Universidade de São Paulo
Faculdade de Filosofia Letras e Ciências Humanas
Departamento de filosofia

Explicating logicality

Dissertação de mestrado apresentada ao Departamento de Filosofia

Candidato:
Daniel A. Nagase
Orientador:
Rodrigo Bacellar da Costa e Silva

São Paulo
2017

Daniel Arvage Nagase

Explicating logicality

Dissertação apresentada ao Programa de Pós-Graduação em Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, para obtenção do título de Mestre em Filosofia sob a orientação do Prof. Dr. Rodrigo Bacellar da Costa e Silva

São Paulo
2017

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

N147e

Nagase, Daniel Arvage
Explicating Logicality / Daniel Arvage Nagase ; orientador Rodrigo Bacellar da Costa e Silva. - São Paulo, 2017.
182 f.

Dissertação (Mestrado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Filosofia. Área de concentração: Filosofia.

1. Lógica. 2. Metafísica. 3. Filosofia da matemática. 4. Tarski. 5. Carnap. I. Costa e Silva, Rodrigo Bacellar da, orient. II. Título.

Nagase, D. A. **Explicating Logicality**. 2017. Dissertação. Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia, Universidade de São Paulo, São Paulo, 2017.

Aprovado em:

Banca Examinadora

Prof. Dr. Rodrigo de Alvarenga Freire Instituição: Universade de Brasília
Julgamento: _____ Assinatura: _____
Prof. Dr. Marco Antonio Caron Ruffino Instituição Universidade de Campinas
Julgamento: _____ Assinatura: _____
Prof. Dr. Edélcio Gonçalves de Souza Insituição: Universidade de São Paulo
Julgamento: _____ Assinatura: _____

To the departed:
Carlos, Sônia, Miguel, and Naides

Acknowledgments

It is a pleasure to start this "thank you" note by expressing my gratitude towards my supervisor, Rodrigo Bacellar. When I first approached him with the idea of working under his supervision, I was interested in an inferentialist framework, couched in proof-theoretic terms and close to intuitionism. After some talk, Rodrigo suggested that I should broaden the horizons of my research and consider other approaches to the logicality question. By following his advice and example, I ended up defending in this dissertation a version of platonism—neo-Fregean, to be sure, which I hope is not too exotic for his tastes. I also owe him some of the impetus to register a bit of my own thoughts on some matters; I still vividly remember when, after approaching him with an idea for a historical article, he said "But what about *Nagase*'s ideas?". It is hard to overstate the enormous influence his thoughts and research program have exerted on my own philosophical development.

Second, I wish to express my warmest thanks to Edélcio Gonçalves. Edélcio started working at USP at roughly the same time when I started this dissertation. Since then, I have learned much from his classes, seminars (much of my current mathematical maturity was developed by studying model theory for these seminars), and from the much cherished chats in his office about mathematical matters and other amenities. He was also in my examination committee, offering valuable advice that I unfortunately have not always followed, to my own peril.

Like Edélcio, Rodrigo Freire was also in my examination committee, giving helpful suggestions that I am again not sure of having thoroughly followed. I have always admired his technical expertise and philosophical acumen, as well as his readiness to share his vast knowledge. The notes which I have taken during his course on forcing are a small treasure that I still consult from time to time.

I have met with Alexandre Costa Leite, Freire's colleague at the Universidade de Brasília, only a few times, but such times were, shall we say, *memorable*. On those occasions, he always expressed interest in my work and encouraged me to go further. I am very grateful for this encouragement and even more for the valuable time spent together.

If it were not for my friends Felipe Salvatore and Pedro Falcão, I would probably never have started studying logic, much less completed this dissertation. Both Felipe and Pedro

encouraged me to change areas and warmly recommended me to Rodrigo Bacellar. They also mentored me in my early stages, helping me to overcome the difficult entrance barrier to a more mathematical way of thinking. I fondly remember pestering Felipe with doubts about this or that exercise, as well as the hours Pedro and I spent tackling McGee's difficult theorem about invariant operations and $\mathcal{L}_{\infty,\infty}$.

I must also thank my colleagues at Edélcio's Tuesday seminars: Diogo Dias, Lucas Baccarat, for suffering through my early attempts at model theory (and especially through my presentation of Ehrenfeucht–Fraïssé games). This year we have been joined by Guilherme Lima and Matheus Cury, who are suffering through my group theory musings. Again, thank you for your patience.

More recently, Lucas Mussnich and Tiago Royer have been invaluable studying partners. Although we are still at the beginning of Lang's famous *Algebra*, the amount I have learned from tackling this mammoth with these two is incalculable. This study group was moreover a unique opportunity to reunite with an old friend—Lucas and I studied together from elementary to high school.

Lucas and Tiago, together with André Coggiola, Pedro Falcão, Victor Coelho, and Victor Moraes, were also part of a mini-seminar that I ran about Gödel's incompleteness theorems. They patiently endured my hopelessly confusing explanations about the nested interval property, Cantor's set, the Fueter–Pólya theorem, and course-of-values recursion. I hope they have managed to retain *something* from this experience, especially since it was so much fun.

In the age of the Internet, it is much easier to get help when you are stuck in a technical problem than before. I am *extremely* grateful for the Math StackExchange and Mathoverflow communities for the immense help they gave me in answering my (sometimes obtuse) questions. In particular, I would like to thank Asaf Karagila, Alex Kruckman, and Noah Schweber for taking their time to guide me through, to use Borges's expression, the *delicate labyrinth* of set theory and model theory, with its "vast numbers that an immortal man would not reach even if he exhausted his eternities counting" and "whose imaginary dynasties have the letters of the Hebrew alphabet as cyphers". It is fitting that some of the keys to this labyrinth should lie in an invisible network with billions of intangible nodes, one that would surely have pleased Borges's aesthetic imagination.

