{4.50}$$

be theories over Σ_1 and Σ_2 respectively. Clearly, every model of either T_1 or T_2 will consist of a single element. This fact has two particular consequences:

- Within $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$, there exists at most one morphism from any one object to another.
- For any two structures M, N in $\mathrm{Mod}(T_1)$ or $\mathrm{Mod}(T_2)$, we can easily invert any elementary embedding from M to N. Thus, every morphism in $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ will be an isomorphism.

By lemma A.2.11 from appendix A, the above two facts now imply that $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are discrete categories. This means that both $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are equivalent to categories whose only morphisms are identity morphisms. Let us refer to these categories as S_1 and S_2 for $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$, respectively. Clearly, any bijection between the objects of S_1 and S_2 would naturally induce an equivalence of categories between S_1 and S_2. Invoking the transitivity of equivalences of categories, we could then conclude that $\mathrm{Mod}(T_1)$ and $\mathrm{Mod}(T_2)$ are equivalent.

To see that we can indeed construct a bijection from $\mathrm{ob}(S_1)$ to $\mathrm{ob}(S_2)$, we will show both to be equinumerous to the set of all infinite binary sequences. First, we define a map f from $\mathrm{ob}(S_1)$ to $\{0,1\}^{\mathbb{N}}$ as follows. Let a be an object in S_1, i.e. a is an isomorphism class of models of T_1. Let $M = (\{m\}, P_0^M, P_1^M, \dots)$ be a model in a. Then, let $f(a)$ be the sequence $(\alpha_n)_{n\in\mathbb{N}}$ such that

$$\alpha_n = \begin{cases} 0, & \text{if } m \notin P_n^M, \\ 1, & \text{otherwise.} \end{cases}$$

Since, by definition, every model in a is isomorphic to M, we see that this definition does not depend on the choice of model M. It is now readily verified that f is a bijection.

In a similar fashion, we define a map g from $\mathrm{ob}(S_2)$ to $\{0,1\}^{\mathbb{N}}$ as follows. Let b be an object in S_2, i.e. b is an isomorphism class of models of T_2. Let $K = (\{k\}, Q_0^K, Q_1^K, \dots)$ be a model in b. We define the map g as follows:

(i) If K is the structure $(\{k\}, \varnothing, \{k\}, \{k\}, \dots)$, then $g(b) = (0, 1, 1, \dots)$.

(ii) If K is the structure $(\{k\}, \{k\}, \{k\}, \{k\}, \dots)$, then $g(b) = (1, 1, 1, \dots)$.

(iii) If none of the above hold, then $g(b)$ is defined to be the sequence $(\beta_n)_{n\in\mathbb{N}}$ such that

$$\beta_n = \begin{cases} 0, & \text{if } k \notin Q_{n+1}^K, \\ 1, & \text{otherwise.} \end{cases}$$

Again, it is easily checked that this definition is well-defined and gives us a bijection between $\mathrm{ob}(S_2)$ and $\{0,1\}^{\mathbb{N}}$. We conclude that $\mathrm{ob}(S_1)$ and $\mathrm{ob}(S_2)$ are both of cardinality $2^{\aleph_0}$ and, hence, that there exists a bijection between them. Thus, we see that T_1 and T_2 are categorically equivalent.

Our next aim is now to show that the theories are not T-Morita equivalent. For sake of contradiction, let us suppose that T_1 and T_2 in fact *are* T-Morita equivalent. Then there exists a common Morita extension T of T_1 and T_2 to some signature Σ. Now, consider the Σ_2-sentence $\forall_{\sigma_2} y Q_0(y)$. Trivially, it holds that this is also a Σ-sentence. By theorem 4.2.17, this implies there exists a Σ_1-sentence φ such that

$$T \models \forall_{\sigma_2} y Q_0(y) \leftrightarrow \varphi. \tag{4.51}$$

We show that this bi-implication leads to absurdity. The idea here is that T_2 permits a sentence that completely determines the structure of its models, viz. the sentence $\forall_{\sigma_2} y Q_0(y)$, while we do not have such a sentence for T_1. That is to say, the sentence φ as defined above cannot exist.

We demonstrate that φ has the following property. Let ψ be a Σ_1-sentence and suppose that $T_1 \models \psi \to \varphi$. Then, we have either $T_1 \models \neg\psi$ or $T_1 \models \varphi \to \psi$. This is sufficient to arrive at a contradiction, since we can then take

$$\psi := \varphi \wedge \forall_{\sigma_1} x P_i(x), \tag{4.52}$$

where P_i is a relation symbol not occurring in φ, and observe that while obviously $\psi \to \varphi$, we can ascertain neither $T_1 \models \neg\psi$ nor $T_1 \models \varphi \to \psi$.

Let us now set about showing that φ indeed has the property stated above. Let ψ^* be the Σ_2-sentence, obtained from theorem 4.2.17, such that

$$T \models \psi \leftrightarrow \psi^*. \tag{4.53}$$

Combining bi-conditionals (4.51) and (4.53) with the fact that $T \models \psi \to \varphi$, we get that $T \models \psi^* \to \forall_{\sigma_2} y Q_0(y)$. Moreover, since $\psi^* \to \forall_{\sigma_2} y Q_0(y)$ is a Σ_2-sentence, we have, by theorem 4.2.23, that $T_2 \models \psi^* \to \forall_{\sigma_2} y Q_0(y)$. Now, suppose $T_1 \not\models \neg\psi$, i.e. ψ holds in some model of T_1. Then it is also the case that ψ^* holds in some model M of T_2. We show this implies $T_2 \models \forall_{\sigma_2} y Q_0(y) \to \psi^*$.

Let M' be an arbitrary model of T_2 and suppose $M' \models \forall_{\sigma_2} y Q_0(y)$. Inspecting T_2, we see that the sentence $\forall_{\sigma_2} y Q_0(y)$ completely determines the structure of the models of T_2. Hence, the models M and M' will be isomorphic. Now, by assumption, $M \models \psi^*$. Thus, we infer that also $M' \models \psi^*$. We conclude $T_2 \models \forall_{\sigma_2} y Q_0(y) \to \psi^*$. Translating back again to T_1, this yields $T_1 \models \varphi \to \psi$, which establishes that φ indeed has the above described property.

■

We see that categorical equivalence indeed provides us with a more liberal criterion for theoretical equivalence than T-Morita equivalence. But should we view this as a point in favor for T-Morita equivalence or, alternatively, as an advantage of categorical equivalence? Barrett and Halvorson opt for the former. As they point out (2015, 23), we could well argue that the theories T_1 and T_2 given in the proof of theorem 4.2.31 are sufficiently different so as to be considered inequivalent. In particular, the theory T_2 contains a unary relation symbol Q_0 that, if satisfied, complete determines the truth values of all the other relation symbols $Q_1, Q_2, \ldots$ in the theory. The theory T_1, by contrast, contains no such 'special' relation symbol.

It would thus seem that with T-Morita equivalence we have arrived at just the right level of generality to capture equivalence between theories. Now, however, the following concern arises: how are we to incorporate the notion of T-Morita equivalence in the program of the categorical school of metascience? Indeed, we have started our discussion of the categorical approach by looking at how to transform bare sets of sentences/models into nice, structured categories. Yet, throughout our entire discussion of T-Morita equivalence, we have assumed theories to be given by bare sets of first-order sentences. Moreover, the only thread connecting our ruminations on theoretical equivalence with the categorical view on theories/models, i.e. the the notion of categorical equivalence, has been shot down as being too general for our purposes! So how exactly *are* we supposed to bring T-Morita equivalence and the categorical research program together in a single overarching view? We shall take up this matter in the following subsection.

4.2.2 A Category of Theories

Let us forget for a moment the notion of T-Morita equivalence and return to the fundamental concept of the categorical approach, i.e. the notion of a syntactic category.[15] Now, given that the categorical view prescribes that we formalize theories by means of their respective syntactic categories, we might very well expect that intertheory relations between theories are similarly explicated in terms of syntactic categories. In particular, we would expect our notion of theoretical equivalence to be defined as a relation holding between syntactic categories. The most straightforward manner in which we could define a notion of equivalence between theories would then be given by:

Definition 4.2.32. Let T_1, T_2 be coherent theories. T_1 and T_2 are said to be *equivalent* if the syntactic categories C_{T_1} and C_{T_2} are equivalent in the category-theoretic sense.

With this definition for theoretical equivalence in place, we could then take our study of intertheory relations one step further by considering the *category of theories* **Th**, with syntactic categories of coherent theories as objects and morphisms determined by choosing the relation of category-theoretic equivalence as the isomorphism relation within **Th**. In the literature, this category is also known as the *coherent category* and denoted **Coh**. Thus, assuming our notion of theoretical equivalence to be given by definition 4.2.32, the study of intertheory relations reduces to the study of the properties of this particular category **Coh**.

Halvorson and Tsementzis (2016) indeed consider **Coh** as a possible candidate for the category of theories. They note, however, that “for reasons too detailed to go into here” (2016, 9) the corresponding notion of equivalence as set out in definition 4.2.32 is too strict a notion of theoretical equivalence. To

[15] The inferiority of semantic categories relative to syntactic categories was already established in subsection 4.1.2.

remedy this situation, we might attempt to liberalize the criterion expressed in definition 4.2.32 by demanding not the equivalence of the *syntactic categories* of the theories in question, but requiring the equivalence of some more general type of categories based on the theories. Here, by 'more general' I mean a type of category that can be considered to contain less information about the underlying theory T than contained in the corresponding syntactic category C_T.

Indeed, it is exactly the above strategy we find implemented in (Halvorson & Tsementzis 2016). More specially, Halvorson and Tsementzis (2016, 7) consider two ways in which we can construct a new type of category from a given theory in coherent logic. Namely:

(i) Given a coherent theory T, consider its associated **classifying topos** $\mathcal{E}_T$.

(ii) Given a coherent theory T, consider the **pretopos completion** $P(C_T)$ of the corresponding syntactic category C_T.

Both of the above notions are rooted in the subfield of category theory known as *topos theory.* Roughly speaking, topos theory may be construed as the study of categories which behave similarly to the category of sets. Unfortunately, the formal explication of the above concepts is a rather involved process and would take us well beyond the scope of the present thesis. The interested reader is invited to consult appendix B for an introduction to the field of topos theory.[16] For our present purposes, it will suffice to sketch here the manner in which the two topos-theoretic concepts formulated above can be connected to the notion of T-Morita equivalence.

Now, let us consider the notion of theoretical equivalence obtained by replacing in definition 4.2.32 syntactic categories by either one of the objects mentioned in (i) and (ii). Say we opt for the former choice. We then obtain a notion of theoretical equivalence closely related to that of T-Morita equivalence. More specifically:

Definition 4.2.33. Let T_1, T_2 be coherent theories. T_1 and T_2 are said to be *J-Morita equivalent* if the classifying topoi $\mathcal{E}_{T_1}$ and $\mathcal{E}_{T_2}$ are equivalent in the category-theoretic sense.[17]

As it turns out, the choice between options (i) and (ii) for defining J-Morita equivalence was inessential, as expressed by the following proposition.

Proposition 4.2.34. *Let T_1, T_2 be coherent theories. Then the classifying topoi $\mathcal{E}_{T_1}$ and $\mathcal{E}_{T_2}$ are equivalent if and only if the pretopos completions $P(C_{T_1})$ and $P(C_{T_2})$ are equivalent.*

The above proposition thus provides us with the following, alternative characterization of J-Morita equivalence.

[16] The concepts involved in the definition of the notion of a topos are of a more general category-theoretic import and will also prove useful later on in the subsequent chapter.

[17] Henceforth, I shall omit this qualification.

Corollary 4.2.35. *Let T_1, T_2 be coherent theories. Then T_1 and T_2 are J-Morita equivalent if and only if the pretopos completions $P(C_{T_1})$ and $P(C_{T_2})$ are equivalent.*

Now, let us finally return to the notion of T-Morita equivalence, as discussed in the previous subsection. The link between the abstract, category-theoretic notion of J-Morita equivalence and the more intuitive, syntactic notion of T-Morita equivalence is established by Tsementzis (2015):

Theorem 4.2.36. *Let T_1, T_2 be coherent theories. Then T_1 and T_2 are T-Morita equivalent if and only if they are J-Morita equivalent.*

Hence, we see that in the coherent fragment of first-order logic, the notion of T-Morita equivalence, through its equivalence with J-Morita equivalence, gives rise to a notion of equivalence defined on category-theoretic entities.

In the same manner as for definition 4.2.32, we see that a definition of theoretical equivalence in terms of T-Morita equivalence now allows us to lift the study of intertheory relations over to the category of theories **Th**. This time, however, this category is not identical to the coherent category **Coh**. In line with the differing notions of theoretical equivalence, we also obtain a different category of theories, viz. the category **Pretop** consisting of objects known as *pretopoi*. While we will not be examining the category **Pretop** in further detail, let us note that it is this category of theories that Halvarson and Tsementzis (2016, 17) consider to be the most suitable for the analysis of intertheory relations. At this stage, we could revisit the doubts expressed in section 4.1 concerning the applicability of notions grounded in coherent logic to cases of actual, concrete scientific theories. For the moment, however , let us conclude our discussion on the categorical school of metascience and redirect our focus to expounding a new approach to logical metascience in the subsequent chapter.

Chapter 5

Second Wave of Logical Abstractivism

In this final chapter, I draw the outline of a new methodology for the application of abstract model theory to metascience. Motivating this revision, I begin by arguing in section 5.1 that the first wave of logical abstractivism has failed to bring together the abstract conception of logic with the study of metascience in a convincing manner. To argue this point, I will present numerous different aspects of the first wave which I believe make it ill-suited for metascientific application. Next, I lay the groundwork for a new abstractivist approach to metascience by considering two recent strands of research within abstract model theory in section 5.2. Having completed the groundwork, I will finally theorize how we can use these new frameworks to generate, what I hope will become, a second wave of logical abstractivism in section 5.3.

5.1 The Failure of FWLA

In this section, I will argue that logical abstractivism is due for a renewal. More specifically, I will contend that the first wave of logical abstractivism, as expounded by Pearce and Rantala, does not provide us with a satisfactory combination of metascience with the abstract conception of logic. This argument will proceed in two stages. In subsection 5.1.1, I argue that the FWLA fails to utilize the most prominent and conceptually interesting aspects of semi-abstract model theory, thereby demonstrating the questionable methodological basis underlying the approach. Next, I will contend that the FWLA also fails in the execution of its methodology. More specifically, I show in subsection 5.1.2 how the FWLA has done little to advance the structuralist program of metascience.

Below, I will frequently call upon the characteristic features (i)-(iv) of the FWLA as discussed in subsection 3.2.3. Recall we had established there that the FWLA may be construed as a flavor of logical abstractivism characterized by a focus on (i) semi-abstract model theory, (ii) extensions of first-order logic,

(iii) logical liberalization and (iv) the structuralist approach to metascience. Of these, I consider (i) to be an inevitable result of the time period in which the FWLA was expounded. Choice (ii), however, is of a quite restrictive and contingent nature, shoehorning us into accepting (iii) and (iv) which, in turn, further restrict the scope of the FWLA. In the following subsections, I will argue that properties (i)-(iv) make the framework of the FWLA into one that fails to fully utilize the potential of its employed formalism, viz. semi-abstract model theory, and in addition does little to advance the framework of structuralism. This, I hold, makes its failure complete and brings to the fore the need for a new brand of logical abstractivism.

5.1.1 FWLA and Abstract Model Theory

To my knowledge, the premise of applying methods from semi-abstract model theory to metascientific considerations is an idea wholly original to Pearce and Rantala. Yet, a cursory view of their writings reveals only a modest utilization of the formal machinery underlying this approach. Recall from section 3.1 that semi-abstract model theory may be seen to be the culmination of three separate strands of research within model theory during the 1950s and '60s, viz. research on cardinality quantifiers, infinitary logics and Lindström-style characterization theorems. In subsection 3.1.2, we familiarized ourselves with the latter of these research traditions, encountering the particularly general abstract maximality theorem.[1] This theorem, we may recall, served to characterize certain *fragments* of first-order logic with, for any signature Σ, the class of Σ-models given by a first-order Δ-elementary class. The abstract maximality theorem, in turn, opened the door to the characterization of a variety of different, non-standard logics, such as the logic of topological structures.

Consider, now, the exposition of the Pearce-Rantala approach offered in section 3.2. Two aspects of its methodology stand out immediately. Namely, within their approach, Pearce and Rantala opt to restrict the concept of abstract logic to those and only those abstract logics whose classes of models consisted of subclasses of first-order models and, most importantly, were extensions of first-order logic. This decision, codified as point (ii) above, already seems to cut off from us the level of generality provided by the abstract maximality theorem. Non-standard logics which may prove to be of significant interest to the formalization of scientific theories, such as topological logic, now become much more difficult to include within the scope of our investigations.[2]

[1] While the choice to focus only on characterization theorems was, for the most part, one of convenience, we may note there is also an intrinsic reason why this strand of research seems more attractive to the aspiring metascientist than the preceding two. Whereas the research on cardinality quantifiers and infinitary logics presupposes already that we are working with a fixed, definite class of logics, usually a class of logics extending first-order logic, the work on Lindström-style characterization results has *in principle* no such a priori restriction. In fact, the abstract maximality theorem itself is a testament to the fact that the search for Lindström theorems provides us with a most natural environment for considering a wide variety of logical systems, not limited to first-order extensions.

[2] To be precise, it prevents us from including topological and other non-standard logics as

Suppose, however, that we accept the restriction to extensions of first-order logic as given. We find then in the FWLA yet another, non-trivial choice being made that further constricts the manner in which we can involve semi-abstract model theory into our metascientific frameworks. This is choice (iii), i.e. the choice to employ abstract logics so as to promote *logical liberalization*, and it may be considered as *the* defining characteristic of the FWLA. For sake of clarity, let me make explicit here what it is I refer to exactly by this mantra. Consider again the propositions offered by Pearce (1985) concerning the structuralist concept of reduction, as described in subsection 3.2.2. We may note that, for these propositions to hold, one only needs to find *one* particular abstract logic satisfying the conditions stated in the propositions' hypotheses. Most importantly, note that the relation $\mathcal{R}$ between models need not necessarily be a first-order definable relation.[3] Instead, $\mathcal{R}$ is merely required to be definable relative to some suitable abstract logic L having a suitable compactness property δ.[4] Which abstract logic L this would then be, is irrelevant. What matters only is the *existence* of such an abstract logic. The crux of the matter here is that the Pearce-Rantala approach 'liberates' metascientific concepts and statements from the yoke of a fixed, underlying logic such as $L_{\omega\omega}$. Thus, if a given logic turns out to be too weak expressively to define, for instance, a relation $\mathcal{R}$ between models, we are free to move to a different logical system strong enough to meet our requirements. With this, we have arrived at the essence of logical liberalization as well as the FWLA in its entirety.

In what sense does adherence to logical liberalization represent a proper restriction of the philosophy of logical abstractivism? As noted in subsection 3.2.3, logical abstractivism may be perceived as necessitating a commitment to this modus operandi. After all, so the dilemma goes, we either have or have not a system of logic that has an adequate combination of expressive power and nice model-theoretic properties. If we have, we simply take that logic as the logic underlying our metascientific frameworks. If we have not, we are left with no choice but to pursue logical liberalization. Such a dilemma, however, would presuppose an overly narrow view on the role of logic in the formalization of scientific theories. This tunnel vision, I hold, can be attributed to adherence to point (ii), i.e. the restriction of abstract logics to extensions of first-order logic. Given this presupposition, logical liberalization seems to be the only plausible way in which abstract logics might be brought into the study of metascience. More specifically, while it is not strictly *necessitated*, adherence to point (ii) does heavily *suggest* to us that the general role of logic in metascientific methodology is similar to the role that has traditionally been assigned to first-order logic, i.e. to represent the contents of scientific theories in a formal language as descrip-

they are construed in (Flum 1985), i.e. as particular *fragments* of first-order logic satisfying some suitable invariance conditions.

[3] We call $\mathcal{R}$ *first-order definable* if its associated class of models $\mathfrak{R}$, defined by $(A, B) \in \mathfrak{R}$ if and only if $\mathcal{R}(A, B)$, is either an elementary or projective class in $L_{\omega\omega}$.