The Internet has also made communications between scholars easier, which benefited me immensely. Numerous scholars have helped me by electronic communication. Special mention goes to Catarina Dutilh Novaes, whose encouragement I much appreciate; Paolo Mancosu, for sharing a draft of *Abstraction and Infinity* with me; Richard Heck, Jr., for not only sharing numerous papers with me, but also engaging in a stimulating discussion about abstract objects; Bob Hale, for searching the depths of his computer for a copy of his in-

valuable book on *Abstract Objects* for me; Gila Sher, for sending me her work on *The Bounds of Logic*; Sam Cowling, for sharing a chapter draft with me and also commenting on some of my incipient thoughts concerning the distinction between abstract and concrete objects; Kenneth Manders, for sharing his manuscript on the absoluteness of first-order logic; Jouko Väänänen, for some crucial help with his own proof of the absoluteness of first-order logic.

On a more personal level, I must thank my friends Fábio Franco, Dario de Negreiros, Rafael Schincariol, Clara Figueiredo, Catarina Pedroso, and Lucas Paolo for keeping me down to earth. Our discussions and activities are an integral part of my life, and give me a glimmer of hope, much needed in these times of political turmoil in our country.

Similarly, I would like to thank Daniel Lago Monteiro and Ana Letícia Adami Batista for the many enjoyable hours discussing literature, cinema, music, and other serious matters. The former is also an esteemed colleague who has taught me much about teaching English as a second language, while the latter is not only an esteemed colleague but also an esteemed student, whom I hope not to have entirely spoiled.

Adriano Correia and I started university together, and have since then had numerous discussion on metaphysics, Plato, Aristotle, Deleuze, textbooks, Bee Gees, and much else. I also owe him apologies, since I have always accepted his invitations for blogging together only to late frustrate him with my lack of participation. Perhaps next time?

Other friends, such as Alexandre Lindenberg, Thor Ribeiro, Davi Mamblona, and Victor Zellmeister were crucial during certain transition periods of my life, especially when I had just entered the university. They have made me see that there was a life out there to be lived, for which I am much grateful.

Beginning at the end of last year, Carlos Gallucci has also proven to be an inestimable friend. His integrity and forthrightness, as well as professionalism and kind heart are an example to us all. I still hope to go to a rock concert with him one day.

On a still more personal note, my extended family, Ana Paula, Marina, Ary, and Victor have welcomed me with arms wide open and full of cats (and, recently, even *two* dogs!). I hope they get as much feline love out of this relationship as I get.

My family has been incredibly supportive of my decisions so far. My uncles, Miguel and Paulo, as well as their respective spouses, Lilian and Simone, raised me, nurtured me, counseled me; they basically made possible that I could keep going after much struggle.

Finally, but definitely not least, my infinite gratitude to the two most important important women in my life, namely my mother, Sandra, and my wife, Olívia. It is to my mother that I owe my discipline, my capacity for concentration, my ethical compass, even my literacy. She encouraged me at every step (though I think I heard a dint of disappointment when I first announced that I was going to study philosophy at the university...), and, through her own example of perseverance, has taught me much about what it is to keep your head up in

the face of ever increasing challenges. As for Olívia, what can I say? I am lucky enough to have her by my side, filling every moment of our marital life with unending joy (and yes, this is an accurate description of our marital life). *Thank you!*

Oh, and, of course, I am grateful to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the research grant that made this dissertation possible!

Resumo

Nagase, Daniel Arvage. *Explicating Logicality*. 2017. Dissertação (Mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia, Universidade de São Paulo, São Paulo, 2017.

O presente estudo tem por objetivo analisar a assim chamada *proposta de Tarski*, a qual visa fornecer uma resposta à pergunta: quais objetos são *lógicos*? Nossa análise consiste em duas partes: uma primeira, mais histórica, compara a metodologia de Tarski àquela de Carnap e de Quine, se atentando principalmente às diferentes acepções que cada um deles atribui à noção de *explicação* (*explication*). A segunda parte, mais argumentativa, procura mostrar que um ambiente natural para essa proposta é uma metafísica platônica de franca inspiração neo-fregeana.

Palavras-chave: Tarski – Logicalidade – Explicação – Carnap – Quine – Platonismo

Abstract

Nagase, Daniel Arvage. *Explicating Logicality*. 2017. Dissertation (Master Degree) – Faculdade de Filosofia, Letras e Ciências Humanas. Departamento de Filosofia, Universidade de São Paulo, São Paulo, 2017.

The present study aims at analyzing the so-called *Tarski proposal*, a proposal about which objects should be considered as logical. My analysis has two parts: the first part, more historically oriented, compares Tarski's evolving methodology to Carnap's and Quine's, in particular with the different conceptions of these latter two regarding that which they called *explication*. The second, more argumentative part, attempts to show that the most natural environment for this proposal is a platonic metaphysics of a neo-Fregean variety.

Key words: Tarski – Logicality – Explication – Carnap – Quine – Platonism

Contents

Introduction

This study has been occasioned by my reading of a famous lecture by Tarski (1966/1986).[1] It is not, however, a close reading or a detailed analysis of this text. Rather, it is best thought of as a collection of reflections prompted by my engagement with this text. Two of these pieces are more historical in nature, whereas in the other two I allow myself to indulge in a bit of philosophical fancy. This is reflected in the organization of this dissertation: the first part contains the two historical pieces, and the second part contains the more indulgent reflections. The reader should be aware that such indulgence does not necessarily pay off, in order not be disappointed by some of the meager results here collected.

The first chapter of the historical part was occasioned not so much by Tarski's own piece, as by the way the relevant literature has constantly *ignored* many of its key statements. Although I am one of the first to agree that the tales of Tarski's legendary clarity and precision are greatly exaggerated, he is nonetheless particularly clear about two things in this lecture: his proposal should not be judged as a kind of description of some platonic essence and the target of his proposal are logical objects. Unfortunately, much of the literature proceeds as if he was proposing just some such description and often takes as his target for analysis something other than logical objects. Therefore, I proposed in this first chapter to study some parallels between Tarski's methodological remarks in this piece with the more famous methodology of *explication* as devised by Carnap, in the hopes of contributing to the dispelling of the first confusion. In the course of researching for this chapter, I was also surprised to discover that there were fundamental differences between Tarski's methodological outlook in his famous 1936 paper on "The Concept of Truth in Formalized Languages" and later papers. Given that the story of the evolution of Tarski's views on this matter has not yet been told, I decided to include this material in the chapter as well.

The second chapter in a sense grew out of the first. During my research on Tarski's philosophy, I discovered that not only was he a committed nominalist, but that this nominalism was expressed in his rejection of higher-order logic. This created a puzzle: the most natural

[1] Tarski's lecture is from 1966, but it was only published in 1986 by John Corcoran. As noted by Feferman (1999), Mautner (1946) can be considered as a kind of forerunner to Tarski's proposal, though with important differences. Mautner's article has not been much discussed in the literature.

setting for the proposal of the 1966 lecture is higher-order logic, in particular type-theory. How then can one reconcile this setting with Tarski's nominalism? The chapter explores a possible answer to this solution, one inspired in part by the more developed thoughts one finds in Quine's writings.

Conceptually, then, Tarski seems to be located between Carnap and Quine, sometimes drawing from one, sometimes from the other. Historically, however, it may be that it is Carnap and Quine who drew much inspiration from these conflicting strands found in Tarski's private philosophy, which would thus justify an in-depth analysis of his rather inchoate philosophical remarks—if nothing else, as an important prelude to what has been considered as one of the most important debates of analytic philosophy.

The third chapter, then, which is the first of the more indulgent part, analyzes Tarski's proposal itself. The indulgence consists in my attempt to read his proposal against the background of a platonic metaphysics, thus going completely against Tarski's own nominalist inclinations. The result is that I give much greater weight to his reference to Klein than most critics, using this framework to timidly suggest an alliance with the neo-Fregeans over these matters. The suggestion is timid, since no more than a brief sketch of the idea is offered—clearly one of the places in which the reader's hopes for a more fully developed theory will be disappointed.

The final, short chapter—more a coda to the third chapter than an autonomous piece—could be considered as an attempt at selling off my metaphysical approach to an unfortunately naive customer who does not care to check all the details of the product he is buying. To wit, I try to show some of the advantages of the poorly sketched metaphysical picture from the previous chapter, by showing that it can at least resist criticism from some of its most prominent competitors. Perhaps out of shame for such lowly commercial tactics, I try also to provide the reader with some extra stuff he did not ask for, in the form of three lengthy technical appendices—another one of my indulgences.

A word, then, about the appendices. First, I decided to append them to *each* chapter, instead of collecting them at the end, because some of them were really short and contained just whatever technical results were mentioned in the main text. Second, I decided to *append* them to each chapter, instead of including them in the main text, in order not to clutter the exposition. That is not to say that they are uninteresting or unimportant. But, this being a *philosophical* work after all (however poorly it performs in this task), it seemed fitting to devote the main text to *philosophical* points, and hence to leave the juicy mathematical bits to appendices. Their length is in part because of my desire to make them as self-contained as possible: a more mathematically inclined reader could get some enjoyment out of reading *just* the appendices, referring to the main text just for motivation.

As for the conclusion, since each chapter has its own conclusion, instead of presenting a

kind of wrap-up of the preceding, it rather points forwards, into directions of research that were not fully developed in this study. It can be thus thought as a series of promissory notes, which I hope to one day acquire the necessary funds to honor, if the reader ever decide to cash them out.

PART I

Tarskian Explication

Chapter 1

Tarski's Conceptual Analyses

In the course of his career, Tarski proposed three conceptual analyses that would prove to be enormously influential: his analyses of truth, of logical consequence, and of "logical notions". Although proposed at different moments in his career, these analyses are generally treated as obeying the same overall logic[1]: starting from a given intuitive notion, Tarski would proceed to give formal analogs to them, always obeying the twin criteria of "material adequacy" (glossed as extensional agreement) and "formal correctness" (glossed as consistency of the resulting theory, along with other formal requirements, e.g., on the form of the definition). Not surprisingly, there is a lot of controversy regarding whether or not Tarski succeeded in capturing the intuitive counterparts of his *analysans*; in particular, Tarski has been charged of failing his "material adequacy" criterion, both by overgenerating (capturing too much) and by undergenerating (capturing too little). Most famous among these is probably Etchemendy's claim that:

> (...) Tarski's analysis is wrong, that his account of logical truth and logical consequence does not capture, or even come close to capturing, any pretheoretic conception of the logical properties. (...) Applying the model-theoretic account of consequence, I claim, is no more reliable a technique for ferreting out the genuinely valid arguments of a language than is applying a purely syntactic definition. Neither technique is guaranteed to yield an extensionally correct specification of the language's consequence relation. (Etchemendy 1999, p. 6)

Indeed, even an overall sympathetic account such as Patterson (2012) agrees with Etchemendy, saying that "the modal and extensional failings of the analysis [of logical consequence—D.N.] are manifest" (Patterson 2012, p. 220).

The aim of this chapter is to show that this standard story is incorrect. Specifically, I'll argue for two claims: (1) Tarski's groundbreaking analyses do not all obey a unified procedure,

[1] Cf., e.g., Dutilh Novaes and Reck (2015, Section 1.1), which groups together the analysis of truth and the analysis of consequence as a kind of "proto-Carnapian" explication.

but rather they can be distinctly grouped in two separate phases, corresponding, briefly, to his early acceptance and later abandonment of "intuitionistic formalism"; (2) Tarski's mature work, beginning with his analysis of logical consequence, does not intend to capture some kind of intuitive concept, but rather to *replace* a vague concept with a formally defined one. In fact, given (2), it seems there are some similarities between Tarski's procedure and Carnap's. Since Carnap is much more forthright about his project than Tarski, and since there is already an extensive literature on his concept of explication[2], there is some hope that a comparison with Carnap may help to shed light on Tarski's own procedure.