[4] We call $\mathcal{R}$ *definable* relative to some abstract logic L with compactness property δ if its associated class of models $\mathfrak{R}$ is δ-elementary or δ-projective in L.

tively precise as possible.[5] If this particular methodology is taken for granted then we are automatically committing ourselves to a view on logics which exalts expressive power as one of the most desirable features a logic may possess. With this in mind, a natural question then presents itself: how strong of a logic can we permit ourselves to work in without throwing away most of the useful results of first-order model theory? Slightly rephrased, this becomes: does there exists a logic having such and such model-theoretic properties that is also strong enough to express such and such? This line of inquiry, however, leads the metascientist seeking to import the tools of semi-abstract model theory straight to the methodology of logical liberalization. Hence, we see that logical liberalization is by no means necessitated by logical abstractivism and that adoption of this methodology is in fact a contingency particular to the Pearce-Rantala approach, stemming from its focus on extensions of first-order logic.[6]

Briefly reflecting on this subsection's title reveals that besides semi-abstract model theory there also exist other model-theoretic approaches to the abstract study of logic, having both many similarities and deviations from the field of *semi*-abstract model theory. In my usage, the term *abstract model theory* will serve as an umbrella term meant to designate the entirety of such model-theoretic approaches.[7] Now, we might, in addition to the criticism expounded above, also scrutinize the FWLA with respect to these alternative approaches. Such an assessment, however, would necessarily be of an anachronistic nature. For most of the second half of the twentieth, semi-abstract model theory stood as the sole framework for model-theoretic investigations into universal logic.[8] Hence, it seems hardly appropriate to assign blame to the FWLA for not fully utilizing the potential of frameworks that had not even come into existence at the time of its own genesis. This being said, the more recent frameworks *do*, in fact, possess certain properties which would make them most suitable to serve as the formal basis for a new, second wave of logical abstractivism. I shall return to this point in section 5.2 and again in section 5.3. This should, however, be interpreted strictly as an argument *in favor* of the second wave, rather than an argument directed *against* the first wave.

5.1.2 FWLA and Structuralism

The last subsection makes clear there are some serious reservations to be had about the manner in which the FWLA utilizes the resources of semi-abstract model theory to the metascientific cause. However, *given* their particular methodology, how well do Pearce and Rantala succeed in revitalizing the metascience

[5] See (Montague 1974) for a paradigmatic example of this tradition.

[6] Recalling the terminology of *first-order fixation* from section 1.1, we might summarize the preceding discussion by the mantra that *the methodology of the FWLA fails to transcend the bounds of first-order fixation.*

[7] Usually, the term *abstract model theory* is used in a more narrow sense to refer to what I call here *semi-abstract model theory.*

[8] Recall that *universal logic* serves as an umbrella term with which to designate *all* frameworks seeking to understand logical systems from an abstract point of view, including but not limited to model-theoretic vantage points.

of old?[9] By point (iv), we know that the Pearce-Rantala approach focuses primarily on applying semi-abstract model theory to reinvent the framework of structuralism. Thus, the question of how well the FWLA meshes with the metascience of old can be seen to reduce to the query of how well it succeeds in improving structuralism. Taking up this question, I shall confine my attention to the work of Peace (1985) on the structuralist notion of reduction, as discussed in section 3.2, which appears to be the most formally elaborate entry in the literary corpus of the FWLA as well as a reasonably representative member of this body of work.

Inspecting the results set out in section 3.2, we may be left feeling somewhat disappointed. The elaborate formal machinery of abstract semantic systems and Feferman's uniform reduction theorem, do not seem to pay off in terms of mathematically interesting results. Looking at proposition 3.2.20, we see that it manges only to characterize the notion of reduction in the sense that it provides us with some sufficient conditions for Pearce's model-theoretic notion of translation, as defined for theory-cores, to coincide with the structuralist concept of reduction. This result thus seems to provide us with a rather lopsided characterization. Moreover, the sufficient conditions presented in this proposition are of a rather specific and ad hoc nature, which contributes to an overall impression of artificialness of the obtained result.

Putting to the side the above 'internal' criticism of the Pearce-Rantala approach, we might next ask to what extent the development of structuralism has been impacted concretely by the revisions proposed in the FWLA. For an answer, we can turn to the discussion presented in section VI.7 of (Balzer et al. 1987), which is one of the few occasions where the structuralists explicitly consider the potential role of language in their framework. Accordingly, it is also in this place that Bazler et al. reflect on the possible value of logic and the FWLA for the explication of structuralism.[10] More specifically, they consider three aspects of their concept of reduction which may stand to benefit from the introduction of language. Of these, the first two points relate directly to the FWLA, while the third one seems to be of a more general nature. I shall thus confine my discussion here to the first two points pertaining to the interplay of language and reduction.[11]

First, we can note that Balzer et al. (1987, 308) acknowledge one point of criticism set forth by Pearce (1985, 135–40), which pertains to a paraphrasing of condition (ii) of definition 2.4.27. This condition, as we may recall, required

[9] That is, how well do they succeed in revitalizing the metascientific frameworks existing concurrently with their own approach?

[10] To be precise, the bibliography following chapter VI contains two references to works falling within the FWLA. Of these, one paper can be seen to coincide approximately with the material of (Pearce 1985) as discussed in section 3.2.

[11] The third point taken up by Balzer et al. (1987 313–20) pertains to the philosophically significant notion of *incommensurability*. Indeed, this notion has also been at the center of a fair number of treatises within the FWLA. Nevertheless, I have opted to omit the discussion of this notion from this thesis, as it would bring with it a host of tangential philosophical issues. The discussion of incommensurability in (Balzer et al. 1987), moreover, does not seem to be informed to any significant extent by the FWLA.

that

$$\text{for all } x' \in M_p'\text{: if } x' \in M' \text{ and } (x', x) \in \rho\text{, then } x \in M.$$

As remarked in subsection 2.4.2, Balzer et al. motivate this requirement by the informal reading that the 'laws' of theory-element T can be 'derived' from the 'laws' in theory-element T'. Pearce, however, notes the problematic nature of such a characterization in the absence of any language within the structuralist account. For without language, the argument goes, any analogy employing the concepts of law and derivability can only be considered disingenuous. In response, Balzer et al. essentially subscribe to Pearce's proposed manner of introducing language into their framework and note that such a treatment is indeed beneficial in elucidating the nature of the employed paraphrasing.

A second aspect of (Pearce 1985) discussed by Balzer et al. (1987, 311–3) is the content of proposition 3.2.20. In this instance, however, the judgment of the structuralists is more reserved. More specifically, they note the translation produced by this proposition, through the formal machinery of the uniform reduction theorem, fails in some crucial regards to qualify as a proper type of translation. More specifically, it is argued that the notion of translation as employed in the Pearce-Rantala approach, inadequately reflects those aspects of translation pertaining to 'preservation of meaning' of terms in scientific languages and hence can not be considered a 'proper translation'. As a result, Balzer et al. do not subscribe to the view that *reduction implies translation*, as stated in proposition 3.2.20. Thus, in contrast to the previous point, Balzer et al. do not consider proposition 3.2.20 as providing us with a natural incentive for introducing language into the structuralist framework.

Taking stock, we see that the impact of the FWLA on the program of structuralism, as explicated in (Balzer et al. 1987), is limited to the justification of a particular paraphrasing of condition (ii) in the structuralists' definition of direct reduction. In no essential way does language, let alone abstract logics, enter into their account. In practical terms, it thus seems that the influence of the FWLA on the structuralist approach to metascience, and by extension to the entire metascience of old, can be considered negligible.

5.2 New Formalisms

In the preceding section, it was noted how new developments within the field of abstract model theory can help us transcend some of the perceived drawbacks of the FWLA and be of much value in erecting a new methodology for logical abstractivism. The present section is therefore devoted to the exploration of these new formalism, in much the same way section 3.1 was devoted to obtaining an overview of semi-abstract model theory. The frameworks in question are two in number. The first falls squarely within the scope of semi-abstract model theory; the study of **abstract modal logics** is an extension of the methods for obtaining Lindström-style characterization theorems to the domain of modal logic. By contrast, the second framework we will be taking up represents an entirely new direction in the study of abstract model theory. **Institution-independent model theory**, or *institutional model theory* for short, presents us with a formalization of the concept of logical system that may be labeled *fully abstract*, in contrast to the semi-abstract treatments we have encountered up until this point.

Having completed our survey tour, we will in section 5.3 be concerned with the application of the above frameworks, as well as 'old' results of semi-abstract model theory, to the metascientific enterprise. It shall then become clear just how much the first-order fixation of both the old and new approaches to metascience, as well as the FWLA itself, has constrained the application of logic to the formal analysis of science.

5.2.1 Abstract Modal Logics

The field of modal logic is one of myriad mathematical and philosophical applications.[12] For our present purposes, we may think of 'a modal logic' as a logic (propositional, first-order, etc.) whose syntax is augmented by the presence of some additional, unary connective symbols $\Diamond_1, \ldots, \Diamond_n$ called *modal operators*.[13] Of particular fundamentality is the logical system of *basic modal logic*, symbolically: $\mathcal{BML}$. Notions such as formula, model and truth are defined for $\mathcal{BML}$ in much the same way as they were in the first-order case, modulo a number of characteristic differences. Throughout this subsection, it will be assumed we are working with a fixed number of modal operators $\Diamond_1, \ldots, \Diamond_n$ for all modal logics under consideration. As is customary in modal logic, we define a dual operator $\Box_i := \neg\Diamond_i\neg$ for every $1 \leq i \leq n$.

Definition 5.2.1. A *propositional letter* P is a meaningless, syntactic symbol. It is assumed we have a countably infinite supply of propositional letters.

Definition 5.2.2. A *signature* Σ for $\mathcal{BML}$ is a set of propositional letters.

Definition 5.2.3. Let Σ be a signature. Then the set of Σ-*sentences* for $\mathcal{BML}$ is the smallest set containing every $P \in \Sigma$ that is closed under $\wedge, \vee, \rightarrow, \neg$ and the modal operators $\Diamond_1, \ldots, \Diamond_n$.

Definition 5.2.4. A *frame* $\mathcal{F}$ is a tuple $(W^{\mathcal{F}}, R_1^{\mathcal{F}}, \ldots, R_n^{\mathcal{F}})$ where $W^{\mathcal{F}}$ is a set, the elements of which are referred to as *nodes*, and $R_i^{\mathcal{F}}$ is a binary relation on $W^{\mathcal{F}}$ for every $1 \leq i \leq n$.

Definition 5.2.5. Let Σ be a signature. Then a Σ-*model* for $\mathcal{BML}$ consists of a frame $(W, R_1^M, \ldots, R_n^M)$ together with a mapping $V^M : W^M \times \Sigma \rightarrow \{0, 1\}$. The mapping V^M is referred to as the *valuation* mapping of the model.

Definition 5.2.6. Let Σ be a signature. Then a *pointed* Σ-model is a pair (M, u) where M is a Σ-model and $u \in W$.

[12] Small portions of this subsection, such as definition 5.2.9 and the statements of theorems 5.2.11, 5.2.12 and 5.2.13, also occur in (Vos 2016).

[13] Though less common, it is also possible for modal operators to be k-ary, for arbitrary $k \in \mathbb{N}$. In this subsection, we will be concerned only with the unary variety.

Definition 5.2.7. Let Σ be a signature. Then the relation $\models$ between pointed Σ-models and Σ-sentences is defined inductively as follows:

- $(M, u) \models P$ iff $V^M(P, u) = 1$ for every $P \in \Sigma$,
- $(M, u) \models \neg\varphi$ iff not $(M, u) \models \varphi$,
- $(M, u) \models \varphi \vee \psi$ iff $(M, u) \models \varphi$ or $(M, u) \models \psi$,
- $(M, u) \models \varphi \wedge \psi$ iff $(M, u) \models \varphi$ and $(M, u) \models \psi$,
- $(M, u) \models \varphi \to \psi$ iff $(M, u) \models \varphi$ implies $(M, u) \models \psi$,
- $(M, u) \models \Diamond_i\varphi$ iff there exists some $u' \in W^M$ such that we have $uR_i^M u'$ and $(M, u') \models \varphi$ for every $1 \leq i \leq n$.

If $(M, u) \models \varphi$, we say (M, u) *satisfies* or *makes true* φ. If Γ is a set of sentences then we say (M, u) *satisfies* or *makes true* Γ, and write $(M, u) \models \Gamma$, if (M, u) satisfies every sentence $\varphi \in \Gamma$.

These definitions complete the specifications of the system of basic modal logic. It is not my intent to dwell on the independent import of this logical system or to motivate thoroughly the above definitions. The reader wishing for an accessible introduction to this topic may consult (Van Benthem 2010). In our present discussion, we will be interested in Lindström-style characterizations for certain modal logics, $\mathcal{BML}$ included. In particular, it is of great interest to determine whether we can formulate a theorem for modal logics comparable the abstract maximality theorem in terms of generality.

A natural course of action might now seem to take the definition of abstract logic as presented in subsection 3.1.1 and see how basic modal logic fits within it. Note, however, that definition 3.1.3 tacitly that any abstract logics, by and large, behave as predicate logics. For instance, the functions L^1 and L^2 of some abstract logic L are both defined on the domain *Sig* of first-order signatures. Accordingly, the closure conditions of definition 3.1.3 presuppose a syntax that behaves roughly like that of predicate, requiring L to be closed under existential quantification. It is thus necessary to revisit the notion of an abstract logic and determine how we can apply it to the family of modal logics.[14] To this end, we introduce the notion of an *abstract modal logic.*

Definition 5.2.8. An *abstract modal logic* is a triple $(L, Sent_L, \models_L)$, consisting of a map $L : Sig \to \mathcal{P}(Sent_L)$ sending each modal signature Σ to a set of objects called the Σ-sentences for L and $\models_L$ is a relation between pointed models and sentences of L, such that we have

[14] Note that modal logics come both in a propositional variety, $\mathcal{BML}$ being the prime example, as well as in a predicate variety, such as first-order modal logic. At any rate, even if we are concerned with characterizing some system of predicate modal logic, we still require a notion of abstract logic that includes the propositional case as well. After all, we would like to characterize such a predicate modal logic relative to the class of *all* modal logics, both predicate and propositional.

(i) For any pointed model (M, u) and $\varphi \in \mathit{Sent}_L$: if $(M, u) \models_L \varphi$ then $\varphi \in L(\Sigma_M)$, where Σ_M is a signature of cardinality equal to the number of binary relations of M.

(ii) *Monotonicity Property.* For any two signatures Σ, Σ': if $\Sigma \subseteq \Sigma'$, then $L(\Sigma) \subseteq L(\Sigma')$.

(iii) *Finite Occurrence Property.* For any signature Σ and $\varphi \in L(\Sigma)$, there exists a smallest, finite signature $\Sigma_\varphi \subseteq \Sigma$ such that $\varphi \in L(\Sigma_\varphi)$.

(iv) *Isomorphism Property.* For any signature Σ, pointed Σ-models $(M, u), (N, v)$ and Σ-sentence φ, if it holds that $(M, u) \models \varphi$ and $(M, u) \models_L \varphi$ then $(N, v) \models_L \varphi$.

(v) *Reduct Property.* For any signature Σ, pointed model (M, u) and $\varphi \in \mathit{Sent}_L$: if $\varphi \in L(\Sigma)$ and $\Sigma \subseteq \Sigma_M$ then $(M, u) \models_L \varphi$ if and only if $(M, u)|_\Sigma \models_L \varphi$, where $(M, u)|_\Sigma$ is the pointed model identical to (M, u) except in that its valuation map has been restricted to the domain $\Sigma \times W^M$.

(vi) *Renaming Property.* For any two signatures Σ, Σ', injection $\rho : \Sigma \to \Sigma'$ and $\varphi \in \mathit{Sent}_L$: if $\varphi \in L(\Sigma)$, then there exists $\varphi^\rho \in L(\Sigma')$ such that for all Σ-models A we have $A \models_L \varphi$ if and only if $A^\rho \models_L \varphi^\rho$.

(vii) *Closure Properties.* For any signature Σ, the set $L(\Sigma)$ is closed under Boolean connectives and modal operators.

As we can see, definitions 3.1.3 and 5.2.8 are largely identical, with minor modifications to account for the fact that we are now working with propositional signatures and pointed models instead of their first-order equivalents. The most significant alteration is that the we no longer require closure under existential quantification, since this condition is no longer always meaningful. Rather, we now require closure under modal operators.[15]

Notions such as extensions and equivalence for abstract modal logics are defined in complete analogon with those for abstract logics and will thus not be repeated here. Quintessential to all modal-logical investigations is the following preliminary:

Definition 5.2.9. Let Σ be a modal signature and let $(M, u), (N, v)$ be pointed Σ-models. Then a *bisimulation* Z between (M, u) and (N, v) is binary relation $Z \subseteq W^M \times W^N$ such that

- uZv,
- for all $P \in \Sigma$, we have $u_0 Z v_0$ implies $V^M(P, u_0) = 1$ iff $V^N(P, v_0) = 1$,

[15] Most definitions of abstract modal logics do not require the full range of properties (i)-(vii) I have included in definition 5.2.8. For instance, Enqvist (2013, 235) omits the finite occurrence and isomorphism properties because these are not required for the results he goes on to derive. Nevertheless, it seems to me prudent to adhere to the more stringent definition above, since otherwise we might open the door the unwanted ‘counterexamples’ to the notion of abstract modal logic, e.g. a system of ‘logic’ which fails to be invariant under isomorphism.

- for any $1 \leq i \leq n$, it holds that if $u_0 Z v_0$ and $u_0 R_i^M u_1$ then there exists v_1 such that $v_0 R_i^N v_1$ and $u_1 Z v_1$ (*forth condition*),
- for any $1 \leq i \leq n$, it holds that if $u_0 Z v_0$ and $v_0 R_i^N v_1$ then there exists u_1 such that $u_0 R_i^M u_1$ and $u_1 Z v_1$ (*back condition*).

Definition 5.2.10. Let Σ be a modal signature and let $(M, u), (N, v)$ be pointed Σ-models. Then (M, u) and (N, v) are called *bisimilar*, symbolically $(M, u) \sim (N, v)$, if there exists a bisimulation between (M, u) and (N, v).

The abstract study of modal logics can be said to have originated with De Rijke (1995), who proved the following Lindström theorem:

Theorem 5.2.11. *Let L be an abstract modal logic extending $\mathcal{BML}$. If L has a notion of finite rank and is invariant under bisimulations then L is equivalent to $\mathcal{BML}$.*

For a proof, along with the definition of a *notion of finite rank*, see (Vos 2016).

In a bid to transform De Rijke's result into a more traditional format, Van Benthem (2007) demonstrated how compactness and bisimulation invariance for an abstract modal logic L lead L having a notion of finite rank, thereby establishing:

Theorem 5.2.12. *Let L be an abstract modal logic extending $\mathcal{BML}$. If L has the compactness property and is invariant under bisimulations then L is equivalent to $\mathcal{BML}$.*[16]

The form of theorem 5.2.12 is indeed extremely reminiscent of the abstract maximality theorem, characterizing a logic in terms of compactness and an invariance property. It is, however, not evident how we are to merge the different scopes involved in both theorems: one ranges over a class of compact *abstract logics* satisfying some invariance property while the other ranges over a class of compact *abstract modal logics* satisfying a certain invariance condition. While it would be of much interest to investigate the compatibility of these two results in more detail, we shall presently turn out attention to another type of general Lindström theorem aimed specifically at modal logics.

The basis for this result is formed by a third modal Lindström theorem, formulated Otto and Piro (2008), characterizing basic modal logic augmented by the *global modality* $\mathcal{BML}[\forall]$.

Theorem 5.2.13. *Let L be an abstract modal logic extending $\mathcal{BML}[\forall]$. If L has the compactness property, the Tarski union property and is invariant under global bisimulations then L is equivalent to $\mathcal{BML}[\forall]$.*

For our present purposes, the precise definitions of the global modality and global bisimulations are unimportant. What is crucial here, is the inclusion of

[16] In proving this statement, Van Benthem presupposes an extended definition of abstract modal logics that includes, in addition to conditions (i)–(vii) above, an extra property called the *relativization property*.

the *Tarski union property*. We may recall that, intuitively, the Tarski union property expresses that the union of any elementary chain is an elementary extension of each model in the chain relative to an abstract logic L, cf. the definitions given in subsection 3.1.2. These definitions can now be carried over to the case of abstract modal logics in a straightforward manner. In particular, consider:

Definition 5.2.14. Let L be an abstract modal logic. Then L is said to have the *Tarski union property* if for every elementary chain relative to L it holds that

$$M_k \preceq_L \bigcup_{m \in \mathbb{N}} M_m \tag{5.1}$$

for every $k \in \mathbb{N}$, where the model

$$\bigcup_{m \in \mathbb{N}} M_m = (W^M, R_1^M, \ldots, R_n^M, V^M) \tag{5.2}$$

is defined by setting

- $W^M = \cup_{m \in \mathbb{N}} W^{M_m}$,
- $R_i^M = \cup_{m \in \mathbb{N}} R_i^{M_m}$ for each $1 \leq i \leq n$,
- $V^M(P, w) = 1$ iff $V^{M_m}(P, w) = 1$ for all $m \in \mathbb{N}$.