Accordingly, this chapter is structured as follows. First, I'll give a brief account of Carnap's notion of explication, contrasting it with what we may call the "classical picture" of conceptual analysis. Second, building on Patterson (2012), I'll analyze Tarski's approach to conceptual analysis in his famous truth paper, showing how it conforms to the central tenets of "intuitionistic formalism". Finally, I'll show how Tarski conceives his task in his paper on logical consequence and other articles written after his abandonment of "intuitionistic formalism" in a way that resembles Carnapian explication. My central claim is thus that Tarski moved from a decidedly un-Carnapian view of conceptual analysis towards an account that is broadly Carnapian in spirit.

1.1 Carnapian Explication

In his book *Logical Foundations of Probability*, Carnap characterizes "explication" in the following manner:

> According to these considerations, the task of explication may be characterized as follows. If a concept is given as explicandum, the task consists in finding another concept as its explicatum which fulfills the following requirements to a sufficient degree:
>
> 1. The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
> 2. The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explicatum into a well-connected system of scientific concepts.

[2] Cf. in particular the book-length study by Carus (2007) and the collection of essays by Wagner (2012).

3. The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, theorems in the case of a logical concept).
4. The explicatum should be as *simple* as possible; this means as simple as the more important requirements (1), (2), and (3) permit. (Carnap 1962, p. 7)

In order to understand this better, let's focus on Carnap's own example of a successful explication, namely the scientific refinement of the pre-scientific concept of "fish". As reflected, e.g., in Linnaeus classification from 1735, most animals that lived under water were classified as fish, so that whales, dolphins, etc., were considered as fish, in spite of their similarities to mammals being known since at least Aritstotle. However, by 1758 Linnaeus refined his classification and included a new type, the "Mammalia", under which we could now find whales, dolphins, etc., with the appropriate restriction of the category of "pisces".[3] Thus, the pre-scientific concept of *fish*, meaning roughly *animal living underwater*, gave way to the scientific concept of *pisces*, meaning roughly *sharing anatomical features such as the presence of gills, etc.* (note that the pre-scientific concept of *quadruped* also gave way to the more refined scientific concept of *mammal*). Taking then the concept of *fish* as the explicandum and the concept of *pisces* as the explicatum, we can then see how this proposed explication sheds light on the Carnapian requirements outlined above.

First, there is a lot of overlap between *fish* and *pisces*, so that the new concept passes the similarity test. *Pisces* is also an exact concept, since, in order to be a pisci, a specimen must possess a number of features exactly specified. Being anatomically more precise, the new concept is also much more fruitful, especially since it connects taxonomy to other branches of biology. It's also relatively simple, as features which are not common to all *pisces* are not mentioned in the classification. Therefore, we can identify a certain *minimal* conception of explication:

Minimal explication: Given a concept C as an *explicandum*, a *successful minimal explication* for C is another concept C' which fulfills the similarity, exactness, fruitfulness, and simplicity criteria.

This conception of explication is already sufficient to distinguish it from the more traditional "conceptual analysis", which generally requires some kind of content identity between analysans and analysandum.[4] In contrast to this identity, it's not necessary for an

[3] In this he was apparently influenced by John Ray, who had already strongly opposed the classification of whale as fish. Cf. the instructive quotation from Ray in Raven (2009, p. 366), which can also be taken as a good contrast between what Ray calls strict and philosophical use of a concept, on the one hand, and common usage, on the other hand.

[4] For the more traditional conceptual analysis, specially during the early phase of analytic philosophy, cf.

explication to be successful that explicatum and explicandum be identical, only that they be similar (and, as we will see, even this requirement will turn out to be a loose one). Nevertheless, it's still *minimal* in that it allows very different construals, depending on the philosophy in which is embedded. So, for instance, a more realist inclined philosopher may also require that the explication somehow track the essence of the phenomenon in question, whereas someone impressed by anti-realist considerations may eschew this extra requirement altogether.[5] There are also possible differences in the purposes to which someone may put explication: Quine, for instance, thinks that one of the main tasks of explication is to reduce ontological commitment, which certainly is not among Carnap's own worries.[6] In the next sections, I'll show how this minimal idea is developed by Carnap into a more full-blown *Carnapian explication* by focusing first on a puzzle related to the similarity requirement.

1.1.1 A Puzzle about Similarity

Let's take a deeper look at the biological example introduced by Carnap in order to better understand the similarity criterion. Clearly the two concepts, *fish* and *pisces*, differ in intension, so this type of similarity can't be what Carnap is requiring. On the other hand, when commenting on their dissimilarity, Carnap says that the concept of *pisces* "is much narrower than the [concept of *fish*]; many kinds of animals which were subsumed under the concept Fish, for instance, whales and seals, are excluded from the concept Piscis" (Carnap 1962, p. 6).[7] So it seems that, by similarity, Carnap has in mind extensional adequacy. Nevertheless, as this example itself makes clear, one should not expect complete extensional adequacy for a concept to be similar to another. So how much deviance should one allow? Specifically, is there any kind of "conceptual core", be it in the form of important conceptual features or of paradigmatic instances of the explicandum, which should be preserved?

Unfortunately, Carnap is not very explicit on this matter; as far as I know, this was simply not discussed by him. Indeed, there appears to be a kind of primacy of the fruitfulness criterion over the similarity criterion, insofar as the latter is clearly determined by the former. That would explain why Carnap slips so easily from one to the other:

> The [concept of *fish*] has been succeeded by [the concept of *pisces*] in this sense: the former is no longer necessary in scientific talk; most of what previously was said with the former can now be said with the help of the latter (though often

the essays collected in the volume edited by Beaney (2007). Cf. also section 1.1.2 for more on the contrast between conceptual analysis and explication.

[5] Note that this is independent of the similarity criterion: the similarity in question is between explicatum and explicandum, not between the explicatum and reality.

[6] I'll say more about Quine's own version of explication in the next chapter.

[7] Note that "piscis" is the singular of "pisces".

> in a different from, not simply by replacement). (...) What was their motive for (...) artificially constructing the new concept Piscis far remote from any concept in the prescientific language? The reason was that they realized the fact that the concept Piscis promised to be much more fruitful than any concept more similar to Fish.

Notice how Carnap answers a question about the similarity criterion (why not choose a concept that tracks more closely the extension of the explicandum) with considerations about fruitfulness. This is in agreement with Dutilh Novaes and Reck (2015), who also identify this same primacy operating in Carnap. As they note, however, this creates a problem for Carnap. Since fruitfulness demands that the new concept shed light on phenomena not explained by the previous concept, and this entails that there will be a mismatch of some sort between the two concepts, it follows that the more we push towards fruitfulness, the more we push away from similarity. So, if we consider similarity a separate requirement that should be respected at least in some core cases, there seems to be a tension between the two criteria. That's why they go so far as to call Carnap's explication project "inherently paradoxical":

> In addition, if what accounts for the fruitfulness of an explication is such a mismatch between explicandum and explicatum, then it becomes clear why fruitfulness is at odds with the requirement of similarity. As is evident now, a less-than-perfect degree of similarity between explicandum and explicatum is not only a tolerable, contingent upshot of explication—it is the *very goal* of a fruitful explication. This in turn implies that, insofar as it is accountable both to similarity and to fruitfulness, Carnapian explication is an inherently paradoxical enterprise. (Dutilh Novaes and Reck 2015, Section 1.5)

Of course, that may be putting matters too strongly. There may be some cases in which fruitfulness does not inevitably pull away from similarity considerations. To use a rather contentious example, but which may help to drive the point home, consider the so-called "squeezing" arguments.[8] This type of argument generally proceeds in the following way. Suppose we're given an informal notion I, which we want to analyze as S. Since I is informal, we may not be able to give a direct proof that I and S agree extensionally; nevertheless, the concept I may be sufficiently precise that we may confidently state that S is a sufficient condition for I, i.e. every instance of S is also an instance of I. Moreover, we may also be able to isolate a precise necessary condition for I, say N, so that I's extension is *squeezed*

[8] This type of argument was famously developed by Kreisel (1967) and has recently become a renewed source of interest. For some discussion, cf. Field (2008, chap. 2), Smith (2011), Dutilh Novaes and Andrade-Lotero (2012), and Smith (2013, chap. 45).

between the extension of S and the extension of N. Finally, since S and N are both precise, it may be possible to show that they're co-extensional, in which case we also have that S and I are co-extensional. Here's an example, taken from Smith (2013, chap. 45): we're given the informal notion of effective computability, and we want to analyze it in terms of Turing computability. It's not difficult to see that every Turing computable function is effectively computable; in order to obtain the converse, Smith introduces a series of conditions which collectively define what he calls Kolmogorov-Uspenskii (KU) computability (the exact nature of these conditions are not important for us here). He then argues that every effectively computable function is KU computable and finally that every KU computable function is Turing computable, completing the squeezing argument. If successful, this would amount to a vindication of the Turing-Church Thesis. The details of the argument are not really relevant; what matters is that this type of argument should not be ruled out on principle, and hence that our concept of explication needs to leave some room for it. Accordingly, it may be the case that fruitfulness and similarity are not *inherently* in conflict.

One obvious rejoinder to this reasoning is that it presupposes that extensional agreement is the only relevant measure of similarity, which may not be the case. Again, Carnap is not very explicit about this, and his own example seems to point towards this direction, but that doesn't mean extensional agreement is the only measure available. Depending on your favorite semantic theory, it may be possible to think about some kind of partial intensional overlap. Dutilh Novaes and Reck also bring up "continuity of purposes", which would in fact be more in line with Carnap's pragmatism. Regardless of the measure chosen, however, it seems to me that there will be cases, such as the above, which won't deviate much (if at all) from the original concept, while at the same time being fruitful.

In any case, that doesn't affect Dutilh Novaes and Reck's main point, which is that in many cases the requirement of fruitfulness will be in conflict with the similarity requirement, thus generating a tension within Carnap's account. Carnap himself, however, seemed to be unconcerned with this. In order to understand why, it will be useful to take a closer look at his pragmatism.

1.1.2 Carnap's Pragmatism

A remark by his student, Howard Stein, may point us to the right direction. Stein reminds us that "*any* question about the relation of the explicatum to the explicandum is an 'external question'; this holds, in particular, of the question whether an explication is adequate" (Stein 1992, p. 281). This points us to "Empricism, Semantics, and Ontology", an article where Carnap (1950/1956) distinguishes between internal or theoretical questions that can be decided within the framework of a given science, and *external* questions about which framework to adopt; the latter can only be solved pragmatically, in accordance with each

context.[9] Perhaps a contrast with what Wilson (2006) calls the "classical picture of concepts" may help to highlight some of the work this distinction is doing here.[10]

According to Wilson, the "classical picture of concepts" maintains that each concept possess a certain core, which gives it its unity:

> In the classical tradition, the conceptual content associated to a predicate—the same stuff that binds it to the world—is intended to serve as an *invariant core* that controls the instructive directivities that attach to the predicate. As explained before, I employ "directivity" as a non-technical means for capturing the loose bundle of considerations that we might reasonably cite, at various moments in a predicate's career, in deciding how the term should be *rightly applied*. (Wilson 2006, p. 95)

As Wilson emphasizes, it is this belief that each concept carries with it a kind of invariant core that underwrites the associate belief that exercises of conceptual analysis should bring this core to the fore, eliminating inessential elements which may get into the way of our fully grasping it. Something like this belief seems to also underwrite Dutilh Novaes and Reck's diagnosis of an "inherent paradox" in Carnap's explication project. If, as they say, "some weaker form of 'faithfulness' or 'matching' remains relevant" (Dutilh Novaes and Reck 2015, section 3) for the success of an explication, that's because there must be some kind of invariant core in the explicandum to which the explicatum must remain faithful.