Now, generalizing the methods involved in the proof of theorem 5.2.13, Enqvist (2013) has formulated a general Lindström theorem for $\mathcal{BML}$. The generalization here lies in the fact we obtain a characterization for $\mathcal{BML}$ relative to a given class of frames satisfying certain general constraints. To understand what it means for an abstract modal logic to be characterized *relative* to some class of frames C, consider the following.

Definition 5.2.15. Let C be a class of frames. Then we say that a pointed model (M, u) is *in* C if it based on a frame in C.

Definition 5.2.16. Let C be a class of frames and X a class of pointed models in C. Then we say X is L-*definable* over C if $X = E \cap C$ for some elementary class E in L.

Definition 5.2.17. Let L, L' be abstract modal logics and C be a class of frames. Then we say L and L' are *equivalent* over C if for every subclass $C_0 \subseteq C$ and every class X of pointed models in C_0, we have: X is L-definable over C if and only if X is L'-definable over C.

Similarly, we can also define model-theoretic properties of some abstract modal logic relative to a given class of frames C.

Definition 5.2.18. Let L be an abstract modal logic. Then we say L has the *compactness property* over C, if for every signature Σ and set of Σ-sentences Γ we have: if every finite subset of Γ is satisfied by some pointed model in C then Γ itself is satisfied by some pointed model in C.

Definition 5.2.19. Let L be an abstract modal logic. Then we say L is *bisimulation invariant* over C, if for any signature Σ, pointed Σ-models $(M, u), (N, v)$ in C and Σ-sentence φ, if it holds that $(M, u) \models \varphi$ and $(M, u) \models_L \varphi$ then $(N, v) \models_L \varphi$.

Definition 5.2.20. Let L be an abstract modal logic. Then we say L has the *Tarski union property* over C, if for every elementary chain $(M_m)_{m\in\mathbb{N}}$ contained in C such that $\cup_{m\in\mathbb{N}} M_m$ is in C as well we have:

$$M_k \preceq_L \bigcup_{m\in\mathbb{N}} M_m \tag{5.3}$$

for every $k \in \mathbb{N}$.

Why would we be in interested in Lindström theorems relativized to some class of frames? First, we may note that is part and parcel of modal logical practice to consider logics defined for particular frames. Consider:

Example 5.2.21. Consider the system of modal logic obtained from $\mathcal{BML}$ by letting the corresponding frames contain a single binary relation. That is, consider all frames of form (W, R). Then we obtain:

- If no constraints are placed on R, the modal logic K.
- If R is reflexive, the modal logic known as T.
- If R is a preorder, the modal logic called S4.
- If R is an equivalence relation, the modal logic S5.

Of particular interest is the fact that the frame classes for T, S4 and S5 can all be specified by means of first-order sentences.

- The class of frames for the modal logic T is given by the class $C = \{(W, R) : (W, R) \models \forall x[xRx]\}$.
- The class of frames for the modal logic S4 is given by the class $C' = \{(W, R) : (W, R) \models \forall x\forall y\forall z[xRy \wedge yRz \to xRz]\} \cap C$.
- The class of frames for the modal logic S5 is given by the class $C'' = \{(W, R) : (W, R) \models \forall x\forall y[xRy \to yRx]\} \cap C'$.

Here, $\models$ denotes the first-order satisfaction relation and frames (W, R) are construed as first-order models.

In each of the above examples, the frame class under consideration can be specified by a particular first-order sentence. More specifically, the frame classes are definable by what is known as a *strict universal Horn formula.*

Definition 5.2.22. A first-order sentence is called a *strict universal Horn formula* if it is of the form $\forall\vec{x}[\varphi_1 \wedge \ldots \wedge \varphi_k \to \psi]$ for atomic formulas $\varphi_1, \ldots, \varphi_k, \psi$.

Setting $k = 0$ in the above definition, we see that sentences of the form $\forall\vec{x}\psi$, for atomic ψ, constitute strict universal Horn formulas as well.

It turns out that frame classes definable by strict universal Horn formulas are of special interest when it comes to Lindström theorems for $\mathcal{BML}$. More precisely, Enqvist (2013) provides us the following characterization result:

Theorem 5.2.23. *Let L be an abstract modal logic extending $\mathcal{BML}$ and C be a class of frames definable by a set of strict universal Horn formulas. Suppose L has the compactness property over C, the Tarski union property over C and is bisimulation invariant over C. Then L is equivalent to $\mathcal{BML}$ over C.*

As noted previously, the proof of theorem 5.2.23 relies to a large extent on a generalization of techniques employed in the proof of theorem 5.2.13. The starting point is provided by the following observations.

Lemma 5.2.24. *Let L be an abstract modal logic extending $\mathcal{BML}$, C be a class of frames, Γ a set of basic modal sentences and φ be a sentence in L. Suppose there exists no basic modal sentence ψ such that any pointed model in C satisfying Γ satisfies $\varphi \leftrightarrow \psi$ as well. Then for any basic modal sentence α, the same holds for either $\Gamma \cup \{\alpha\}$ or $\Gamma \cup \{\neg\alpha\}$. That is, either $\Gamma \cup \{\alpha\}$ or $\Gamma \cup \{\neg\alpha\}$ preserves the inexpressibility of φ.*

Proof. Suppose neither set preserves the inexpressibility of φ, i.e. there exist basic modal ψ, ψ' such that any pointed model in C satisfying $\Gamma \cup \{\alpha\}$ will satisfy $\varphi \leftrightarrow \psi$ and any pointed model in C satisfying $\Gamma \cup \{\neg\alpha\}$ satisfies $\varphi \leftrightarrow \psi'$. Then it also holds that any pointed model in C satisfying Γ will satisfy $\varphi \leftrightarrow (\alpha \wedge \psi) \vee (\neg\alpha \wedge \psi')$ as well. But this contradicts our original assumption on Γ, since the right side of this equivalence is a basic modal formula. ■

Proposition 5.2.25. *Let L be an abstract modal logic extending $\mathcal{BML}$ and C be class of frames. Suppose that L has the compactness property over C and that L is not equivalent to $\mathcal{BML}$ over C. Then there exist pointed models (M, u), (N, v) and a sentence φ in L such that (M, u) and (N, v) agree on all basic modal sentences, but do not agree on φ. That is, $(M, u) \models \psi$ iff $(N, v) \models \psi$ for all basic modal ψ, but $(M, u) \models_L \varphi$ and $(N, v) \not\models_L \varphi$.*

Proof. Since L and $\mathcal{BML}$ are not equivalent over C, we know there exists some subclass $C_0 \subseteq C$ such the class X of pointed models in C_0 is definable over C relative to either L or $\mathcal{BML}$ but not definable over C relative to the other. Now, since L is assumed to extend $\mathcal{BML}$, we know that any class EC in $\mathcal{BML}$ is EC in L as well. Hence, it is impossible for X to be definable over C relative to $\mathcal{BML}$ but not relative to L. Thus, it must be the case that X is L-definable over C, but not $\mathcal{BML}$-definable over C. This, in turn, implies there exists some L-sentence φ such that there exists no basic modal sentence ψ for which

$$X = \{(M, u) : (M, u) \models_L \varphi\} \cap C = \{(M, u) : (M, u) \models \psi\} \cap C. \qquad (5.4)$$

Using the fact that (5.4) does not hold for any basic modal ψ, we see that the hypothesis on Γ from lemma 5.2.24 is satisfied for φ if we set $\Gamma = \varnothing$. Hence, iteratively applying lemma 5.2.24 on the empty set, we obtain a chain of sets of basic modal sentences $\varnothing \subseteq \{\psi_0\} \subseteq \{\psi_0, \psi_1\} \subseteq \ldots$, for an appropriate choice of $\psi_0, \psi_1, \ldots$, every set of which preserves the inexpressibility of φ. Since L has the compactness property over C, the union of any such chain will also preserve the inexpressibility of φ. Invoking a well-known mathematical result known as *Zorn's lemma*, the closure under unions of any chain preserving the inexpressibility of φ, implies the existence of a maximal set (relative to set inclusion) of basic modal sentences preserving the inexpressibility of φ. This maximal set is readily seen to be a complete theory T for $\mathcal{BML}$.

Next, consider the set $T \cup \{\varphi\}$. We would like to show now that this set has a pointed model in C. For sake of contradiction, suppose it does not. Then for any pointed model in C satisfying T, we have that this pointed model satisfies $\varphi \leftrightarrow \bot$. But this is a contradiction, since $\bot$ is a basic modal sentence. Hence, $T \cup \{\varphi\}$ has a pointed model (M, u) in C. In the same vein, we see that $T \cup \{\neg\varphi\}$ has a pointed model (N, v) in C, since otherwise φ would be equivalent to $\top$. Clearly, (M, u) and (N, v) both satisfy the theory T. And since T is complete, (M, u) and (N, v) agree on all basic modal sentences as a result. Moreover, we also have $(M, u) \models_L \varphi$ and $(N, v) \models_L \neg\varphi$, concluding the proof.

■

The crux of the proof of theorem 5.2.23 consists of the observation that for any abstract modal logic L extending $\mathcal{BML}$ that the three properties and is non-equivalent to $\mathcal{BML}$ over C, we can transform the pointed models (M, u) and (N, v) from proposition 5.2.25 in such a manner that they retain the properties expressed in that proposition but simultaneously are made to agree on all sentences in L. The first step of this transformation consists of 'unraveling' of (M, u) and (N, v). The so-called *unraveling technique* was put to prominent use by Sahlqvist (1975) and is rather well-known among modal logicians. A description of the technique, for frames with relations of arbitrary arity, may be found in (Vos 2016, 12–3) and will not be repeated here. For our present purposes, the following two properties of the unraveling (M^U_u, u) of a pointed model (M, u) will suffice.

Proposition 5.2.26. *For any pointed model (M, u), it holds that (M, u) is bisimilar to (M^U_u, u).*

Proposition 5.2.27. *For any pointed model (M, u) it holds that the frame M^U_u is free in the class of frames.*

Understanding the second proposition requires the following definitions:

Definition 5.2.28. Let Σ be a signature and M and N be Σ-models and let $h : M \to N$ be mapping. Then h is called a *homomorphism* if we have:

- $wR^M_i w'$ implies $h(w)R^N_i h(w')$ for all $w, w' \in W^M$ and $1 \leq i \leq n$,
- $V^M(P, w) = 1$ if and only if $V^N(P, h(w))$ for all $w \in W^M$ and $P \in \Sigma$.

Definition 5.2.29. Let M, N be models and $Z \subseteq W^M \times W^N$ be a binary relation. Then Z is called a *simulation* if it satisfies the conditions for a bisimulation[17] except possibly the back condition. Furthermore, Z is said to be total if for every $w \in W^M$ there exists some $w' \in W^N$ such that wZw'.

Definition 5.2.30. Let C be class of frames. Then a model M in C is said to be *free* in C if for every model N in C and total simulation Z from M to N, there exists a homomorphism $h : M \to N$ such that $wZh(w)$ for all $w \in W^M$.

Enqvist (2013, 242) explains the intuition underlying the notion of a free model as being that the model in question has 'many' homomorphisms, in the sense that every total simulation from the model 'contains' a homomorphism.

The next step in the transformation process of the models (M, u) and (N, v) of proposition 5.2.25 will be to apply some transformation map F, satisfying a number of nice properties, to their respective unravelings. More precisely, we will be considering a transformation F on the class of all frames. Such a transformation will then naturally induce a transformation F' on the class of all models, obtained by setting $F'(M) = (F(W^M, R_1^M, \ldots, R_n^M), V)$. For convenience, let us henceforth simply write $F(M)$ to denote the model $F'(M)$. Let us consider some of 'nice' properties such a transformation F may possess.

Definition 5.2.31. Let Σ be a signature and M, N be Σ-models. Then N is staid to be an *expansion* of M if $W^M = W^N$, $V^M = V^N$ and $R_i^M \subseteq R_i^N$ for all $1 \leq i \leq n$.[18]

Definition 5.2.32. Let F be a transformation on the class of all frames. Then F is said to be *expansive* if for every model M it holds that $F(M)$ is an expansion of M. If, in addition, for every homomorphism $h : M \to N$ it holds that the mapping $h : F(M) \to F(N)$ is a homomorphism as well, we say F *preserves homomorphisms.*

Definition 5.2.33. Let F be a transformation on the class of all frames. Then F is called *idempotent* if for every model M it hold that $F(F(M)) = F(M)$.

The desirability of the above properties is made clear by the subsequent lemmas. Fist, define:

Definition 5.2.34. Let F be a transformation on the class of all frames. Then C_F denotes the class of frames invariant under F, i.e. the class of frames $(W, R_1, \ldots, R_n)$ for which $F(W, R_1, \ldots, R_n) = (W, R_1, \ldots, R_n)$.

Now, let us note proposition 5.2.26 allows us to derive the following lemma.

Lemma 5.2.35. *Let F be a transformation on the class of all frames and suppose F preserves homomorphisms. Then for any model M in C_F and node $u \in W^M$, it holds that $(F(M_u^U), u) \sim (M, u)$.*

[17] Note that bisimulations are defined for *pointed* models. However, since only the first condition in definition 5.2.9 depends on the choice of u and v, the definition is easily lifted from pointed models to models proper.

[18] Here, the term *expansion* should not be construed as the inverse of the term *reduct*, as defined for models of $\mathcal{BML}$.

Furthermore, we have:

Lemma 5.2.36. *Let F be a transformation on the class of all frames and suppose F preserves homomorphisms. Then for any pointed model (M, u) it holds that the model $F(M_u^U)$ is free in C_F.*

Proof. Let N be an arbitrary model in C_F and suppose we have total simulation Z from $F(M_u^U)$ to N. Since F is expansive, it is readily verified that Z is a total simulation from M_u^U to N as well. By proposition 5.2.27, we know that M_u^U is free in the class of all models. Hence, there exists a homomorphism $h : M_u^U \to N$ such that $wZh(w)$ for all w. Applying the fact that F preserves homomorphisms, we infer the map $h : F(M_u^U) \to F(N)$ is also a homomorphism with $wZh(w)$ for all w. Now, since N is in C_F, we know that $F(N) = N$. Thus, we see that we have a homomorphism $h : F(M_u^U) \to N$ with the desired property, concluding the proof. ■

Definition 5.2.37. Let Σ be a signature and M be a Σ-model. Let Σ^* be signature obtained by adding to Σ a fresh propositional letter P_w for each $w \in W^M$. Then M^* is the Σ^*-model obtained by setting $W^{M^*} = W^M$, $R_i^{M^*} = R_i^M$, for every $1 \leq i \leq n$, and defining V^{M^*} to be the map identical to V^M on $\Sigma \times W^M$ and letting $V^{M^*}(P_w, w') = 1$ iff $w = w'$ for all $w, w' \in W^M$.

Lemma 5.2.38. *Let L be an abstract modal logic extending $\mathcal{BML}$, F be an idempotent mapping on the class of all frames that preserves homomorphisms, (M, u) be a pointed model in C_F and (N, v) be a pointed model in C_F with the same signature as $(F(M_u^U)^*, u)$. Suppose $(F(M_u^U)^*, u)$ and (N, v) agree on all sentences in L. Furthermore, define the binary relation Z between $F(M_u^U)^*$ and N as follows. Let v be the only node in N such that uZv. For any other node u' in $F(M_u^U)^*$ we let $u'Zv'$ for any v' such that $V^N(P_{u'}, v') = 1$ and v' is finitely reachable from v.*[19] *Then, Z is total simulation from $(F(M_u^U)^*, u)$ to (N, v).*

Proof. First, observe we have trivially that uZv. Next, let u', v' be two nodes in $F(M_u^U)^*$ and N respective and let P be an a propositional letter in the signature of $F(M_u^U)^*$. Suppose that $u'Zv'$. Our goal is then to ascertain that $V^*(P, u') = 1$ iff $V^N(P, v') = 1$, where V^* denotes the valuation mapping of $F(M_u^U)^*$. Let us consider here the if-direction, the converse implication proceeding analogously. Hence, suppose that $V^*(P, u') = 1$. In case $u' = u$ and $v' = v$ it follows immediately that $V^N(P, v') = 1$ as well, since $(F(M_u^U)^*, u)$ and (N, v) agree on all L-sentences. Thus, suppose $u' \neq u$ and $v' \neq v$. Then, by definition of Z, we know that v' is a node such that $V^N(A_{u'}, v') = 1$ and v' is finitely reachable from v. From this, we know there exist some appropriate choice of $i_1, \dots, i_k$ such that $(N, v) \models_L \Diamond_{i_1} \dots \Diamond_{i_k} A_{u'}$. Consequently, we then also have $(F(M_u^U)^*, u) \models_L \Diamond_{i_1} \dots \Diamond_{i_k} A_{u'}$, which means there exists some u'' reachable from u such that $V^*(A_{u'}, u'')$. But this, in turn, implies that $u'' = u'$. So we see that u' is finitely reachable through the same sequence of binary relations

[19] To be *finitely reachable* means there exist finite sequences $R_{i_1}^N, \dots, R_{i_k}^N$ of binary relations in N and $w_1, \dots, w_k$ of nodes in N such that $vR_{i_1}^N w_1, w_1 R_{i_2}^N w_2 \dots w_k R_{i_k}^N v'$.

as v'. Now, since $V^*(P, u') = 1$ holds by hypothesis and u' is the unique node in $F(M_u^U)^*$ making true $A_{u'}$, we see that

$$(F(M_u^U)^*, u) \models_L \Box_{i_1} \ldots \Box_{i_k}(A_{u'} \to P) \tag{5.5}$$

and consequently also that

$$(N, v) \models_L \Box_{i_1} \ldots \Box_{i_k}(A_{u'} \to P). \tag{5.6}$$

From 5.6, we can infer now that $V^N(P, v') = 1$, as desired.

To establish the forth condition, let us u_0, u_1, v_0 be nodes such that $u_0 Z v_0$ and $u_0 R_i^* u_1$, for some binary relation R_i^* of $F(M_u^U)^*$, and assume u_0 and v_0 are finitely reachable by binary relations labeled by $i_1, \ldots, i_k$. Our task is then to prove the existence of some node v_1 such that $v_0 R_i^N v_1$ and $u_1 Z v_1$. This can be accomplished swiftly by noting that

$$(F(M_u^U)^*, u) \models_L \Box_{i_1} \ldots \Box_{i_k}(A_{u_0} \to \Diamond_i A_{u_1}) \tag{5.7}$$

and hence that

$$(N, v) \models_L \Box_{i_1} \ldots \Box_{i_k}(A_{u_0} \to \Diamond_i A_{u_1}). \tag{5.8}$$

Thus, we see that for every node v' in N reachable by a path labeled by $i_1, \ldots, i_k$ for which $V^N(A_{u_0}, v') = 1$ there exists a nodev'' such that $v' R_i^N v''$. So, in particular, there exists such a node v_1 for v_0, establishing the forth condition.

The proof that Z is total relies on the observation that M_u^U is an unraveling of the underlying model M. In brief: in unraveling a model we, among other things, delete any nodes of the original model that are not finite reachable. And since the mapping F is expansive and the operation $*$ only affects the valuation mapping, the model $F(M_u^U)^*$ will contain no unreachable nodes. Thus, for any node u' in $F(M_u^U)^*$ we find a path to u' from u by a sequence of binary relations with labels $i_1, \ldots, i_k$ and accordingly

$$(F(M_u^U)^*, u) \models_L \Diamond_{i_1} \ldots \Diamond_{i_k} A_{u'}, \tag{5.9}$$

from which it follows that

$$(N, v) \models_L \Diamond_{i_1} \ldots \Diamond_{i_k} A_{u'}. \tag{5.10}$$

It thus holds that $u' Z v'$ for some v' in N. We conclude Z is indeed a total simulation from $(F(M_u^U)^*, u)$ to (N, v).

■

For the next lemma, we require an extension of the concept of elementary embedding to the case of abstract modal logics.

Definition 5.2.39. Let L be an abstract modal logic and $(M, u), (N, v)$ be pointed models. Then a function $f : (M, u) \to (N, v)$ is called an *L-elementary embedding* if f is an injective homomorphism such that

- $f(u) = v$,

- for all nodes u_0, u_1 it holds that $f(u_0)R_i^N f(u_1)$ implies $u_0 R_i^M u_1$,
- for every node u' the pointed models (M, u') and $(N, f(u'))$ agree on all sentences in L.

Lemma 5.2.40. *Let L be an abstract modal logic extending $\mathcal{BML}$, F be an idempotent mapping on the class of all frames that preserves homomorphisms, (M, u) be a pointed model in C_F and (N, v) be a pointed model in C_F with the same signature as $(F(M_u^U)^*, u)$. Suppose $(F(M_u^U)^*, u)$ and (N, v) agree on all sentences in L. Then there exists an L-elementary embedding*

$$f : (F(M_u^U)^*, u) \to (N, v).$$

Proof. Let Z be the total simulation as defined in lemma 5.2.38. We will show how to obtain an L-elementary embedding f from Z. Now, it is straightforwardly verified that Z is also a total simulation from $F(M_u^U)^*$ to the reduct N' of N to the signature of $F(M_u^U)$. By lemma 5.2.36,we have that $F(M_u^U)$ is free in C_F. Hence, there exists a homomorphism f from $F(M_u^U)$ to N' such that $u'Zf(u')$ for any node u' in $F(M_u^U)$.

By an argument similar to those in the proof of lemma 5.2.38, it now follows for every u in $F(M_u^U)^*$ that $(F(M_u^U)^*, u')$ and $(N, f(u'))$ agree on all L-sentences. This observation, in turn, can be used to ascertain the other properties of elementary embeddings. First, note that f is injective. For if $u_0 \neq u_1$ then $(F(M_u^U)^*, u_0) \models_L A_{u_0}$ and $(F(M_u^U)^*, u_1) \not\models_L A_{u_0}$ and thus also $(N, f(u_0)) \models_L A_{u_0}$ and $(N, f(u_1)) \models_L A_{u_0}$. Consequently, we see that $f(u_0) \neq f(u_1)$. All that is remains is to verify that $f(u_0)R_i^N f(u_1)$ implies $u_0 R_i^* u_1$ for every $1 \leq i \leq n$. Hence, suppose that $f(u_0)R_i^N f(u_1)$. Then we know $(N, f(u_0)) \models_L \Diamond_i A_{u_1}$ and thus $(F(M_u^U)^*, u_0) \models_L \Diamond_i A_{u_1}$. So there exists some node u' such that we have $u_0 R_i^* u'$ and $(F(M_u^U)^*, u') \models_L A_{u_1}$. Now, the latter fact implies that $u' = u_1$ and hence $u_0 R_i^* u_1$. Thus, we see that f is indeed an L-elementary embedding. ■

Finally, we require:

Definition 5.2.41. Let T be a theory in basic modal logic, M be a model and w a node in M. Then, for every $1 \leq i \leq n$, the theory T is said to be *satisfiable among R_i^M-successors of* w, if there exists a node w' in M such that $wR_i^M w'$ and $(M, w') \models T$. In a similar vein, the theory T is called *finitely satisfiable among R_i^M-successors* of w if every finite subset of T is satisfiable among R_i^M-successors of w.

Definition 5.2.42. Let M be some model. Then we say M is *modally saturated* if for every node w in M, every basic modal theory T and every $1 \leq i \leq n$ we have: if T is finitely satisfiable among R_i^M-successors of w then T is satisfiable among R_i^M-successors of w.

It may now noted that the class of modally saturated models has the so-called *Hennessy-Milner property.* That is:

Proposition 5.2.43. *Let (M, u) and (N, v) be pointed models. Suppose (M, u) and (N, v) are modally saturated and agree on all basic modal sentences. Then it holds that $(M, u) \sim (N, v)$.*

We are now ready to consider the crucial lemma for the proof of theorem 5.2.23.

Lemma 5.2.44. *Let L be an abstract modal logic and F be a transformation on the class of all frames. Suppose that the following holds:*

- *L has the compactness property, the Tarski union property and is bisimulation invariant over C_F.*
- *F is idempotent and preserves homomorphisms,*
- *C_F is closed under union of elementary chains relative to L.*

Then L is equivalent to $\mathcal{BML}$ over C_F.

Proof. Let Σ be a signature and let $(M, u), (N, v)$ be two arbitrary pointed Σ-models in C_F. By proposition 5.2.25, it is sufficient to show that if (M, u) and (N, v) agree on all basic modal sentences then they agree on all L-sentences as well. For this, in turn, it suffices to prove that there exist modally saturated models (M^s, u) and (N^s, u) which agree on all L-sentences with (M, u) and (N, v) respectively: the result then follows from proposition 5.2.43 and the observation that L is bisimulation invariant relative to C_F. We proceed by constructing modally saturated L-elementary extensions of $F(M_u^U)$ and $F(N_v^U)$ in C_F. This will conclude the proof, since we know by lemma 5.2.35 that $(M, u) \sim (F(M_u^U), u)$ and $(N, v) \sim (F(N_v^U), v)$. Below, I sketch how to obtain such an extension for $F(M_u^U)$, with the construction for $F(N_v^U)$ proceeding in a completely analogous manner.

Let T denote the theory of $(F(M_u^U)^*, u)$ and denote by $\underline{\text{FinSat}}(w, i)$ the set of all basic modal theories that are finitely satisfiable among the R_i^*-successors of w. Next, we define for every $1 \leq i \leq n$, node w in $(F(M_u^U)^*$ and Φ in $\underline{\text{FinSat}}(w, i)$ a propositional letter $P_{\Phi,i}^w$ and let $\Gamma_{\Phi,i}^w$ denote the set consisting of all sentences either of the form

$$\Box_{i_1} \dots \Box_{i_k} (A_w \to \Diamond_i P_{\Phi,i}^w) \tag{5.11}$$

or of the form

$$\Box_{i_1} \dots \Box_{i_k} (P_{\Phi,i}^w \to \varphi), \tag{5.12}$$

with $i_1, \dots, i_k \in \{1, \dots, n\}$ and $\varphi \in \Phi$. Now, consider the theory

$$T' := T \cup \{\Gamma_{\Phi,i}^w : w \in F(M_u^U),\ 1 \leq i \leq n,\ \Phi \in \underline{\text{FinSat}}(w, i)\}^{\text{cl}}. \tag{5.13}$$

By extending the valuation mapping $(F(M_u^U)^*, u)$ to the propositional letters $P_{\Phi,i}^w$ in an appropriate manner, we can find a model for any finite subset of the theory T'. Using the fact that basic modal logic satisfies the compactness property, we thus obtain a model (M', u') of the entire theory T'. By the bisimulation invariance of basic modal logic, we may assume this model to be unraveled. From the idempotence of F, it follows that $F(M')$ is a model in C_F. Moreover, by lemma 5.2.40, we know that there exists an L-elementary embedding of $(F(M_u^U, u)$ into $(F(M'), u')$. Hence, we see that $F(M')$ is an L-elementary extension of $F(M_u^U)$.

Applying the same procedure as above now to $F(M')$ to construct an L-elementary extension $F(M'')$ of $F(M')$ and continuing in this manner ad infinitum, we obtain an L-elementary chain

$$F(M_u^U) \preceq_L M' \preceq_L M'' \preceq_L \dots \tag{5.14}$$

Let M^∞ denote the union of this chain. Now, we know that M^∞ is a model in C_F, since C_F is assumed to be closed under unions of L-elementary chains. Moreover, since L has the Tarski union property over C_F, it follows that M^∞ is an L-elementary extension. Finally, although we do not go into the details here, it can be checked that M^∞ is modally saturated. Hence, we see that M^∞ is the desired extension of $F(M_u^U)$, concluding the proof. ■

To now arrive at theorem 5.2.23 from lemma 5.2.44, all that is left is to relate transformations F to classes of frames definable by means of strict universal Horn sentences in some meaningful manner. This is accomplished as follows.

Definition 5.2.45. Let M be some model and let φ be a strict universal Horn sentence of the form $\forall\vec{x}[\alpha(\vec{x}) \to yR_iz]$, where y, z are included in the tuple $\vec{x}$ and R_i is the relation symbol corresponding to R_i^M, and let $\vec{a}$ be a tuple of elements in M in which we denote the elements corresponding to to the variables y, z as v, w respectively. Then we say $\vec{a}$ *forces* the the pair (v, w) with respect to the operator $\Diamond_j$ and formula φ if $i = j$ and $M \models \alpha[\vec{a}]$.[20] By extension, we say for an arbitrary set of strict universal Horn sentences Γ that $\vec{a}$ *forces* (v, w) with respect to $\Diamond_j$ and Γ if for some $\psi \in \Gamma$ we have that $\vec{a}$ forces (v, w) with respect to $\Diamond_j$ and ψ.

Definition 5.2.46. Let Γ be some set of strict universal Horn sentences. Then we define F_Γ to be the transformation[21] sending each model M to the model $F_\Gamma(M)$ with identical domain W^M and valuation mapping V^M and, for every $1 \leq i \leq n$, its i-th binary relation R_i^Γ defined by

$$R_i^M \cup \{(v, w) : \vec{a} \text{ forces } (v, w) \text{ w.r.t. } \Diamond_i \text{ and } \Gamma \text{ for some } \vec{a} \text{ in } M\}. \tag{5.15}$$

Intuitively, we can view Γ as prescribing a number of conditions and the transformation F_Γ as adding to any model M all additional links between nodes which satisfy one of these conditions. Naturally, the addition of more links by the application might result in more of the conditions in Γ being met by the obtained model. Once more applying the transformation might then lead to a new model containing even more links between nodes than its predecessor. Building on this observation, we have:

[20] Here, we write '$\models$' for the satisfaction relation in $L_{\omega\omega}$.

[21] As before, we do not distinguish notionally between a transformation on the class of all models and the corresponding transformation on the class of all underlying frames.

Definition 5.2.47. Let Γ be a set of strict universal Horn sentences. Denote by F^n_Γ the n-th iteration of the transformation F_Γ. Then, for any model M, define

$$F^\omega_\Gamma = \bigcup_{n\in\mathbb{N}} F^n_\Gamma(M). \tag{5.16}$$

The significance of this construction is now displayed in the following lemma:

Lemma 5.2.48. *Let Γ be a set of strict universal Horn sentences. Then a model M satisfies Γ if and only if M is invariant under F^ω_Γ.*

Proof. For the if-direction, note that it can be facilely verified that any model M such that $M \models \Gamma$ will be invariant under F_Γ and hence under F^ω_Γ as well. For the converse implication, suppose M is invariant under F^ω_Γ. Now, if there were to exist some $\forall\vec{x}[\alpha(\vec{x}) \to yR_iz] \in \Gamma$ not satisfied by M then we would have a sequence $\vec{a}$ such that $M \models \alpha[\vec{a}]$ and $M \not\models yR_iz[\vec{a}]$. Then, since $\vec{a}$ forces the pair (v, w) corresponding to y and z, we see that $vR^\Gamma_i w$ but not $vR^M_i w$. That is, $F_\Gamma(M) \neq M$. Since F_Γ is expansive, it then also follows that $F^\omega_\Gamma(M) \neq M$. This, however, contradicts our assumption that M is invariant under F^ω_Γ. ■

Theorem 5.2.23 is now readily obtained from the following lemmas:

Lemma 5.2.49. *For any set of strict universal Horn sentences Γ, the transformation F^ω_Γ is idempotent and preserves homomorphisms.*

Lemma 5.2.50. *Let L be an abstract modal logic and C be a class of frames definable by a set of strict universal Horn sentences. Then C is closed under unions of elementary chains relative to L.*

Reflecting on theorem 5.2.23, we see once again the potential considering logics for non-standard classes of models. As noted in subsection 3.1.2, one of the two significant ways in which the abstract maximality theorem for abstract logics generalizes typical Lindström theorems is that is provides us with a characterization result for logics formulated over different classes of models than the usual class of first-order models. In the same vein, Enqvist's general Lindström theorem shows us there is also much to be gained in the study of abstract modal logics by considering certain specialized classes of frames. What value does this hold for the metascientific enterprise and, indeed, the creation of a SWLA? We will return to this point in section 5.3. Before this, however, there is another recent framework within abstract model theory that awaits our attention.

5.2.2 Institutional Model Theory

There is no doubt that the framework of semi-abstract model theory provides us with a valuable methodology for the analysis of logical systems from a general point of view. The notion of *abstract logic*, as well as the more recent notion of *abstract modal logic*, can be seen to provide us with generalizations of the notions of *sentence*, *model* and *truth*. Indeed, almost all of the usual components of first-order logic/basic modal logic find a generalization in the realm of semi-abstract model theory. In one aspect, however, abstract (modal) logics remain yoked to standard logical practice, viz. through their reliance on the usual notion of **signature**. For example, we may note in definition 3.1.3, the map L is defined on the set of all *first-order* signatures. Similarly, we see in definition 5.2.8 that the otherwise abstract maps L^1, L^2 take as domain the set of all propositional signatures. Hence, we might observe that the level of generality offered by semi-abstract model theory is still one step short of a truly general approach to logic. In fact, it is for this reason that I have followed Diaconescu (2008, 3) in referring to the framework covered in section 3.1 as *half-* or *semi-abstract* model theory.

To fill the void of a **fully abstract** abstract approach to the model-theoretic study of logics, a select group of researchers have taken it upon themselves to develop the framework of institutional model theory. Within this framework, the informal concept of *a logic* is formalized by the formal notion of an *institution.*

Definition 5.2.51. An *institution* is a tuple $I = (\mathrm{Sig}^I, \mathrm{Sen}^I, \mathrm{Mod}^I, \{\models^I_\sigma\}_{\sigma\in\mathrm{Sig}^I})$, where

- Sig^I is a category, the objects of which are called *signatures*,
- $\mathrm{Sen}^I : \mathrm{Sig}^I \to \mathrm{Set}$ is a functor, giving for each signature Σ a set whose elements are called *sentences* over Σ,
- $\mathrm{Mod}^I : (\mathrm{Sig}^I)^{\mathrm{op}} \to \mathrm{Cat}$ is a functor, giving for each signature Σ a category whose objects are called Σ-*models* and whose arrows are called Σ-*(model) homomorphisms*,
- for each signature Σ, $\models^I_\Sigma$ is a subset of $|\mathrm{Mod}^I(\Sigma)| \times \mathrm{Sen}^I(\Sigma)$, called Σ-*satisfaction*, such that for each morphism $\rho : \Sigma \to \Sigma'$ in Sig^I the *satisfaction condition*

  $$M' \models^I_{\Sigma'} \mathrm{Sen}^I(\rho)(\varphi) \text{ if and only if } \mathrm{Mod}^I(\rho)(M') \models^I_\Sigma \varphi$$

  holds for each $M' \in |\mathrm{Mod}^I(\Sigma')|$ and $\varphi \in \mathrm{Sen}^I(\Sigma)$.

As is evident from the above definition, a fundamental aspect of the methodology employed in institutional model theory is the usage of category-theoretic notions.[22] Institutional model theory finds its origins in computer science, being first developed by computer scientists Joseph Goguen and Rod Burstall. Diaconescu (2008, 2) provides with an apt description of the motivation underlying the development of the framework:

> "The notion of institution was introduced by Goguen and Burstall in the late 1970s . . . in response to the population explosion of specification logics with the original intention of providing a proper abstract framework for specification of, and reasoning about, software systems."

[22] It should be noted that semi-abstract model theory has, in fact, also received a categorical formulation at the hands of Barwise (1974). In this context, however, category theory is employed only at superficial level as a convenient language for showcasing the most important definitions. As evidenced by (Barwise & Feferman 1985), we can just as well do semi-abstract model theory in entirely non-categorical terms.

Chronologically, this would place the genesis of institution model theory in the same period of time as that of the FWLA. It should be noted, however, that major interest and developments in the institutional model theory did not seem to take off until the publication of (Burstall & Goguen 1992). The most comprehensive treatment of the field, including some of its more recent developments, is given by Diaconescu (2008). For a primer on the aims and methods employed in the study of institutions, see (Diaconescu, Mossakowski & Tarlecki 2014). It is from these latter two sources that the material of the current subsection has been drawn.

To illustrate the potential for abstraction innate in the institutional approach to abstract model theory, let us consider here how one particularly well-known model-theoretic property of first-order logic can be generalized to the setting of institutions. More specifically, recall from section 3.2 the *interpolation property* as defined for abstract logics. For the case of single-sorted first-order logic, the interpolation property can be restated in the following, better known format:

Theorem 5.2.52. *Let Σ_1, Σ_2 be signatures and let φ, ψ be sentences over Σ_1, Σ_2 respectively. Suppose that for the $(\Sigma_1 \cup \Sigma_2)$-sentence $\varphi \to \psi$ we have $\models \varphi \to \psi$. Then there exists a $(\Sigma_1 \cap \Sigma_2)$-sentence χ such that $\models \varphi \to \chi$ and $\models \chi \to \psi$.*

This result is widely known as the *Craig interpolation theorem*, named for its progenitor. Intuitively, it expresses the fact that the only non-logical symbols relevant to the derivation of ψ from φ are those non-logical symbols occurring in both φ and ψ.

Now, suppose we are interested in lifting theorem 5.2.52 to the institutional setting. One problem immediately presents itself: with 'signatures' now being objects of *any* arbitrary category Sig^I, it is not obvious what the generalized counterparts of the objects $\Sigma_1 \cup \Sigma_2$ and $\Sigma_1 \cap \Sigma_2$ ought to be. It turns out, however, that we can readily generalize these set-theoretic constructs by, in typical category-theoretic fashion, formulating their properties in terms of morphisms. To this end, let us first note that the relation between the different signatures in theorem 5.2.52 can be represented pictorially as follows:

$$\begin{tikzcd}
\Sigma_1 \cap \Sigma_2 \arrow[r, "\rho_1"] \arrow[d, "\rho_2"'] & \Sigma_1 \arrow[d, "\zeta_1"] \\
\Sigma_2 \arrow[r, "\zeta_2"'] & \Sigma_1 \cup \Sigma_2
\end{tikzcd} \tag{5.17}$$

Here, $\rho_1, \rho_2, \zeta_1, \zeta_2$ denote the obvious inclusion maps. Now, what can we say about the objects $\Sigma_1 \cup \Sigma_2$ and $\Sigma_1 \cap \Sigma_2$ in terms of the above diagram? By the nature of unions, we find that if we take any other signature Σ such that $\Sigma_1, \Sigma_2 \subseteq \Sigma$, with inclusion maps $i_1 : \Sigma_1 \to \Sigma$ and $i_2 : \Sigma_2 \to \Sigma$, then there must also exist a corresponding inclusion map $\zeta : \Sigma_1 \cup \Sigma_2 \to \Sigma$. Moreover, for this and only this inclusion map ζ, we see that we have the identities $i_1 = \zeta \circ \zeta_1$ and $i_2 = \zeta \circ \zeta_2$. In this sense, the signature Σ with maps i_1, i_2 can be said to 'factor' through the signature $\Sigma_1 \cup \Sigma_2$.

It turns out that the square (5.17) has almost all of the properties of a so-called *pullback square*, as defined in appendix B. In fact, (5.17) is an instance of what is known as a *pushout square*, the dual construction of a pullback square. Moving to an arbitrary institution I, we see that in the hypothesis of theorem 5.2.52, we can replace the signatures $\Sigma_1, \Sigma_2, \Sigma_1 \cap \Sigma_2, \Sigma_1 \cup \Sigma_2$ by any four objects $\Sigma_1, \Sigma_2, \Sigma, \Sigma'$ in Sig^I such that the commutative diagram

$$\begin{tikzcd}
\Sigma' \arrow[r, "\rho_1"] \arrow[d, "\rho_2"'] & \Sigma_1 \arrow[d, "\zeta_1"] \\
\Sigma_2 \arrow[r, "\zeta_2"'] & \Sigma
\end{tikzcd} \tag{5.18}$$

is a pushout square.