Going back to the distinction between internal and external questions, since the explicandum is generally not part of the scientific framework of the explicatum (otherwise, there would be no need for explication), the question of how similar they are is not an internal question, but an external one. It's a question of *which framework to adopt*, and, as such, it is to be solved on pragmatic grounds. To go back to our taxonomic example, which selection of core features produces the most fruitful results, selecting anatomic or phylogenetic features? This obviously depends on the purposes of our classification; moreover, however we answer that, it's clear that for Carnap the core thus selected is "not assumed to be somehow naturally given, or to be *discoverable*, independently identifiable, as a natural kind within the total semantic field; it is *solely* a provisional singling out of certain uses for the purpose of explication. It is relative to the specific purposes of those who are singling it out." (Carus 2007, p. 285)

[9] For a recent defense of this distinction, cf. Warren (2016).

[10] Wilson's work is much more complex than the thumbnail sketch I'm using as a springboard here may suggest. Indeed, taking full measure of his account of the "life of the concepts", so to speak, would require a work on its own; here, I won't even be able to spell out its consequences for Carnap's project. For an interesting exchange about these consequences, cf. Carus (2012) and Wilson (2012); the latter, in spite of being printed before the former, is actually a reply to Carus's essay.

Thus, Carnap would reject that there is some kind of *given* invariant core to which explication must be faithful. On the contrary, we must *choose* the features we will treat as "core", in part by deciding to which purposes we will put our new, engineered concept (the explicatum). As Carus, reminds us, *this* process of coming to an agreement on which features will be considered as core features, is what Carnap calls "clarification". Here, it may be useful to contrast the account developed in this section with the one developed by Dutilh Novaes and Reck (2015, section 1.4). Tellingly, although they also put heavy emphasis on clarification, Dutilh Novaes and Reck reject that it is mostly a question of intersubjective agreement by claiming that "ultimately more [than intersubjective agreement] is at stake [in the clarification process], namely an adequate understanding of the explicandum and its original/intended uses, which will be fed into the explication" (Dutilh Novaes and Reck 2015, section 1.4). In other words, Dutilh Novaes and Reck apparently think that the clarification process should bring to light the "original/intended uses" of the explicandum, which would then form a kind of invariant core according to which the proposed explication would be judged. They are thus much closer to the classical picture of concepts than Carnap, who holds that ordinary language concepts generally have many different and not necessarily compatible uses,[11] and hence that in most cases there is no way to capture all their "original/intended uses", as those may be in conflict.[12] In those cases, we must first agree on what is the purpose of the explication and, given that purpose, which features or uses of the target concept are the most relevant, with the understanding that different purposes may call for a different selection of features. This may result in the same explicandum having several different explicata, all equally successful according to the metric established in their respective clarifications.

Let's illustrate the above points by pushing Carnap's own taxonomic example a bit further. Note that the new taxonomy which separated whales from fish was based on certain anatomical features of these animals. But why should we adopt anatomical features as a basis for biological taxonomy? In particular, why not adopt, say, *common ancestry* as the only relevant criteria? This would result in abandoning the category of *pisces* in favor of *two* other kinds, namely the *cartilaginous vertebrates* and the *bony vertebrates*, thus also separating, e.g., shark from fish, and regrouping fish with other animals such as mammals and birds at another level. It would also result in grouping together birds and crocodiles, separating the latter from lizards and snakes, thus also abandoning the concept of *reptile*. Given such differences between the two choices, a natural question to ask is, which is the right framework to adopt?

[11] Cf. Carnap (1963), especially his replies to Strawson and Bar-Hillel.

[12] Carnap would probably then find congenial Waismann's remarks on the open-texture of concepts as developed by Shapiro (2006).

Note that, in a sense, the question of which framework to adopt is prior to the explication process. It amounts to which features of the concepts to be formalized will be deemed as core, namely anatomic or phylogenetic features. The explication process, which in this case is basically the construction of the relevant category, can only really get off the ground *after* this preliminary decision has been made. Now, it's possible to read this as a debate about invariant core features of fish: should we privilege anatomic or phylogenetic information in determining the essence of *fish*? But I think a more Carnapian way of interpreting it is by considering it as asking a question about which framework is more useful. And answering this requires us to get clear on the purposes of biological taxonomy.[13] Obviously, we may wish to fulfill several purposes, to which different frameworks are suited; as Stein makes clear, Carnap's talk of tolerating alternative frameworks, embodied in his famous principle of tolerance, is not idle talk, but is supposed precisely to accommodate this type of situation. That is, there may be no need for choosing one framework over another once and for all: rather, we may work with different frameworks at the same time, since those frameworks may be put to use with different tasks in mind.

An interesting consequence of this pragmatic approach to the clarification process is that, in a sense, after a new concept has been engineered by the explication process, the old one drops out of the picture, which is why Carnap emphasizes that we're *replacing* the old concept by the new one. As Stein puts it:

> If one asks what such an explicatum is the explication of, more than one reply is possible. One can say that the exact characterization proposed is just the explication of the very concept in question (as a definition defines the concept whose definition it is); or that it explicates a presystematic idea, not previously in general use, but vaguely entertained by the inquirer when groping for clarity. I hope it is clear that all this is peripheral: what counts in the end—still in Carnap's view of things—is the clarity and the utility of the proposal; whether part of that utility has to do with an earlier, vaguer, general usage is distinctly a secondary matter. (Stein 1992, p. 282)

Let's take stock. We started with a puzzle about the similarity criterion, as emphasized by Dutilh Novaes and Reck, according to which there is a tendency for the fruitfulness criterion to conflict with similarity considerations. In order to solve this, I pointed out, following Stein, that for Carnap similarity considerations were external considerations, to be decided following an initial agreement on the purposes of the explication and the relevant

[13] Which is still hotly debated. For a quick survey, cf. Ereshefsky (2008), and, for a more in-depth treatment, cf. Ereshefsky (2003, esp. chap. 2). The proposal for taking common ancestry as the sole criterion for belonging to the same kind goes by the name of "process cladism".

features of the concept to be analyzed, in the step called "clarification". After this step, we proceed, in Carus's apt phrase, to some conceptual engineering;[14] that is, we create a new concept that is supposed to better meet that particular need (this is the explication step). The ultimate standard for the success of our explication is how fruitful the new concept is in relation to the agreed upon purposes in the clarification step. It's therefore possible to offer the following characterization of a typically *Carnapian* explication:

Carnapian explication: Given a target concept C (the explicandum) and a community[15] of researchers interested in C, the Carnapian explication process has two step:

1. Clarification: in this step, the researchers reach an agreement concerning the goal G to be reached by the explication and, based on G, on the relevant features $F_1, \ldots, F_n$ of C which best serve this goal and what are the best metrics to assess the success of the new concept to replace C;

2. Conceptual engineering: in this step, the researchers create a new concept C' in order to better achieve G. C' is then deemed successful if it fulfills the similarity (with respect to the selected features $F_1, \ldots, F_n$), exactness, fruitfulness, and simplicity criteria according to the metric established in the previous step.