An obvious generalization of the interpolation property to arbitrary institutions now presents itself. That is, we might venture to say that an institution I has the *interpolation property* if for any pushout square, as given by (5.18), we have: for any sentences φ, ψ, if $\mathrm{Sen}^I(\zeta_1)(\varphi) \models_\Sigma \mathrm{Sen}^I(\zeta_2)(\psi)$ then there exists some χ in $\mathrm{Sen}^I(\Sigma')$ such that $\varphi \models_{\Sigma_1} \mathrm{Sen}^I(\rho_1)(\chi)$ and $\mathrm{Sen}^I(\rho_2)(\chi) \models_{\Sigma_2} \psi$. This proposal, however, turns out to be overly restrictive. For example, the system of logic known as *equational logic*[23], and hence also its corresponding institution, would fail the criteria of the interpolation as proposed here. Yet, there is a distinct sense in which equational logic *does* exhibit interpolation. Namely, by replacing in the above proposal the sentence χ by a *finite set* of sentences Γ, equational logic would in fact meet the requirements of the interpolation property. The crux of matter now is that while for first-order logic, as well as many other traditional systems of logic, the conjunction property guarantees that we can identify single sentences with finite sets of sentences, this does not necessarily hold true in a generalized setting. Thus, by adhering to the aforementioned proposal for the generalized interpolation property, without noticing the implicit assumptions based on first-order logic still present within it, we run the risk of putting forth an unduly restrictive generalization.[24]

To ensure we achieve a proper level of generality, let us replace the sentences φ, ψ, χ in theorem 5.2.52 by arbitrary sets of sentences and introduce:

Definition 5.2.53. Let I be an institution, $\Sigma_1, \Sigma_2, \Sigma, \Sigma'$ be objects in Sig^I and

[23] *Equational logic*, as the term is used here, refers to the logic obtained from first-order logic by restricting the class of signatures to only those signatures containing no relation symbols, making '=' the only relation symbol in the language, and furthermore restricting the syntax by only allowing sentences having the form of a universally quantified equation. Notably, this logic does not have the conjunction property, as defined for abstract logics, since the conjunction of two universally quantified equations is not necessarily equivalent to a universally quantified equation.

[24] Incidentally, we might note that this presents us with a superb example of the debilitating influence of first-order fixation within the field of logic itself.

$\rho_1, \rho_2, \zeta_1, \zeta_2$ be morphism in Sig^I. Suppose

$$\begin{tikzcd}
\Sigma' \arrow[r, "\rho_1"] \arrow[d, "\rho_2"'] & \Sigma_1 \arrow[d, "\zeta_1"] \\
\Sigma_2 \arrow[r, "\zeta_2"'] & \Sigma
\end{tikzcd} \tag{5.19}$$

is a commutative diagram. Then this diagram is said to be an *interpolation square* if for all sets of sentences Γ_1, Γ_2 we have: if

$$\mathrm{Sen}^I(\zeta_1)(\Gamma_1) \models_\Sigma \mathrm{Sen}^I(\zeta_2)(\Gamma_2) \tag{5.20}$$

then there exists some set Γ of sentences in $\mathrm{Sen}^I(\Sigma')$ such that

$$\Gamma_1 \models_{\Sigma_1} \mathrm{Sen}^I(\rho_1)(\Gamma) \text{ and } \mathrm{Sen}^I(\rho_2)(\Gamma) \models_{\Sigma_2} \Gamma_2, \tag{5.21}$$

where expressions of the form $\mathrm{Sen}^I(\rho_1)(\Gamma)$ denote the set of sentences resulting from applying the morphism ρ_1 to every sentence in Γ.

This now suggests the following definition:

Definition 5.2.54. Let I be an institution. Then I is said to have the *interpolation property* if it holds that every pushout square of signatures is also an interpolation square.

The above definition goes a long way in providing us with a proper generalization of the interpolation property for arbitrary institutions. Yet, it still proves to be overly restrictive for a number of significant cases. In particular, many-sorted first order logic only satisfies the conditions for interpolation if we allow for the additional relaxation that in definition 5.2.54 we only consider those pushout squares in which either the morphism ρ_1 or ρ_2 is an injection when restricted to the sort symbols in Σ'. This gives rise to the final liberalization of the notion of interpolation within the context of institutions:

Definition 5.2.55. Let I be an institution and $\mathcal{L}, \mathcal{R}$ be classes of signature morphisms. Then I is said to have the $(\mathcal{L}, \mathcal{R})$-*interpolation property* if it holds that every pushout square of signatures such that $\rho_1 \in \mathcal{L}$ and $\rho_2 \in \mathcal{R}$ is also an interpolation square.

Let us conclude here this brief exposition of the theory of institutions. It is clear that institution-independent model theory presents with avenues of abstraction unparalleled by even the more abstract regions of semi-abstract model theory. To contrast the two formalisms, we can thus refer to the institutional framework as being a *fully* abstract model theory. The crucial point to keep in mind is that by generalizing the notion of signature we are longer bound to consider only signatures consisting of meaningless, syntactic objects. This, in turn, greatly enhances the potential of logic as a tool of formalization. Hence, any new attempt at developing a framework for logical abstractivism should undoubtedly place the study of institutions at the center of its attention. We shall return to this observation in the subsequent section, in which a new approach to logical abstractivism will be given shape.

5.3 Making a Second Wave

In this section, I bring together all our prior considerations and outline a new way of combining abstract model theory with the field of metascience, which is to form the basis of a second wave of logical abstractivism. Similarly to how the first wave could be typified by means of four characteristic features, so too will I present a general program for model-theoretic metascience based around four alternative central tenets. The most significant aspect of this new program consists of moving away from logical liberalization as our central methodological principle and adopting a new modus operandi in its stead. Sketching this general program will be the focus of subsection 5.3.1. Following this, we will briefly return to the ruminations set out in the preceding chapter in subsection 5.3.2. That is, we will consider how well and to what extent the newly proposed program for logical abstractivism is compatible with the categorical approach to metascience. Finally, I take up the question of what a second wave of logical abstractivism might hope to achieve within a more general metascientific context. To this end, I introduce in subsection 5.3.3 a number of problems pertaining to the contemporary philosophy-of-science literature and argue that the SWLA puts us in a favorable position to resolve these issues.

5.3.1 A General Program

Throughout the preceding text, I have foreshadowed on numerous occasions the manner in which I conceive a more successful flavor of logical abstractivism may be given shape. Expounding these arguments now in systematic fashion, let us consider what such a second wave of logical abstractivism would look like on the whole. To showcase the differences between the FWLA and the prospective SWLA, let us take up each of the former's central tenets in turn and consider how it relates to a corresponding feature of the latter. Proceeding in this way, we find counterparts (i')–(iv') of the characteristics (i)–(iv) of the Pearce-Rantala approach, which then provide us with a programmatic outline of the novel framework.

We commence by reconsidering characteristic feature (i) of the FWLA, i.e.

(i) The usage of semi-abstract model theory as formalism of choice.

As noted in section 3.2.3, semi-abstract model theory represented the only substantial model-theoretic approach to the abstract study of logics at the time Pearce and Rantala first developed their account. Nowadays, however, we have available to us a somewhat wider array of model-theoretic frameworks, making an a priori adherence to point (i) less forgivable.

Which alternatives now suggest themselves? A glance at the preceding subsection may make the answer to this query seem of a trivial nature. We must take note, however, that adherence to logical abstractivism does *not* entail a commitment to the usage of *model-theoretic methods* per se. Rather, logical abstractivism espouses the notion that the abstract study of logics *in general* is of value to the study of metascience. This leaves open the question of whether such frameworks are of a model-theoretic, proof-theoretic, algebraic or other kind of nature. Recalling the fact that universal logic serves as an umbrella term for abstract approaches to logic, logical abstractivism thus boils down to the conviction that *universal logic*, not *abstract model theory* exclusively,[25] is of value to metascience.

There is thus no a priori imperative which states that a new approach to logical abstractivism should employ a framework of a model-theoretic nature. In section 3.2.3, the choice to expound logical abstractivism in a model-theoretic setting was defended by noting that other abstract frameworks at the time usually only allowed us to study the features of propositional logics, thus being too impoverished to be of much use in the formal analysis of science. To what extent does this defense still hold up today? Without entering into too much details, it may be noted that the field of *algebraic logic* has started to see an increasing number of metascientific applications.[26]

It is not the case that I believe a model-theoretic approach to be *inherently* superior to an algebraically inspired approach. A comprehensive argument for one approach over the other would require a deeper investigation of both frameworks and will not be taken up here. I do hold, however, that there is good reason to believe that the new model-theoretic approach as outlined in this subsection can be combined nicely with at least some aspects of algebraic logic. Incidentally, this would constitute an additional advantage the SWLA would have over the FWLA. We shall return to this point below. For now, let us simply accept that abstract model theory provides us with the desirable backdrop for logical abstractivism. The revised version of characteristic (i) is then evident. Namely, we should adopt a methodology based around

> (i') The usage of *fully* abstract model theory as formalism of choice.

More precisely, I use here the designation *fully abstract* to refer to the frameworks of

- semi-abstract model theory, where we allow our abstract logics to have a class of models Mod_L different from the usual class of first-order models;
- semi-abstract model theory, where we expand its definition to include the study of abstract modal logics as well as the study of abstract logics;
- institution-independent model theory.

[25] To make it absolutely clear, let us note once more that, just as I employ *universal logic* as an umbrella term for any framework for the abstract study of logics, I use *abstract model theory* as an umbrella term for any model-theoretic approach to universal logic. Consequently, both semi-abstract model theory and institution-independent model theory are considered special cases of abstract model theory. This is contrary to common usage, in which *abstract model theory* is identified with *semi-abstract model theory*.

[26] Cf. (Dewar 2017). It may also be noted that the categorical view of scientific theories, as observed in chapter 4, has some significant connections to algebraic logic through its reliance on notions as the syntactic category and the field of topos theory.

Let us leave open here the exact manner in which these three are to be combined. If desired, we could view semi-abstract model theory as being a special case of institutional model theory, though such a reduction would fail to do justice to the different methodologies underlying each framework. At any rate, we see that tenet (i') provides us with a strictly more general formal toolbox than the one employed by Pearce and Rantala.

With this increase in generality, we can now also reconsider central characteristic (ii), which prescribes:

(ii) A focus on extensions of first-order logic.

While semi-abstract model theory already provided us with the tools to analyze the properties of non-standard logics, i.e. logics that are not some extension of first-order logic such as probabilistic logic or logics for topological structures, the inclusion of institutional model theory increases our ability to include such non-standard logics into our framework exponentially. After all, we only need to take any category Sig^I different from a conventional category of signatures to arrive at an instance of non-standard logic. Thus, we are free to put forward an alternative tenet for the SWLA, namely:

(ii') A focus on non-standard logics.

From such a myriad of possibilities, selecting the most suitable type of logic to employ in the analysis of science is no straightforward task and will require careful considerations and scrutiny. Below, I make some tentative suggestions as to the general direction in which we should search for an appropriate type of logic for metascientific applications.

We are now well-poised to consider the revision of the most significant aspect of the FWLA:

(iii) A focus on logical liberalization.

Recall, now, the dilemma we faced when trying to dispel this tenet from our list of necessary conditions for logical abstractivism. "If we have a logical system of adequate expressive power and nice model-theoretic properties, we use this logic for our metascience. If we do not have such a logic, we are left with no option but to purse logical liberalization." As noted in subsection 5.1.1 above, the dilemma dissolves when we let go our traditional view of logic in the analysis of science. If taken for granted that the only use of logic in the study of metascience is to codify in a formal language the statements of some scientific theory, such as quantum mechanics, then it is implicit that we evaluate the usefulness of logics by their expressive capabilities. However, once we free ourselves from the restrictions placed on us by tenet (ii) of the FWLA and with it the traditional view on the role of logic in metascience, an easy way out of the dilemma presents itself: use a logic which does not have adequate expressive power or necessarily has nice model-theoretic properties. Take, for instance, a fragment of first-order logic, e.g. basic modal logic, and determine how we may use it to formalize certain aspects of science.[27]

This strategy might puzzle the traditionalist: for how could we possibly hope to encode even the simplest of scientific statements in a logic that is expressively weaker than first-order logic? The answer lies in the observation that we are no longer interested solely in encoding statements from a particular scientific theory in the language of the logic under consideration. Instead, we use the *structure* of the logical system to formalize the *structure* of some corresponding scientific theory. Of course, there are many different scientific theories, which means that our metascience would involve a multitude of logical systems as well. Relations between different theories can be formulated as relations between the corresponding systems of logic. This is, of course, extremely reminiscent of the modus operandi encountered in abstract model theory and, indeed, universal logic as a whole. This methodology, which I refer to as **logical pluralization**, thus represents an approach to metascience that is much closer to the spirit of abstract model theory than the logical liberalization employed in the FWLA.

As an added bonus, the method of logical pluralization enables us to present the traditionalist with a more nuanced answer to their above inquiry. For if we decide to use a plurality of logics with which to model different areas of science then, in particular, we can use logics with non-standard classes of models. Now, by defining a logic with respect to some fixed class of models, thereby removing any unwanted or meaningless models from our discourse, it is quite conceivable that we can positively impact the expressive power of the underlying system of logic. The usefulness of Enqvist's theorem for abstract modal logics relativized to particular classes of frames now also comes to the fore, since it demonstrates how we can we restrict a well-known system of logic to a more limited class of models while retaining a number of powerful model-theoretic properties.[28]

Summing up the above considerations, we can discard tenet (iii) of the FWLA in favor of

(iii') A focus on logical pluralization.

Finally, consider the fourth tenet of the Pearce-Rantala approach. That is:

(iv) A focus on improving the structuralist approach to scientific theories.

Now, while I do not believe that (iv) by itself is necessarily incompatible with tenets (i')–(iii') established above, I do hold that another approach to scientific theory-structure, viz. the state-space approach, connects to these revisions in a much more natural fashion. This is based on the straightforward observation

[27] The connection between tenet (ii) and the traditional view on the role of logic in metascience now also becomes more apparent. While tenet (ii) strictly speaking does not force us to adhere to the traditional view, a focus on extensions of first-order logic still brings with it implicitly a conception that it is expressive power that is the most important feature a logic can possess for the metascientific enterprise.

[28] We shall presently return to this observation in explicating the fourth central tenet of the SWLA.

that the state-space approach includes sentences (in the form of elementary statements)[29] as well as models interpreting these sentences (in the form of state spaces) and hence already has all of the fundamental ingredients required for it to be linked with the model-theoretic study of logics. Moreover, as witnessed in section 2.3, the usage of state spaces naturally gives rise to a certain *modal-logical* structure in the formalization of scientific theories.

An extremely natural and elegant way of combining the state-space approach with the study of abstract modal logics now presents itself. Consider an abstract modal logic L identical to basic modal logic except in that we restrict its class of frames to consist of only state spaces. In a very definite sense, the logic L would then represent the 'logic of classical physics'. Investigating the properties of L would then simultaneously reveal to us corresponding properties of the knowledge-system of classical physics itself. If we are feeling particularly bold, we might even attempt to formulate a Lindström theorem for this logic L, providing us with nothing less than a model-theoretic characterization of classical physics. This is but one of many possible examples. Replacing the state spaces of classical physics by their quantum counterparts, i.e. *Hilbert spaces*, we might venture to characterize *quantum physics*. Or, if we wish to consider disciplines outside of physics, we might turn to Llyod (1994) and Thompson (1989) and see how we can apply the state-space approach to the field of evolutionary biology.

Enqvist's (2013) treatment of abstract modal logics now becomes of great interest. It shows us how to obtain a Lindström-style characterization theorem for basic modal logic relative to particular classes of frames. The question of whether we could formulate a similar theorem when taking, instead of frame classes definable by Horn sentences, frame classes consisting of state spaces now thus very naturally presents itself. Of course, it is by no means certain that the methods employed in the proof of Enqvist's theorem would be transferable to a characterization result relative to classes of state spaces. The theorem, however, does provide us with a valuable starting point in the search for metascientifically relevant characterization results by showcasing the general methodology involved in relativizing abstract modal logics to specific frame classes.

An additional point in favor of the state-space approach is that it provides us with a natural interface between model-theoretic and algebraic approaches to logical metascience. Without entering into details, let us note that algebraic logic can be seen as the study of logics by means of a specific type of mathematical structure known as a *lattice*. Classical propositional logic, for instance, can be associated with the structure known as a *Boolean lattice*. The crucial observation here is that in some instances, there exists a connection between lattices on the one hand and state spaces on the other. For example, by considering the set of all subsets of a given state space for classical physics, we arrive at a Boolean lattice and the corresponding logic that is classical propositional logic. In the same vein, we can associate to a state space for quantum physics, i.e. a Hilbert space, a special kind of lattice to which we may then associate a

[29] Note that this applies only to the early incarnation of the state-space approach, as presented in (Van Fraassen 1970).

logical system known as *quantum logic*. Further investigation into the general correspondence between state spaces and lattices might thus reasonably be expected to yield some results on the interrelations between model-theoretic and algebraic approaches to the logical analysis of science.

Taking stock, we see that there are ample reasons to prefer the state-space approach over the structuralist framework in setting up a new wave of logical abstractivism. Hence, we conclude our revisionary efforts by propounding:

(iv') A focus on improving the state-space approach to scientific theories.

Putting together the above observations, we now arrive at a foundation for a second wave of logical abstract abstractivism. That is, the SWLA should be expounded around the following four central tenets:

(i') The usage of fully abstract model theory.

(ii') A focus on non-standard logics.

(iii') A focus on logical pluralization.

(iv') A focus on improving the state-space approach.

Of course, tenets (i')–(iv') still leave us with much freedom in specifying the exact framework of the envisioned second wave of logical abstractivism. The next step in explicating this framework would be to settle on a precise definition of the class of logics we would like to employ in our metascientific investigations. It seems reasonable to suppose that this prospective family of logics would sit somewhere between the notions of abstract (modal) logic and institution in its level of generality. Let us refer to this new species of logical systems as **systems of scientific knowledge**, or **knowledge-systems** for short, to reflect its intended aim of formalizing the structure of scientific knowledge. To make this notion of knowledge-system somewhat more tangible, let us briefly consider its four constituents and the manner in which these can reflect salient properties of a given body of knowledge:

- **Vocabulary.** That is, the set/category of signatures of the underlying logic. Whereas signatures were traditionally taken to consist of purely syntactic, meaningless objects, this restriction no longer exists in our generalized setting. Intuitively, we can think of this component as representing the core concepts with which a knowledge-system is concerned. This could be some collection of mathematical quantities, such as position and momentum, or some other, more loosely defined set of objects.

- **Syntax.** That is, the map sending each signature to a class of statements. This tells us how the practitioners of a given knowledge-system, given their core concepts, use these concepts to make statements about their subject matter. In classical physics, a given set of position and momentum variables is 'processed' through the language of calculus so as to result in differential equations of real-valued functions together with some initial conditions. In quantum physics, statements are formed in largely the same manner with the exception that real-valued functions are replaced by self-adjoint operators.

- **Semantics.** That is, the class of models associated to each signature which serves to provide an interpretation of the statements corresponding to that signature. This component of a knowledge-system reflects the portion of the empirical world with which it is concerned.
- **Satisfaction conditions.** That is, the model-theoretic satisfaction condition we define for our given system of knowledge. This captures the criteria employed by the practitioners of a knowledge-system to determine whether a given statement concerning a relevant portion of the empirical world is satisfactory. For contemporary natural science, for instance, such satisfaction conditions would undoubtedly involve some criteria pertaining to issues of scientific experimentation.

As already noted in the discussion of tenet (iv'), systems of knowledge would most likely have a certain modal-logical structure, with the class of models of any given knowledge-system consisting of a particular variety of state spaces. As the vocabulary and syntax of knowledge-systems are concerned, it is apparent from the above descriptions that these will differ wildly from their counterparts in traditional logic. The crucial point here is that since institutional model theory allows us to take *any* arbitrary category to serve as a 'category of signatures' we are, by extension, also free to consider *any* set of objects defined over these signatures to serve as a set of 'sentences'.

In particular, we might employ a logic of which the sentences are given not by merely syntactic expressions, as usual, but already have some (mathematical) information imbued in them. As noted above, we might employ a category in which 'signatures' are given by a particular set of variables representing some physical quantities and the 'sentences' over this signature are given by all possible differential equations over this set of variables. Regardless of the exact implementation, we may note already that explicating the concept of knowledge-system gives rise to a potentially beautiful interplay between the study of abstract modal logics, institutional model theory and the state-space approach to scientific theories. Moreover, it is entirely conceivable that the semi-abstract approach to logics such as topological and probabilistic logic may be of great assistance in developing a 'logic' for state spaces.

The further explication of the above four components and, subsequently, a rigorous formulation of the notion of knowledge-system should be the first priority of any future research on the SWLA. In subsection 5.3.3, we will consider how already at this stage of development the SWLA can be seen to hold much potential for resolving several outstanding issues with extant approaches to metascience. In the process, some additional manners in which the above four components can be used to represent certain aspects of scientific knowledge will be addressed. Before this, however, let us reflect on the role the proposed SWLA would play in the contemporary scene of metascientific research. That is, let us reflect on the relation between the SWLA, as well as logical abstractivism more generally construed, and the categorical school of metascience encountered in the previous chapter.

5.3.2 Categorical Compatibility

Any new research program for metascience would do well to take note of the contemporary research being done within the field. Such comparative studies, not only valuable for providing new vantage points upon one's own framework, also have a very practical dimension: if one wishes to get one's fledgling new framework to become widely accepted within a certain field of study, it seems advisable to connect this framework, for better or worse, to the most recent developments within that field. Of course, this situation is no different for the SWLA. Thus, let us consider how it relates to the only contemporary school of metascience currently under active development: the categorical approach to scientific theories.

On first sight, there would appear to exist a rather large gap between the two approaches. In particular, the methodology employed by the east-coast brand of categorical metascience seems to go against the basic tenets of the SWLA in the following sense. As noted several times in chapter 4, this categorical account of *scientific* theory-structure appears to be limited to the analysis of 'toy theories' formulated within either first-order logic or its coherent fragment. Consequently, one's adherence to this style of categorical metascience is hinged almost entirely on one's conviction that the analyses of such 'toy theories' can reasonably be extended to include cases of actual, concrete scientific theories. And indeed, Halvorson and Tsementzis (2016, 14) make it no secret they subscribe to this conviction:

> "[W]hat we are claiming here clearly amounts to a methodology based on an analogy between the categorical metamathematics of first-order theories and the philosophy of scientific theories. Is there any reason to believe in the fruitfulness of this analogy? We believe so. Category theory brings to the table new constructions and concepts with which to study the metamathematics of first-order theories. And the metamathematics of first-order theories – if anything – is rich in concepts ("theory", "axioms", "interpretations" etc.) that are used heavily in the philosophy of scientific theories. So as long one is not a complete skeptic with respect to the use of logical methods to come up with idealized versions of scientific theories, there is every reason – it seems to us – to take seriously the analogy on which our method relies."

In the final sentence of this quotation, it seems to me that Halvorson and Tsementzis fall prey to the fiend of first-order fixation: the very preconception against which logical abstractivism is directed. A tentative conclusion might thus be that the categorical approach to metascience and the second wave of logical abstractivism represent two radically different viewpoints on the methodology for logical metascience.

Reviewing once more the methods of the categorical approach, one might object to the above polarization. After all, was it not tenet (ii') of the SWLA which stated that we should look for metascientific applications of non-standard logics, including *fragments* of first-order logic? And does the categorical school not employ one particular fragment, viz. the *coherent* fragment of first-order logic, in an essential way in explicating its metascientific concepts? While we might indeed recognize in this description some superficial degree of similarity between both frameworks, we quickly find a fundamental discrepancy when taking the *usage* of this fragment of first-order logic into account. Recall that the characteristic feature of the SWLA is the usage of the properties of logics to reflect certain properties of scientific knowledge. By contrast, the role of coherent logic within the categorical approach still fits largely within the traditional view of logical metascience, in which logic serves to translate the statements of a scientific theory into a formal language as accurately as possible. Indeed, one of the attractive features of coherent logic, as cited by Halvorson and Tsementzis (2016, 3), was that it was roughly as descriptively powerful as first-order logic (as could be seen through the process of Morleyization).

While the differences between the SWLA and the categorical approach seem to run deep and we may not reasonably expect much meaningful interactions to occur between them in their early stages of development, it is still conceivable that the approaches might one day enter into a symbiotic relationship. This is especially true for the *west-coast style* of categorical metascience, which until this point has been left unmentioned. As noted in chapter 4, the west-coast style may be distinguished from its east-coast counterpart in that the former employs categories of structures, which usually do not and *cannot* serve as the categories of models of some coherent or first-order theories. One of the advantages of the abstract approach to logic, however, was exactly the fact that it allowed us to freely specify what kind of mathematical structures we wanted to serve as the models of the logic under consideration. Moreover, the freedom granted to us by institution-independent model theory in defining classes of sentences may very well enable us to define an appropriate sentential counterpart of the categories of structures considered in the west-coast framework.

In addition, recent work by Teh and Tsementzis (2015) falling within the east-coast approach has demonstrated potential for fruitful interrelations with the SWLA. More specially, Teh and Tsementzis (2015, 14) expound an analogy between signatures and theories on the one hand, and (co)tangent bundles of configuration spaces and hyperregular Hamiltonians/Lagrangians on these bundles on the other. Without entering into details, we may note that institutional model theory, with its extreme generalization of the notion of signature, provides us with an interesting vantage point for the further study of the analogy proposed by Teh and Tsementzis. More generally, the fact that institutional model theory offers us a truly *category-theoretic* framework for abstract model theory bodes well for its compatibility with the categorical approach.

In sum, we might conclude that a combination of the categorical school of metascience with the proposed second wave of logical abstractivism is by no means evident. Closer inspection, however, reveals that a proper hybridization of particular components of both approaches might lead to interesting new avenues of research that may be beneficial to the categorical framework as well as the SWLA. Of course, whether this potential can truly be actualized depends heavily on the future development of both strands of metascience.

5.3.3 Applications

Up until this point, the focus has been mainly on arguing for the superiority of the SWLA over the FWLA. We may note, however, that such argumentation holds little value for the metascientist without a prior commitment to the ideology of logical abstractivism. To demonstrate the value of the proposed SWLA on more independent grounds, let us consider in which ways it can be of service to the metascientific enterprise generally construed. To this end, let us consider three problems which I perceive to exist within the field of metascience as well as the manner in which I believe the SWLA can help remedy these deficiencies. Taking up each problem in turn, we have:

The Problem of Lost Beings

When asked to identify the objects of scientific inquiry, broadly conceived, the answer would typically require little deep thought: science is concerned with understanding the behavior of *beings* in the empirical world. Yet, when reflecting on the hitherto encountered frameworks of metascience, we see that the relation between science and the world, any by extension beings in the world, in these frameworks is often elusive at best. Following Muller (2011, 97–8), I refer to this as the problem of *lost beings*. Briefly reviewing, we have:

- **Syntactic approach:** In the positivist framework, beings serve as interpretations of the observational terms in our logical language. How this process proceeds exactly, however, is left to the imagination: at no point are the beings that science is about formally introduced into the account.
- **State-space approach:** The physical system modeled by the state space H was given in this account by the object X. We may recall, however, that this object X only served to fix the initial state in H trough the 'location map' *loc*. No additional structure is imposed on X, effectively making it a meaningless object.
- **Structuralist approach:** Here, the empirical world enters into the mix through the inclusion of the set I of intended applications in the pair $T = (K, I)$ representing a scientific theory. Now, since $I \subseteq M_{pp}$, we see that beings in the structuralist account are effectively represented by partial potential models. The link between theories and the world is then accounted for by the empirical claim of a theory-element (cf. definition 2.4.21) that $I \in Cn$.

- **Categorical approach:** While the categorical view of scientific theories assigns much value to the explication of relations between individual theories, the relation of these theories to beings in the world as of yet remains completely unspecified.

Indeed, we see that the role of beings and the theory-world relation remain sorely underdeveloped in the syntactic and state-space approaches to scientific theories, while they are entirely absent in the categorical approach. The only framework which seems to impose a non-trivial amount of structure on beings is that of the structuralists. Ultimately, however, the structuralist formalization proves to be equally unsatisfactory.

To see this, it is sufficient to exclaim the following, obvious observation: *empirical beings are different from their formal representation in theories, describing these beings.* To name but one salient aspect in which beings differ from mathematical entities found in scientific theories, we might note that beings in the empirical world are 'infinitely complex' as opposed to the more self-contained, idealized entities produced by mathematical abstraction. For example, while we may ascribe infinitely many properties to a ball, e.g. shape, size, color, material, year of production, number of goals scored with it, while in the theory of classical mechanics this list is reduced to a few variables pertaining to its motion. Without wishing to get involved here in the elaborate discussions concerning the nature of scientific representation, modeling, abstraction and idealization, the banal point I wish to emphasize is that differences between beings and their formal representations exist and should be reflected accordingly in any adequate account of metascience.

In this sense, the structuralist framework fares no better than its competitors. By representing beings by means of partial potential models, objects which simultaneously represent internal components of scientific theories, the structuralist fail to account for the differences that exist between beings existing in the empirical world and formal entities in abstract, mathematical realms. We can thus conclude that no approach to metascience, past or present, provides us with a satisfying account of the role of beings in the scientific endeavor.

How can the SWLA be service here? This may be seen by first drawing a parallel with the field of metamathematics. Moving down one rung on the ladder of abstraction, we see that just we can use mathematical entities to talk about and formalize empirical beings, so too can we analyze mathematical beings by means of logical, syntactic sentences. But this, we may note, is exactly the aim of model theory, a formalism with which we have at this stage become quite acquainted. In analogy with model theory, the appropriate place for beings in metascientific frameworks thus seems to be at the place of models providing empirical meaning to abstract, mathematical statements of scientific theories. With this goal in mind, the potential of the SWLA for retrieving the lost beings now becomes apparent. Making full use of the freedom offered to us by the abstract model-theoretic view of logic, we can simply define a logic the sentences of which are given by abstract, mathematical statements and whose class of models consists of structures which, in one way or another, represent beings from the empirical world.

The exact manner in which this vision can be brought to fruition will require further investigation. In particular, the precise way in which the representation of beings at the level of models is to be combined with the focus of the SWLA on the state-space approach to theories, in which the role of models is seemingly played by the mathematical structures of state spaces, warrants additional consideration. It is conceivable, however, that we can modify the state spaces of the Beth-Van Fraassen framework in such a manner that they reflect the infinitely complex nature of the empirical world. The 'mathematical' state spaces would then still appear in the model class of the abstract logic as special, simplified instances of the more complicated models. Another point worth mentioning is that the above discussion for a large part presupposes a physics-centered view of science, as reflected in the branding of scientific statements as being *abstract* and *mathematical* in nature. How these considerations are to be extended to different, less mathematized areas of science, such as the field of biology, will need to be examined at a later stage.

The Problem of Historical Myopia

Norwood Russell Hanson famously remarked that 'history of science without philosophy of science is blind' and 'philosophy of science without history of science is empty' (1962, 580).[30] Over half a century later, we see Hanson's statement has not fallen on deaf ears. In recent times, the field of *integrated history and philosophy of science* has made much headway and seems to have gained significant support within the relevant academic circles.[31] Replacing, however, *philosophy of science* with *formal philosophy of science* in Hanson's credo and our outlook becomes a lot less rose-colored.[32] Indeed, it can be noted that, on the whole, formal philosophers of science have been blind to historical considerations. Notable exceptions here are the structuralist, who devote a significant portion of their works to the study of 'theory dynamics'.[33], as well as Pearce and Rantala (1983b). To my knowledge, however, there exists as of yet no formal analysis of any scientific theory predating the sixteenth century.

Why should we be interested in formalizing the science from this span of history? After all, it may be argued that *science*, as we know it, did not take form until the sixteenth or perhaps even seventeenth century. It is, however, precisely this fact that makes this historical period so appealing. Looking at, for instance, the 'science' of the medieval period or classical antiquity, the great conceptual disparity with the modern-day approach to science immediately stands out. Hence, any metascientist wishing to understanding the formal structure of science *at the most general level possible* would do well to include these time periods within the scope of their investigations.

Naturally, the task of constructing a historically all-encompassing framework

[30] The statement is sometimes attributed to Lakatos, who expressed it in his (1970).

[31] As evidenced by the &HPS series of biannual conferences dedicated to the subject.

[32] Let us, for present purposes, equate formal philosophy of science with metascience.

[33] See, in particular, Stegmüller (1976).

for the study of metascience is one of gargantuan proportions and I do not mean to suggest that metascience should necessarily concern itself with the formalization of each and every scientific concoction history has to offer. Rather, I merely put forth the claim that a moderate increase in historical diversity can be of great assistance in dispelling any modern preconceptions we might have regarding the nature of science. A poignant example here is the relation between physics and mathematics. A widely held perception is that the latter is indispensable for the former. But what becomes of this conviction when we consider one of the most influential systems of physics in all of history, viz. **Aristotelian physics**? While the presence of mathematics in our theories of physics is often taken for granted in modern times, Aristotle's physics of the sublunar realm was known for being explicitly anti-mathematical.

The methodology of the SWLA, I hold, is well-suited to accommodate a more historically diverse metascience. The crucial observation here is that by adopting the tenet of logical pluralization, we have available to us a great many more components with which to formalize different aspect of a given system of science than we would have by employing logic in the traditional manner. Take, for instance, the theories of Aristotelian, Newontian and Lagrangian mechanics. Clearly, the latter two are much more closely related to one another than they are to the former. In the traditional approach to logical metascience, however, we are forced to formalize these scientific theories as theories in the same underlying logic, be it first-order logic or some extension thereof. As a result, we are relegated to differentiating between the three theories only by comparing the different content of the statements making up their logical theories.

In the SWLA, by contrast, the choice of logic, or *knowledge-system* as I referred to it in the previous subsection, will now also be representative of the type of scientific theory we are attempting to formalize. Hence, the choice of what kind of objects we employ as our signatures, which type of syntax we use to form classes of sentences from these signatures and what sort of models we allow in our knowledge-system can all be used to reflect some salient aspects of the scientific theory under consideration, allowing us to differentiate between theories in a much more fine-grained manner.

To give but one example of how this might proceed, let us return to the example of the science-mathematics relation. To formalize this difference using traditional methodology, i.e. working in some fixed logical system, we would most likely be required to formulate in our logical theories a host of auxiliary assumptions which express that some terms occurring in our language are to be deemed as *quantitative* or *mathematical*, while some other are to be considered *qualitative* or *non-mathematical*. In our new approach, however, we can do away which such 'clutter' altogether, by simply formulating the theory of Aristotelian mechanics relative to a knowledge-system of which the syntax sends any signature to an appropriately chosen set of non-mathematical statements, while the knowledge-system underlying Newtonian and Lagrangian mechanics would employ a syntax sending each signature to a set of mathematical statements in the language of calculus.

Naturally, the formal concept of knowledge-system will need to be developed in greater detail before we can accurately determine its usefulness to historical considerations. At the same time, we can employ widespread historical applicability as one of our desiderata when explicating the exact definitions of the different components of the knowledge-system concept. It seems reasonable to assume, however, that the great freedom associated to logical pluralization can assist to at least some extent in correcting historical myopia.

The Problem of Theory-Centrism

In section 1.2 it was noted how, for lack of diversity, we could identify the formal analysis of the structure of science, i.e. metascience, with the formal analysis of the structure of scientific *theories*. Now, while it must be noted that formal methods have been applied to some other areas within philosophy of science, the discourse on theory-structure remains the only field which comes close to providing us with a formalized, big picture of science. Let us refer to this preoccupation with the notion of scientific theory as *theory-centrism*. It is my conviction, then, that theory-centrism has hampered the development of metascience. More specifically, I hold that theory-centrism has made it so that our metascientific formalisms are unable to produce a truly *unified* analysis of the structure of the scientific enterprise.

Why would theory-centrism inhibit the unifying power of a metascientific framework? Surely, there is no denying that theories are crucial in the scientific endeavor. However, by insisting that any approach to metascience be theory-centric, we are unknowingly making this account unable to cope with several other salient aspects of science. A case in point here is the relation between physics and mathematics. The nature of this relation has seen much debate both inside and outside the philosophy of science. Yet, no account of scientific theory-structure has been able to incorporate the relation in its framework, for the simple reason that *being mathematical* is not a property of individual physical theories, but of modern-day physics as a whole. Phrased differently: it is a property of the system of knowledge that is modern physics. This observation is crucial. Instead of a theory-centric metascience, we should aim to establish a *systems-based* metascience, i.e. an account of metascience whose primary unit of analysis consists of *systems of knowledge* or *knowledge-systems*.

Of course, since we noted already at the end of subsection 5.3.1 that by using abstract model theory we can arrive at a new type of logic for metascientific research similarly referred to as *knowledge-system*, the connection between the SWLA and systems-based metascience is now clear: by specializing the concept of institution in a metascientifically relevant manner, with appropriate inspiration provided by study of abstract (modal) logics, we obtain a formal notion of knowledge-system, which can then be used to serve as the foundation for a systems-based approach to metascience. Such a metascience would, in turn, enable us to formulate a uniform framework to discuss a number of hitherto disjointed subfields of the philosophy of science, each concerned with a different scientific concept.

As an example, let us consider once more the science-mathematics relation.

Within the philosophy of science, we find devoted literature on the topics of the 'applicability of mathematics' (Pincock 2001), (Rizza 2013), the 'explanatory role of mathematics in the empirical sciences' (Batterman 2010), (Bueno & French 2012) and the 'indispensability of mathematics' to science (Colyvan 2001). None of these debates, however, have been connected to the philosophical literature on scientific theories. Meanwhile, the process of *mathematization* has received no significant philosophical discussion whatsoever and have remained exclusively within the purview of the history of science. There is a clear role to be played by knowledge-systems, since the adoption of a 'mathematical language' of a particular system of science can essentially be formalized as the choice of a knowledge-system with a certain kind of syntax. The challenge for the metascientist then becomes to model in logical terms what it means for syntax to constitute as either *mathematical* or *non-mathematical*.

Another salient example can be found in the concept of scientific explanation. This concept, we may note, is a controversial one that has seen at least as much debate as the notion of theory. In contrast to the previous example, however, there actually have been some limited attempts at extending accounts of scientific theory-structure to include the notion of explanation. Hempel and Oppenheim (1948) famously developed the *deductive-nomological model* for explanation against the backdrop of their syntactic view on theories. Less well-known are attempts by Sneed (1994) and Suppe (1989, 152–201) to incorporate scientific explanation within the structuralist and state-space frameworks respectively. It may be noted, however, that all of these are accounts of *theoretical explanation*, i.e. the manner in which scientific *theories* are used in explaining empirical phenomena. This identification of scientific explanation with theoretical explanation may be viewed as a prime example of the negative influence of theory-centrism on metascience. After all, some instances of explanation are inherent not to individual theories, but to systems of knowledge in their entirety, e.g. teleological explanations in Aristotelian physics. While it is by no means a straightforward task to explicate the notion of explanation in model-theoretic terms, the knowledge-system concept as obtained from the SWLA provides us with a natural environment within which to take up these considerations.

It may be noted that the above examples both involve historical considerations to some degree. We can thus view the problem of theory-centrism as being intimately connected with the problem of historical myopia. This should not be entirely surprising: the further back we go in history, the more 'high-level' the differences with modern science become and, consequently, the greater the odds will be of these differences being more than simple disparities between different scientific theories. In addition, we see that the manner of representing beings in the empirical world and relating these to scientific statements is also definitely an aspect of science transcending the properties of individual theories. Hence, we see that the introduction of knowledge-systems can have a positive impact on the problem of lost beings as well. In sum, we can conclude that a systems-based metascience constructed around the central tenets of the SWLA contains much potential for resolving some of the key issues plaguing the current metascientific enterprise.

Epilegomena

Let us, for the moment, conclude our ruminations on metascience, abstract model theory and the interface between the two. We have encountered a variety of approaches to the metatheoretical study of science, ranging from the syntactic framework of the positivists, to the different flavors of the structural approach, to modern-day category-theoretic explorations. In tandem, we were introduced to a number of different model-theoretic methodologies for the abstract study of logics. Each of these frameworks were seen to include many different concepts and technical results of great interest within their scopes, with the abstract maximality theorem for abstract logics and Enqvist's theorem for abstract modal logics standing out in particular.