The above will hopefully make clear how deeply embedded in Carnap's pragmatism is his own conception of explication and specifically why how successful a particular explication is is an external question for Carnap: it is an external question because the measure of success for a given explication is determined by a *decision* on the part of the researchers both about the goal of the explication and the metrics to be used. Of course, that is not to say that such decision is entirely arbitrary or irrational, as if external questions were to be decided based on the whim of the researchers.[16] But it does mean that it is a decision, and hence not completely determined by the data of the problem, whence there is indeed a certain *voluntarism* in Carnap's way of framing the explication task, to borrow a phrase from Richard Jeffrey (1994). It's precisely these voluntaristic or pragmatic elements that will be found lacking in the early Tarski's conceptual analyses, as we will see in the next sections.

[14] For an analysis of Carnap that takes seriously this engineering aspect, cf. French (2015) and Richardson (2013).

[15] "Community" here is taken in a wide sense, thus allowing a community of just one researcher in extreme cases.

[16] For an examination of Carnap's overarching concept of rationality that is relevant to this point, cf. Carnap (1958/2015) and Carus (2017).

1.2 Tarski's analysis of the concept of truth

In the last section, I briefly described Carnap's project of explication, highlighting its conceptual engineering aspect and the importance of pragmatic considerations in order to understand it. In particular, we saw how Carnap defined the task of explication as being the engineering of a new concept which could better fulfill the purposes of an old, vague one. Given that some[17] have considered Tarski's strictures concerning his analysis of the concept of truth forerunners to Carnap's explication concept, it'll be interesting to see how much in common they have. Rather surprisingly, the answer is " very little". Indeed, we will see that, far from being interested in "explicating" or analyzing the ordinary concept of truth, Tarski was instead interested in *expressing* this intuitive notion in a formal system. This difference is specially prominent when one takes into account the background philosophical projects of both philosophers.

In other words, just like there was a need to take into account Carnap's overall philosophical standpoint to better appreciate what he was getting at with his explication concept, in this section, I'll claim that something analogous holds for Tarski's analysis of the concept of truth, namely that it's important to take a step back and see the context in which such an analysis was embedded. Thus, before analyzing the paper itself, I'll first make a few remarks about Tarski's philosophical views at the time, which are characterized by an adherence to what Tarski himself called "intuitionistic formalism". This will allow us to see how different such a project is from Carnap's use of explication and also from Tarski's own later work.

1.2.1 Intuitionistic Formalism

In his paper "Fundamental Concepts of the Methodology of the Deductive Sciences", Tarski makes the following intriguing remarks:

> In conclusion it should be noted that no particular philosophical standpoint regarding the foundations of mathematics is presupposed in the present work. Only incidentally, therefore[,] I may mention that my personal attitude towards this question agrees in principle with that which has found emphatic expression in the writings of S. Leśniewski and which I would call *intuitionistic formalism*. (Tarski 1930/1983, p. 62, original emphasis)

This passage is intriguing for a number of reasons. First, there's the curious fact, noted by Patterson, that "Tarski insists that certain 'philosophical' views are not strictly relevant to our understanding of certain formal work, yet finds philosophical views worth mentioning anyway." (Patterson 2012, p. 16) Second, he introduces this rather odd term, "in-

[17] E.g. Dutilh Novaes and Reck (2015).

tuitionistic formalism", to describe such views—considering that the opposition between Brouwer's intuitionism and Hilbert's formalism was reaching a boiling point in the early 1930's, when Tarski wrote the above remarks, this would clearly strike his audience as almost an oxymoron. Finally, *late* Tarski also makes clear his disagreement with *young* Tarski in the English edition of the article, by the introduction of a footnote declaring that the view expressed in this paragraph "does not adequately reflect his present attitude." (Tarski 1930/1983, p. 62, dagger footnote)

A first clue to a resolution of these questions lies in a footnote Tarski appended to the above passage, which directs the reader to Leśniewski's "Fundamentals of a New System of the Foundations of Mathematics". Specifically, we are directed to the following passage:

> Perhaps I should add that for many months I spent a great deal of time working systematically towards the formulation of these systems of Protothetic by means of a clear formulation of their directives using the various auxiliary terms whose meanings [*Bedeutungen*] I have fixed in the terminological explanations given above. Having no predilection for various 'mathematical games' that consist in writing out according to one or another conventional rule various more or less picturesque formulae which need not be meaningful or even—as some of the 'mathematical gamers' might prefer—which should necessarily be meaningless, I would not have taken the trouble to systematize and to often check quite scrupulously the directives of my system, had I not imputed to its theses a certain specific and completely determined sense [*Sinn*], in virtue of which its axioms, definitions, and final directives (as encoded for SS5), have for me an irresistible intuitive validity [*intuitive Geltung*]. I see no contradiction, therefore, in saying that I advocate a rather radical 'formalism' in the construction of my system even though I am an obdurate 'intuitionist'.(Leśniewski 1992b, p. 487, original German expressions supplied by Patterson)

So the source of the term "intuitionistic formalism" is most likely the last sentence of the above passage. It's also clear the extent to which Leśniewski considers himself an "intuitionist": unlike the "mathematical gamers", he does not consider his systems as formal calculi, but as interpreted languages (or languages with meaning, to use Sundholm's turn of phrase[18]), that is, languages whose terms have some sort of "intuitive validity". But why does he consider himself a formalist? The passage continues:

> Having endeavored to express some of my thoughts on various particular topics by representing them as a series of propositions meaningful [*sinnvoller Sätze*]

[18] Cf. Sundholm (2003) for the distinction between "languages with meaning" and "languages without use".