Another theorem, viz. the uniform reduction theorem, was seen to form the central point of departure for Pearce and Rantala's attempt at infusing metascience with the methods of abstract model theory. This first wave of logical abstractivism, however, was observed to be of only marginal success. We identified several deficiencies to which we could ascribe this lack of fruitfulness, most poignantly an adherence to a methodology that fails to utilize the full potential of abstract model theory and in a definite sense remains bound to the methodological restrictions of first-order fixation.

In wake of the apparent failure of the FWLA, a different proposal for a logically abstractivist metascience was put forward. In contrast to its predecessor, this new approach would proceed in such a manner that makes use of the full capacities of the traditional framework of abstract model theory and also connects to more recent developments within the field. More specially, it was argued that a combination of institution-independent model theory, the study of abstract modal logics and the more abstract components of semi-abstract model theory together form a fruitful backdrop against which a metascientific framework may be developed, centered around the state-space approach.

This prospective second wave of logical abstractivism, I argued, could be of great assistance in resolving a number of exigent issues in the philosophy-of-science literature. In particular, we have seen how the SWLA can be used to combat the alarming trend of theory-centrism within the philosophy of science and allows us to move to what I have termed a *systems-based* metasience. Relatedly, the SWLA was argued to be of much effect in remedying the problems of lost beings and historical myopia. In this context, the SWLA could be construed as a setup for a logico-historical approach to metascience, filling the void between formal and historical approaches to philosophy of science.

At this stage, I hope the reader finds themselves convinced of the viability and desirability of the SWLA and, most importantly, of the ideology of logical abstractivsm in general. It seems to me a lamentable state-of-affairs that logical abstractivism, over three decades after its inception at the hands of Pearce and Rantala, remains known to a select few researchers in what can already be considered an academic niche. At the same time, we might note that neither abstract model theory, nor any other framework falling within the umbrella of universal logic, are particularly well-known among logicians in general, making it all the more difficult for such formalism to be fruitfully exported to other disciplines such as metascience. This is in firm contrast with fields as homotopy type theory and topos theory which, high levels of abstraction notwithstanding, seem to enjoy great popularity among practitioners of logic, mathematics, as well as philosophy of science. Indeed, it is not *in spite of* high levels of abstraction that these formalisms have seen significant uptake, but exactly *because* of their great degrees of generality. This observation, however, makes the 'passing over' of abstract model theory, or any framework within universal logic for that matter, by such a large portion of the relevant academic communities all the more regrettable.

There is room, however, for reserved optimism. While, to my knowledge, abstract model theory remains known to only a small number of mathematical and philosophical logicians, the wider field of universal logic seems to be gaining in popularity and recognition. The last decade has seen the founding of a dedicated journal *Logica Universalis* as well as the organization of the *UniLog* series of conferences. The time, it seems, is ripe for philosophical application of universal logic. Whether it is abstract model theory that will emerge as the formalism of choice for metascience or it is some different framework, e.g. algebraic logic, that will be bestowed this role, I hope this thesis has demonstrated the great potential of logical abstractivism for the metascientific enterprise.

The clichéd observation, endemic to many a concluding remark, that much work still needs to be done holds true for the present thesis as well. The main order business constitutes the rigorous explication of a notion of knowledge-system, similar to that of *abstract (modal) logic* and *institution*, adequate for metascientific considerations. En passant, this will lead us to consider the problems of lost beings, historical myopia and theory-centrism, each of which has to be carefully analyzed and (partially) resolved before we can arrive at a truly satisfactory notion of knowledge-system.

The epitome of abstract model theory for logical metascience would undoubtedly be the formulation of Lindström-style characterization theorems for particular systems of knowledge, such as classical and quantum physics. Moreover, by noting that such a Lindström theorem would reveal the inherent limitations of the knowledge-system under consideration, we see that this opens up another natural avenue through which the history of science might enter into our logico-philosophical mixture. It is in this manner that I hope to one day arrive at a fully integrated history, philosophy and logic of science. To be sure, there is no shortage of grand ambitions relating to the future development of the SWLA. For the moment, however, let us put such grandiose dreams on hold and conclude here the cerebrations of this master's thesis.

Appendix A

Category Theory

In this appendix, I reiterate some basic facts from category theory, which we require at several different points in the thesis.

A.1 Basic Definitions

We start by considering some basic definitions to lay the groundwork for the subsequent sections.

Definition A.1.1. A category C consists of the following three components:

- A class $\mathrm{ob}(C)$ of *objects*, sometimes also denoted as $|\mathcal{C}|$.
- A class $\hom(C)$ of *morphisms*, such that each morphism f has a *source object* $\mathrm{dom}(f)$ and *target object* $\mathrm{cod}(f)$. A morphism f with source object a and target object b is denoted as $f : a \to b$. We say that a morphism $f : a \to b$ is a morphism *from a to b*. The class of all morphism from an object a to an object b in the category C is denoted as $\hom_C(a, b)$.
- For every three objects a, b, c, a binary operation

  $$\circ : \hom_C(a, b) \times \hom_C(b, c) \to \hom_C(a, c)$$

  called the *composition of morphisms*. If we have the morphisms $f : a \to b$ and $g : b \to c$, we denote the composition of f and g as $g \circ f : a \to c$.

Furthermore, we require C to satisfy the following axioms:

- *Associativity.* For any three morphisms $f : a \to b, g : b \to c, h : c \to d$, $(h \circ g) \circ f = h \circ (g \circ f)$.
- *Identity.* For every object x, there exists a morphism $\mathrm{Id}_x :\to x \to x$, called the *identity morphism* for x, such that for every morphism $f : a \to x$ and $g : x \to b$, we have $\mathrm{Id}_x \circ f = f$ and $g \circ \mathrm{Id}_x = g$.

Definition A.1.2. Let C and D be categories. A *functor* F from C to D is a mapping that

- sends each object in x in C to an object $F(x)$ in D,
- sends each morphism $f : a \to b$ to a morphism $F(f) : F(a) \to F(b)$ in D such that
  - $F(\mathrm{Id}_x) = \mathrm{Id}_{F(x)}$ for every object x in C,
  - $F(g \circ f) = F(g) \circ F(f)$ for all $f : a \to b$ and $g : b \to c$ in C.

There are myriad examples of categories which may be cited to further illustrate the notion. Two examples that are particularly salient are the following.

Example A.1.3. The category of sets, denoted **Set**, having sets as objects, functions between sets as morphisms and the usual composition of functions as morphism composition.

Example A.1.4. The category of categories, denoted **Cat**, having categories as objects, functors between categories as morphisms and functor composition as morphism composition.

Definition A.1.5. Let C be a category. A *subcategory* S of C is a category such that

- $\mathrm{ob}(S)$ is a subclass of $\mathrm{ob}(C)$,
- $\mathrm{hom}(S)$ is a subclass of $\mathrm{hom}(C)$,

and it holds that

- for every object x in C, the identity morphism Id_x is a morphism in S,
- for every morphism $f : a \to b$, the source object a and target object b are in S,
- for every pair of morphisms $f : a \to b, g : b \to c$ in S, the composite morphism $g \circ f$ is in S.

Definition A.1.6. Let C be a category and let $f : a \to b$ be a morphism in C. We say that f is *invertible* if there exists a morphism $f^{-1} : b \to a$ such that $f^{-1} \circ f = \mathrm{Id}_a$ and $f \circ f^{-1} = \mathrm{Id}_b$. We call f^{-1} the *inverse* of f. It is easily checked that every morphism has a unique inverse.

Definition A.1.7. An *isomorphism* is an invertible morphism. Two objects in a category are called *isomorphic* if there exists an isomorphism between them.

Definition A.1.8. A category C is called *skeletal* if for every pair of objects a, b in C: a and b are isomorphic if and only if they are equal.

Definition A.1.9. Let C and D be categories and let $F, G : C \to D$ be functors. A *natural transformation* from F to G, denoted $\alpha : F \Rightarrow G$, is a function sending every object x in C to a morphism $\alpha_x : F(x) \to G(x)$ in D, which is called the *component* of α in x, such that for any morphism $f : a \to b$ in C, the following diagram commutes:

$$\begin{tikzcd}
F(a) \arrow[r, "F(f)"] \arrow[d, "\alpha_a"'] & F(b) \arrow[d, "\alpha_y"] \\
G(a) \arrow[r, "G(f)"] & G(b)
\end{tikzcd}$$

Definition A.1.10. Let C and D be categories, let $F, G : C \to D$ be functors and let $\alpha : F \to G$ be natural transformation. We call α a *natural isomorphism* if, for every object x in C, the component of α in x is an isomorphism.

The employed terminology compels us to inquire about the relation between isomorphisms as pertaining to objects in categories and natural isomorphisms as pertaining to functors between categories. Fittingly, the two notions coincide within the appropriate category:

Definition A.1.11. Let C and D be categories. The *functor category* $[C, D]$ is defined to be the category whose

- objects are the functors from C to D,
- morphisms are the natural transformations between these functors.

Proposition A.1.12. *For any categories C, D, functors $F, G : C \to D$ and natural transformation $\alpha : F \Rightarrow G$, it holds that α is a natural isomorphism if and only if α is an isomorphism in $[C, D]$.*

A.2 Equivalence

With all preliminaries in place, we now turn to the question of how to define a useful notion of equivalence between categories. Let us begin by considering the obvious candidate.

Definition A.2.1. Let C and D be categories and let $F : C \to D$ be a functor. F is called *invertible* if there exists another functor $F^{-1} : D \to C$ such that $F^{-1} \circ F = \mathrm{Id}_C$ and $F \circ F^{-1} = \mathrm{Id}_D$. Here, Id_C and Id_D denote the identity functors for C and D respectively and $\circ$ refers to the composition of functors in the obvious way.

Definition A.2.2. Let C and D be categories and let $F : C \to D$ be a functor. F is called an *isomorphism of categories* if F is invertible. If there exists an isomorphism of categories between C and D, the two categories are said to be *isomorphic.*

Note that this definition of isomorphism aligns perfectly with definition A.1.7 when applied to the category of categories Cat, having categories as objects and functors as morphisms.

Isomorphism, as it turns out, does *not* provide us with a satisfactory notion of equivalence. To see this, consider the following example.

Example A.2.3. Consider the category $\bullet$ consisting of only a single object together with its identity morphism. Next, consider the category $\bullet \leftrightarrows \bullet$ consisting of two objects along with their two identity morphisms, a morphism from one object to the other and the inverse of this morphism. Clearly, the two objects in the latter category are isomorphic. For all intents and purposes, we can thus consider the category $\bullet \leftrightarrows \bullet$ to consist of only a single object, which, in turn, allows us to consider $\bullet$ and $\bullet \leftrightarrows \bullet$ as essentially the same category. However, a brief survey of all possible functors between the two categories reveals there can exist no isomorphism of categories between them.

Rather than isomorphism, we are interested in a less fine grained definition of equivalence. The following definition, as it turns out, is just right:

Definition A.2.4. Let C and D be categories. An *equivalence of categories* between C and D is a functor $F : C \to D$ for which there exists a functor $G : D \to C$ such that there exist natural isomorphisms $\alpha : G \circ F \Rightarrow \mathrm{Id}_C$ and $\beta : F \circ G \Rightarrow \mathrm{Id}_D$.

Definition A.2.5. Two categories are said to be *equivalent* if there exists an equivalence of categories between them.

The crucial point here is that, using this definition of equivalence, we no longer demand the compositions of the functors F and G be *equal* to the identity functors, but we only require their compositions to be *naturally isomorphic* to the identity functors. This liberalization allows us to circumvent the problems typically associated with isomorphisms of categories.

While the above definition provides us with a satisfactory notion of equivalence, in category theory it is often useful to employ an alternative characterization of equivalence of categories. Consider the following properties.

Definition A.2.6. Let C and D be categories and let $F : C \to D$ be a functor. We say that

- the functor F is *full* if for every two objects c_1, c_2 in C and every morphism $g : F(c_1) \to F(c_2)$ in D, there exists a morphism $f : c_1 \to c_2$ in C such that $F(f) = g$.
- the functor F is *faithful* if $F(f) = F(g)$ implies $f = g$ for all morphisms f, g in C.
- the functor F is *essentially surjective* if for every object d in D, there exists an object c in C such that $F(c)$ is isomorphic to d.

This provides us with the following alternative characterization of equivalence between categories:

Proposition A.2.7. *A functor is an equivalence of categories if and only if it is full, faithful and essentially surjective.*

Armed with our notion of equivalence, we are free to define a host of new, important category-theoretic concepts. One such concept is the following.

Definition A.2.8. Let C be a category. A *skeleton* of C is a skeletal subcategory of C that is equivalent to C.

Significantly, we have:

Proposition A.2.9. *Any category has a skeleton.*

Thus, whenever we are studying the properties of a particular class of categories we can, by proposition A.2.9, restrict our attention to the skeletons of these categories, at least as properties up to equivalence are concerned. One type of category for which this strategy proves to be convenient is the following.

Definition A.2.10. A category is *discrete* if it is equivalent to a category whose only morphisms are identity morphisms.

Lemma A.2.11. *Let C be a category. If every morphism in C is an isomorphism and there is at most one morphism from any one object to another, then C is a discrete category.*

Proof. By proposition A.2.9, we know that C has a skeleton S. Let a, b be objects and in S let $f : a \to b$ be a morphism in S. By assumption, we have that f is an isomorphism between a and b. Furthermore, we know S, by definition, is skeletal and thus we must have $a = b$. So f is a morphism from a to a. But we also know, again by assumption, that there is at most one morphism from any one object to another. This implies that f is the identity morphism from a to a. We infer that S contains only identity morphisms. Finally, note that, by definition, S is equivalent to C. Hence, C is a discrete category. ■

Appendix B

Topos Theory

The present appendix serves to present the reader with a brief introduction to the theory of topoi. In short, a topos refers to particular kind of category having certain 'nice' features which ensures the categories in question behave like the category of sets. Topos theory is of deep, fundamental importance in many different ways to many different researchers and is being applied in fields as diverse as logic, algebraic geometry, quantum physics and philosophy of science. The material covered in this appendix is based almost entirely on the excellent introductory text by Goldblatt (1984).

B.1 Preliminary Notions

To establish what exactly *is* the notion of a topos, we first require several preliminary definitions. The general strategy will be to look at properties holding for the category **Set** and see how these may be lifted to an arbitrary category. First, we require a generalized notion of injective and surjective functions.

Definition B.1.1. Let C be a category and $f : a \to b$ a morphism in C. We say f is *monic* if for any object c and morphisms $g, h : c \to a$ in C, we have $f \circ g = f \circ h$ implies $g = h$. We denote the fact that f is monic by writing $f : a \rightarrowtail b$.

Definition B.1.2. Let C be a category and $f : a \to b$ a morphism in C. We say f is *epic* if for any object c and morphisms $g, h : b \to c$ in C, we have $g \circ f = h \circ f$ implies $g = h$. We denote the fact that f is epic by writing $f : a \twoheadrightarrow b$.

It is straightforwardly verified that monic and epic morphisms in **Set** coincide with injective and surjective functions respectively. Next, we want categorical equivalents of the empty set and the singleton sets in **Set**.

Definition B.1.3. An object 0 in a category C is *initial* if for every object a in C, there is one and only one morphism from 0 to a in C.

Definition B.1.4. An object 1 in a category C is *terminal* if for every object a in C, there is one and only one morphism from a to 1 in C.

Again, it is easily seen that the empty set and the singleton take on th role of initial object and terminal objects in **Set**.

Our next order of business is to define the notions of *limit* and *co-limit*. Unlike the preceding definitions, the intuitive interpretation of these concepts in terms of the category of sets is not immediately clear. Let us thus motivate the definitions that are to come by considering two examples: the categorical counterparts of the product set construction and the co-product construction.

Definition B.1.5. Let C be a category and let a, b be objects in C. A *product* of a and b in C is taken to be an object $a \times b$ with a pair of morphisms $\pi_a : a \times b \to a$, $\pi_b : a \times b \to b$ in C such that for any object c and any two morphisms $f : c \to a$ and $g : c \to b$ in C there exists exactly one morphism $(f, g) : c \to a \times b$ making the following diagram commute:

$$\begin{tikzcd}
& c \arrow[ld, "f"'] \arrow[d, dashed, "{(f,g)}"] \arrow[rd, "g"] & \\
a & a \times b \arrow[l, "\pi_a"'] \arrow[r, "\pi_b"] & b
\end{tikzcd} \tag{B.1}$$

A point of notation: the dashed arrow for the morphism $(f, g) : c \to a \times b$ in the above diagrams indicates that this is the only morphism that can make this diagram commute.

The co-product construction can be generalized to arbitrary categories in a similar fashion.

Definition B.1.6. Let C be a category and let a, b be objects in C. A *co-product* of a and b in C is taken to be an object $a + b$ with a pair of morphisms $i_a : a \to a + b$, $i_b : b \to a + b$ in C such that for any object c and any two morphisms $f : a \to c$ and $g : b \to c$ in C there exists exactly one morphism $[f, g] : a + b \to c$ making the following diagram commute:

$$\begin{tikzcd}
a \arrow[r, "i_a"] \arrow[rd, "f"'] & a + b \arrow[d, dashed, "{[f,g]}"] & b \arrow[l, "i_b"'] \arrow[ld, "g"] \\
& c &
\end{tikzcd} \tag{B.2}$$

There is now an extremely important point to be made concerning these two constructions. First, let me introduce some terminology. We start by considering the notion of *diagram*. A completely rigorous explication of this concept is somewhat cumbersome. For present purposes, the following, more informal definition suffices.

Definition B.1.7. A *diagram* D in a category C is a collection of some objects $d_i, d_j, \ldots$ in C, where we allow the same object to appear more than once, together with some morphisms $g : d_i \to d_j$ in C between some of the objects in the diagram, where we allow the same morphism to appear more than once.

Building on the notion of diagram, we now have:

Definition B.1.8. Let C be a category and D a diagram in C. A *cone* for D consists of an object c in C with morphisms $f_i : c \to d$ in C for every d in the diagram D such that for every d_i, d_j and $g : d_i \to d_j$ in D the following diagram commutes:

$$\begin{tikzcd}
d_i \arrow[rr, "g"] & & d_j \\
& c \arrow[ul, "f_i"] \arrow[ur, "f_j"'] &
\end{tikzcd} \tag{B.3}$$

Definition B.1.9. Let C be a category and D a diagram in C. A *co-cone* for D consists of an object c in C with morphisms $f_i : d \to c$ in C for every d in the diagram D such that for every d_i, d_j and $g : d_i \to d_j$ in D the following diagram commutes:

$$\begin{tikzcd}
d_i \arrow[rr, "g"] \arrow[dr, "f_i"'] & & d_j \arrow[dl, "f_j"] \\
& c &
\end{tikzcd} \tag{B.4}$$

While these definitions may seem rather *ad hoc* at first glance, they actually enable us to characterize the notions of product and co-product in an effective manner. To see this, let C be an arbitrary category and consider the diagram D consisting of only two C-objects a and b and no morphisms. For each C-object c, we have a cone for this diagram of the form:

$$\begin{tikzcd}
a \arrow[dr, "f_1"'] & & b \arrow[dl, "f_2"] \\
& c &
\end{tikzcd} \tag{B.5}$$

So, in particular, for the C-object $a \times b$ and C-morphisms $\pi_a : a \times b \to a$, $\pi_b : a \times b \to b$, we have the cone:

$$\begin{tikzcd}
a \arrow[dr, "\pi_a"'] & & b \arrow[dl, "\pi_b"] \\
& a \times b &
\end{tikzcd} \tag{B.6}$$

But *this* cone is not just any cone! By definition of the production construction, we know that, for any C-object c, the diagrams

$$\begin{tikzcd} & a & \\ a \times b \arrow[ur, "\pi_a"] & & c \arrow[ul, "f_1"'] \arrow[ll, dashed, "{(f,g)}"] \end{tikzcd} \qquad \begin{tikzcd} & b & \\ a \times b \arrow[ur, "\pi_b"] & & c \arrow[ul, "f_2"'] \arrow[ll, dashed, "{(f,g)}"] \end{tikzcd} \tag{B.7}$$

commute. That is to say, every other cone for diagram D factors uniquely through the cone B.6. In fact, we might just as well say that the product of the objects a and b *is* exactly the cone for D such that any other cone for D factors uniquely through it. This property turns out to be of such significance that it has been awarded a special name.

Definition B.1.10. Let C be a category and let D a diagram in C. A cone for D is called a *limiting cone* or simply a *limit*, if any cone for D factors uniquely through it.

The notion of limit brings with it a tremendous unifying power. Many other category-theoretic constructions can be characterized as being the limit for the appropriate choice of diagram. An easy example is the case of an empty diagram for any category C. For an empty diagram, a cone is simply given by any object in C, without morphisms. Thus, a limit for an empty diagram, is simply a C-object c such that for any other C-object c', we have that c' factors uniquely through c, i.e. there exists exactly one morphism f such that $f : c' \dashrightarrow c$. This is, however, precisely the requirement for c to be a terminal object in C. Hence, we can view the terminal object construction for a category as simply being the limit for the empty diagram in that category.

By a similar line of reasoning, we can also link together the co-product construction with the notion of a co-cone and, subsequently, the notion of a co-limit. We see, for a given category C and diagram D consisting of the C-objects a and b, that the co-product $a + b$ with morphisms $i_a : a \to a + b$ and $i_b : b \to a + b$ is exactly the co-cone for D which factors uniquely through any other co-cone for D. Thus, we have

Definition B.1.11. Let C be a category and let D a diagram in C. A co-cone for D is called a *co-limiting co-cone* or simply a *co-limit*, if it factors uniquely through any co-cone for D.

Analogously with the exposition for limits, we can look at the co-limit for the empty diagram of a given category C. A co-cone for this diagram will then be given by any object in C. Now, if the C-object c is a co-limit for the empty diagram, then we have, for any other C-object c', that c factors uniquely through c', i.e. there exists exactly one morphism $f : c \dashrightarrow c'$. As we may have expected, this is precisely the definition of an initial object in the category C.

B.2 Defining Topoi

With the above preliminaries in place, we are now poised to formulate two of the defining properties of topoi.

Definition B.2.1. Let C be a category. We say C is *finitely complete* if it has a limit for every finite diagram in C. Similarly, we say C is *finitely co-complete* if it has a co-limit for every finite diagram in C.

The above definition is readily generalized to give us the notions of *completeness* and *co-completeness* by dropping the restriction to finite diagrams. However, we will not be requiring these two stronger concepts for our present purposes.

Besides finite completeness and finite co-completeness, there are two additional properties a category should satisfy for it to be a topos. First, we want to define a notion of *exponential object*. Intuitively, this notion can be thought to generalize the construction in **Set** where we take two sets A and B and with these form the set B^A of functions from A to B. Before we can categorically define this construction, we require a preliminary.

Definition B.2.2. Let C be a category and let $f : a \to b$ and $g : c \to d$ be morphisms in C. We define the *product morphism* $f \times g : a \times c \to b \times d$ to be the unique C-morphism making the following diagram commute:

$$\begin{tikzcd}
 & c \arrow[r, "g"'] & d \\
a \times c \arrow[ur, "\pi_c"] \arrow[rr, dashed, "f\times g"] \arrow[dr, "\pi_a"'] & & b \times d \arrow[u, "\pi_d"'] \arrow[d, "\pi_d"] \\
 & a \arrow[r, "f"] & b
\end{tikzcd} \tag{B.8}$$

Now, we can define the concept of an exponential object as follows.

Definition B.2.3. Let C be a category such that for any two C-objects, C contains the product of these objects. Furthermore, let a, b be objects in C. An *exponential object* for a and b is a C-object b^a together with a C-morphism $ev : b^a \times a \to b$, called an *evaluation morphism*, such that for any C-object c there exists a unique morphism $\hat{g} : c \to b^a$ making the following diagram commute:

$$\begin{tikzcd}
b^a \times a \arrow[dr, "ev"] & \\
 & b \\
c \times a \arrow[uu, dashed, "\hat{g}\times Id_a"] \arrow[ur, "g"'] &
\end{tikzcd} \tag{B.9}$$

For a motivation of the above definition, consult (Goldblatt 1984, 70–4). Now, the third condition we impose on categories in our definition of topos reads as follows.

Definition B.2.4. Let C be a category such that for any two C-objects, C contains the product of these objects. We say C has *exponentiation*, if for any two C-objects, C contains the exponential object of these objects.

Let us now turn to the final property we need to understand in order to state in order to define topoi. This time, the property of **Set** that we will aim to emulate is the existence of a characteristic function χ_A for any given set A. Once again, this calls for several preliminary definitions. To gain a better understanding of the upcoming construction, let us first consider the situation in Set. Say we have been given a set S, a subset $A \subseteq S$ and a function f. Suppose we want to represent the fact that f is the characteristic function of A in a categorical fashion. A necessary condition is given by the demand that the diagram

$$\begin{tikzcd}
A \arrow[r, "1"] \arrow[d, tail, "i_A"'] & \{1\} \arrow[d, tail, "i_1"] \\
S \arrow[r, "f"'] & \{0,1\}
\end{tikzcd} \tag{B.10}$$

commutes, where 1 denotes the constant function from A onto $\{1\}$ and i_A, i_1 denote the inclusion maps from A and $\{1\}$ into S and $\{0,1\}$ respectively. This leaves open, however, the possibility of f being a function that sends some proper superset of A to 1 and the rest of S to 0. We then have $A \subseteq f^{-1}(\{1\})$, while what we want is the identity $A = f^{-1}(\{1\})$. We must thus look for some categorical criterion with which to enforce the latter.

Let $X = f^{-1}(\{1\})$ and consider the diagram

$$\begin{tikzcd}
X \arrow[r, "1"] \arrow[d, tail, "i_X"'] & \{1\} \arrow[d, tail, "i_1"] \\
S \arrow[r, "f"'] & \{0,1\}
\end{tikzcd} \tag{B.11}$$

What is the relation between diagrams B.10 and B.11? Since $A \subseteq X$, we have an inclusion map $j_A : A \to X$. Now, note that this inclusion map is the unique morphism in **Set** such that the diagrams

$$\begin{tikzcd}
& S & \\
X \arrow[ur, tail, "i_X"] & & A \arrow[ul, tail, "i_A"'] \arrow[ll, dashed, tail, "j_A"]
\end{tikzcd}
\qquad
\begin{tikzcd}
& \{1\} & \\
X \arrow[ur, "1"] & & A \arrow[ul, "1"'] \arrow[ll, dashed, tail, "j_A"]
\end{tikzcd} \tag{B.12}$$

commute. This should seem familiar. Indeed, it is turns out we can characterize B.11 and, hence, the concept of characteristic function as giving us the limit of the diagram

$$\begin{tikzcd} & \{1\} \arrow[d, tail, "i_1"] \\ S \arrow[r, "f"'] & \{0,1\} \end{tikzcd} \tag{B.13}$$

in the category **Set**. Limits for this type of diagram occur in many different places in category theory and are referred to as *pullbacks* and the resulting diagrams such as B.11 are known as *pullback squares*. With this, we can now characterize in a purely categorical fashion the notion of a characteristic function in **Set**. To be precise, the characteristic function χ_A for some $A \subseteq S$ is the unique morphism in **Set** such that

This characterization is now readily extended to arbitrary categories. The inclusion maps can simply be generalized to monic morphisms, the set $\{0,1\}$ can be replaced by an arbitrary object and the role of $\{1\}$ can be played by any terminal object. All that is left is for us to state the definition of a pullback in its full generality:

Definition B.2.5. Let C be a category and let $f : a \to c$, $g : b \to c$ be a pair of C-morphisms with a common codomain. A *pullback* for f and g is then a limit $g' : d \to a, f' : d \to b$ for the diagram

$$\begin{tikzcd} & b \arrow[d, "g"] \\ a \arrow[r, "f"'] & c \end{tikzcd} \tag{B.14}$$

The diagram

$$\begin{tikzcd} d \arrow[r, "f'"] \arrow[d, "g'"'] & b \arrow[d, "g"] \\ a \arrow[r, "f"'] & c \end{tikzcd} \tag{B.15}$$

is then referred to as the *pullback square* of the given pullback.

We now formulate what it means for a category to permit constructions that emulate characteristic functions. In the vernacular of category theory, such constructions are referred to as *subobject classifiers*.

Definition B.2.6. Let C be a category with at least one terminal object 1. A *subobject classifier* for C is a C-object Ω together with a C-morphism *true* : $1 \to \Omega$ such that for any monic morphisms $f : a \rightarrowtail d$ there exists exactly one C-morphism $\chi_f : d \to \Omega$ making the diagram

$$\begin{tikzcd}
a \arrow[r, tail, "f"] \arrow[d, "!"'] & d \arrow[d, "\chi_f"] \\
1 \arrow[r, "true"'] & \Omega
\end{tikzcd} \tag{B.16}$$

a pullback square. Here, we write '!' for the unique morphism from a to 1.

At long last, we are now able to give the definition of a topos. Since there exist two differing, widespread usages of the term *topos*, we will restrict ourselves, for the moment, to what are known as *elementary* topoi.

Definition B.2.7. A category is an *elementary topos* if it is finitely complete, finitely co-complete, has exponentiation and has a subject classifier.

Of course, there exists many other, equivalent definitions of the concept of elementary topos, the consideration of which would take us well beyond the scope of the present appendix.

With the concept of topos now defined, we are well-poised to consider more complicated topos-theoretic constructs such as the notions of classifying topos and pretopos completion. The explication of such notions, however, would require us to once again consider a host of prerequisite definitions and constructions, which could easily fill a score of additional pages. Therefore, let us be content here with the exposition presented above. Readers wishing to delve deeper into the trenches of topos theory may consult one of the several introductory works dedicated to the subject. For a particularly accessible introduction to the field of topos theory, see (Goldblatt 1984).

Bibliography

[1] Awodey, Steve and Henrik Forssell. 2013. “First-order Logical Duality.” *Annals of Pure and Applied Logic* 164 (3): 319–48.

[2] Balzer, Wolfgang, C. Ulises Moulines, and Joseph D. Sneed. 1987. *An Architectonic for Science: The Structuralist Program.* Dordrecht: D. Reidel Publishing Company.

[3] Barrett, Thomas William and Hans Halvorson. 2015. “Morita Equivalence.” ArXiv: https://arxiv.org/pdf/1506.04675.pdf

[4] Barrett, Thomas William, Sarita Rosenstock and James Owen Weatherall. 2015. “On Einstein algebras and relativistic spacetimes.” *Studies in History and Philosophy of Modern Physics* 52 (B): 309–16.

[5] Barwise, Jon and Solomon Feferman (eds). 1985. *Model-theoretic Logics.* New York: Springer-Verlag.

[6] Barwise, Jon K. 1974. “Axioms for Abstract Model Theory.” *Annals of Mathematical Logic* 7 (2–3), 221–65.

[7] Batterman, Robert W. 2010. “On the Explanatory Role of Mathematics in the Empirical Sciences.” *Brit. J. Phil. Sci.* 61 (1): 1–25.

[8] Beth, Evert W. 1960. “Semantics of Physical Theories.” *Synthese* 12 (2): 172–75.

[9] Beziau, Jean-Yves (ed). 2007. *Logica Universalis: Towards a General Theory of Logic.* Basel: Birkhäuser Verlag AG.

[10] Bueno, Otávio and Steven French. 2012. “Can Mathematics Explain Physical Phenomena?” *Brit. J. Phil. Sci.* 63 (1): 85–113.

[11] Bunge, Mario. 1959. *Metascientific Queries.* Springfield (IL): C.C. Thomas.

[12] Burstall, Rod and Joseph Goguen. 1992. “Institutions: Abstract Model Theory for Specification and Programming.” *Journal of the Association for Computing Machinery* 39 (1): 95–146.

[13] Colyvan, Mark. 2001. *The Indispensability of Mathematics.* Oxford: Oxford University Press.

[14] Contessa, Gabriele. 2006. “Scientific Models, Partial Structures and the New Received View of Theories.” *Studies in History and Philosophy of Science Part A* 37 (2): 370–7.

[15] De Rijke, Maarten. 1995. “A Lindström Theorem for Modal Logic.” In *Modal Logic and Process Algebra: A Bisimulation Perspective*, Maarten de Rijke, Alban Ponse and Yde Venema (eds), Stanford (CA): CSLI Publications.

[16] Dewar, Neil. 2017. “Towards a Metaphysics of Categorical Equivalence.” Workshop presentation. Erasmus University Rotterdam.

[17] Diaconescu, Răzvan, Till Mossakowski and Andrzej Tarlecki. 2014. “The Institution-Theoretic Scope of Logic Theorems.” *Logica Universalis* 8 (3–4): 393–406.

[18] Diaconescu, Răzvan. 2008. *Institution-Independent Model Theory.* Basel: Birkhäuser Verlag AG.

[19] Dieks, Dennis. 2010. “E.W. Beth as a Philosopher of Physics.” *Synthese* 179 (2): 271-284.

[20] Ebbinghaus, Heinz-Dieter. 1985. “Extended Logics: The General Framework.” In *Model-Theoretic Logics*, Jon Barwise and Solomon Feferman (eds), New York (NY): Springer-Verlag.

[21] Enqvist, Sebastian. 2013. “A General Lindström Theorem for Some Normal Modal Logics.” *Logica Universalis* 7: 233–64.

[22] Feferman, Solomon. 1974. “Two notes on abstract model theory I: Properties invariant on the range of definable relations between structures.” *Fundamenta Mathematicae* 82: 153–65.

[23] Flum, Jörg. 1985. “Characterizing Logics.” In *Model-Theoretic Logics*, Jon Barwise and Solomon Feferman (eds), New York (NY): Springer-Verlag.

[24] Goldblatt, Robert. 1984. *Topoi: The Categorical Analysis of Logic*. Amsterdam: North-Holland.

[25] Halvorson, Hans and Dimitris Tsementzis. 2016. “Categories of scientific theories.” Preprint. PhilSci Archive: http://philsci-archive.pitt.edu/11923/2/Cats.Sci.Theo.pdf

[26] Halvorson, Hans. 2012. “What Scientific Theories Could Not Be.” *Philosophy of Science* 79 (2): 183–206.

[27] Hanson, Norwood Russell. 1962. ‘The Irrelevance of History of Science to the Philosophy of Science.’ *The Journal of Philosophy* 59 (21). 574-86.

[28] Hempel, Carl G. and Paul Oppenheim. 1948. “Studies in the Logic of Explanation.” *Philosophy of Science* 15 (2): 135–75.

[29] Lakatos, Imre. 1970. “History of Science and Its Rational Reconstructions.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1970. 91–136.

[30] Lindström, Per. 1969. “On Extensions of Elementary Logic.” *Theoria* 35, 1-11.

[31] Lloyd, Elisabeth Anne. 1994. *The Structure and Confirmation of Evolutionary Theory.* Princeton (NJ): Princeton University Press.

[32] Ludwig, Günther, and Gérald Thurler. 2006. *A New Foundation of Physical Theories.* Berlin: Springer.

[33] Lutz, Sebastian. 2012. “On a Straw Man in the Philosophy of Science: A Defense of the Received View.” *The Journal of the International Society for the History and Philosophy of Science* 2 (1): 77–120.

[34] Montague, Richard. 1974. “Deterministic Theories.” In *Formal Philosophy: Selected Papers of Richard Montague*, ed. Richmond H. Thomason, 303-59. New Haven: Yale University Press.

[35] Muller, F.A. 2011. “Reflections on the Revolution at Stanford.” *Synthese* 183 (1): 87–114.

[36] Muller, F.A. 1998. *Structures for Everyone.* A. Gerits & Son.

[37] Otto, Martin and Robert Piro. 2008. ‘A Lindström Characterisation of the Guarded Fragment and of Modal Logic with a Global Modality.’ Nancy: *Advances in Modal Logic.* Conference paper. Available online.

[38] Pearce, David. 1985. *Translation, Reduction and Equivalence: Some Topics in Intertheory Relations.* Frankfurt: Verlag Peter Lang GmbH.

[39] Pearce, David, and Veikko Rantala. 1983a. “New Foundations for Metascience.” *Synthese* 56 (1): 1–26.

[40] Pearce, David, and Veikko Rantala. 1983b.“Constructing General Models of Theory Dynamics.” *Studia Logica* 42 (2): 347–62.

[41] Pincock, Christopher. 2001. “A New Perspective on the Problem of Applying Mathematics.” *Philosophia Mathematica* 12 (2): 135–61.

[42] Przełecki, Marian. 1969. *The Logic of Empirical Theories.* London: Routledge and Kegan Paul.

[43] Rantala, Veikko. 1978. “The Old and the New Logic of Metascience.” *Synthese* 39 (2): 233–47.

[44] Rizza, Davide. 2013. “The Applicability of Mathematics: Beyond Mapping Accounts.” *Philosophy of Science* 80 (3): 398–412.

[45] Sahlqvist, Henrik. 1975. ‘Completeness of Correspondence in the First and Second Order Semantics for Modal Logic.’ In *Proceedings of the Third Scandinavian Logic Symposium*, Stig Kanger (ed), Amsterdam: North-Holland Publishing Company.

[46] Scheibe, Erhard. 1979. “On the Structure of Physical Theories.” In *The Logic and Epistemology of Scientific Change*, eds. Ilkka Niiniluoto, and Raimo Tuomela, 205-24. Amsterdam: North-Holland Publishing Company.

[47] Schmidt, Heinz-Juergen. 2014. “Structuralism in Physics.” *Stanford Encyclopedia of Philosophy.*

[48] Sneed, Joseph D. 1994. “Structural Explanation.” In *Patrick Suppes: Scientific Philosopher, Volume 2*, ed. Paul Humphreys, 195–216. Dordrecht: Kluwer Academic Publishers.

[49] Sneed, Joseph D. 1971. *The Logical Structure of Mathematical Physics.* Dordrecht: D. Reidel Publishing Company.

[50] Stegmüller, Wolfgang. 1976. *The Structure and Dynamics of Theories.* Berlin: Springer-Verlag.

[51] Suppe, Frederick. 1989. *The Semantic Conception of Theories and Scientific Realism.* Urbana: University of Illinois Press.

[52] Suppe, Frederick. 1977. *The Structure of Scientific Theories.* Urbana: University of Illinois Press.

[53] Suppes, Patrick. 1967. “What is a Scientific Theory?” In *Philosophy of Science Today*, ed. Sidney Morgenbesser, 55–67. New York: Basic Book, Inc., Publishers.

[54] Teh, Nicholas J. and Dimitris Tsementzis. 2015. “Theoretical Equivalence in Classical Mechanics and its Relationship to Duality.” PhilSci-Archive: http://philsci-archive.pitt.edu/11695/

[55] Thompson, Paul. 1989. *The Structure of Biological Theories.* Albany (NY): SUNY press.

[56] Tsementzis, Dimitris. 2015. “A Syntactic Characterization of Morita Equivalence.” ArXiv: https://arxiv.org/pdf/1507.02302.pdf

[57] Van Benthem, Johan. 2010. *Modal Logic for Open Minds.* Stanford: CSLI Publications.

[58] Van Benthem, Johan. 2007. “A New Modal Lindström Theorem.” *Logica Universalis* 1, 125–38.

[59] Van Fraassen, Bas C. 1989. *Laws and Symmetry.* Oxford: Clarendon Press.

[60] Van Fraassen, Bas C. 1970. “On the Extension of Beth’s Semantics of Physical Theories.” *Philosophy of Science* 37 (3): 325–39.

[61] Van Oosten, Jaap. 2002. “Basic category theory.” Lecture notes. https://www.staff.science.uu.nl/∼ooste110/syllabi/catsmoeder.pdf

[62] Visser, Albert. 2004. “Categories of Theories and Interpretations.” Logic Group Preprint Series (volume 228), Utrecht University. https://dspace.library.uu.nl/handle/1874/26909

[63] Vos, Bobby. 2016. “Model-Theoretic Approaches to Universal Logic: A Modal-Logical Perspective.” Tutorial paper. Available on request.

[64] Vos, Bobby. 2015a. “The Model-Theoretic Structure of Greek Nature-Knowledge.” Final essay. Course: Philosophy of Nature. Available on request.

[65] Vos, Bobby. 2015b. “On an Extension of Van Fraassen’s Semantics of Physical Theories.” Final essay. Course: Seminar Philosophy of Physics. Available on request.

[66] Vos, Bobby. 2014. “Lindström’s Characterizations of First-Order Logic.” Bachelor’s thesis. Available on request.

[67] Weatherall, James Owen. 2016. “Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent?” *Erkenntnis* 81 (5): 1073–91.