> in various deductive theories, and to derive one proposition from others in a way that would harmonize with the way I finally considered "intuitively" binding [*welche Ich "intuitiv" als für mich bindend betrachte*], I know no method more effective for acquainting the reader with my "logical intuitions" [*"logischen Intuitionen"*] than the method of formalizing any deductive theory to be set forth. By no means do theories under the influence of such a formalization cease to consist of genuinely meaningful propositions which for me are intuitively valid. (Leśniewski 1992b, p. 487, original German terms supplied by Patterson, scare quotes restored)

This makes clear the extent to which Leśniewski considers himself a formalist: the best method to communicate his thoughts ("intuitions") is simply to set up a formal system, whose terms are unambiguous. It's thus possible to give the following initial characterization of intuitionistic formalism:

Minimal intuitionistic formalism: The (minimal) intuitionistic formalist is committed to two theses: (a) in order to do any philosophically interesting work, languages should be considered as interpreted, that is, their terms must have "intuitive validity" (the intuitionistic thesis); (b) the best way to communicate one's thoughts is to set up formal languages (the formalist thesis).

Again, this characterization is minimal in the sense that, although it suffices to distinguish the intuitionistic formalist from the Brouwerian intuitionist (who denies (b)) and the cartoonish Hilbertian formalist (who denies (a)),[19] it nevertheless fails to distinguish Lesńiewski's position from that of a Fregean logicist, who would count as an intuitionistic formalist by the above characterization.[20]

In any case, the first thing to notice is that, for Leśniewski (as for Frege), the *purpose* of a formal system is precisely to express his thoughts in a more exact manner than natural language allows. This may sound strange to contemporary ears, accustomed as they are to the idea that the main purpose of setting up a formal system is to study its properties, or to study the interaction between certain syntactic properties of its sentences and structural properties of its models, etc. Indeed, few would today think that the main purpose of formalizing the theory of algebraic closed fields of characteristic 0 in the first-order predicate calculus is to "express" one's thoughts about such structures; rather, what is interesting about such

[19]I say cartoonish because it seems highly unlikely that the historical Hilbert held anything like the view Lesńiewski attributes to his "mathematical gamer".

[20]The similarity with Frege has also been noted, e.g., by Sundholm (2003). For an in-depth analysis of Frege that is congenial to this point, cf. Blanchette (2012). Cf. also Betti (2008) on the classical ideal of science, which could be thought of as a point of convergence between Leśniewski and Frege.

formalization is what it reveals about these fields themselves (e.g. that any two such fields of a given cardinality are isomorphic, etc.).

An example might help the reader to better grasp Leśniewski's position. As remarked by Kotarbińska (1990, p. 54), Leśniewski attempted in his ontology to establish a new foundation for mathematics by replacing a distributive conception of sets, in which a set is an abstract object determined by its elements, with a collective conception of sets, according to which a set is a concrete object having its elements as parts. With this in mind, he set out to establish some general properties of propositions of the type "*A* is *b*", with "*A*" as a singular term (he called those "singular propositions"). Since a symbolic language is "technically much simpler than the colloquial language and at the same time less prone than that language to lead to misunderstandings in the formulation of ideas" (Leśniewski 1992c, p. 365), Leśniewski formulated his new system in one such language, adopting "ε" as a symbol for "is". Notice that this system is not an empty formalism, but has a very clear interpretation in Leśniewski's mind, which he wishes to convey to his readers. Unfortunately, this symbol was already appropriated by the "tradition of 'mathematical logic' and 'the theory of sets'", so Leśniewski had to try "very varied methods of appealing to [his readers'—D. N.] intuitions" (Leśniewski 1992c, p. 375) in order to elicit in them the correct thoughts. Interestingly, although he does mention among these methods showing his interlocutor the (sole) axiom which implicitly defined "ε", most of the explanations take the form of (colloquial equivalents of) *theorems* derived from this axiom. Not surprisingly, Leśniewski considers as a *necessary* condition for his axiomatization that it implies certain key theses which he associates with "ε" (namely, thesis (1)-(6) in Leśniewski (1992c, p. 368)). As Patterson puts it:

> Here, quite clearly, the role of certain theses in the system is to secure agreement as to the meaning of the primitive terms of the system, and the desired axiomatization is held to the standard that they be implied. By contrast, the axiom itself is simply required to imply (1)–(6) by the directives; it need not be intuitive in its own right. (Patterson 2012, p. 29)

Another remarkable observation is made by Leśniewski in connection with this example in the course of preventing further misunderstandings about the meaning of "ε" that he has in mind. After discussing several passages from Kotarbiński describing other usages of "is", Leśniewski observes the "well-known fact that *the expression* 'is' and propositions of the type '*A* is *b*' are *used* in colloquial language in a highly inconsistent way" (Leśniewski 1992c, p. 378, my emphasis). Notice that he doesn't say that the concept expressed by "is" is inconsistent, but rather that some usage of *the expression* is inconsistent. That's to be expected: as the meaning of a term is, for Leśniewski, an associated thought or intuition

(in any case, a subjective experience), there isn't much sense in calling such an experience "inconsistent".

In order to account then for the inconsistencies to be found in association with the ordinary language correlate of "ε", Leśniewski explicitly follows Kotarbiński, in particular his doctrine of "indirect" or "secondary" usage of an expression. The idea seems to be this: Kotarbiński has agreed to use sentences of the grammatical form "A is b" as the colloquial equivalent of "$A\varepsilon b$", which implies, in particular, that A is always a singular term. So, in his system, a sentence such as "man is mammal" expresses the fact that a *particular* man is a mammal. However, certain *conventions* (Kotarbiński's expression) of the English language associate the use of the singular term in that sentence with a *generic* reading, not a singular one, as in "every man is a mammal". Therefore, although (as written by Kotarbiński) the sentence directly expresses Kotarbiński's thought that a particular man is a mammal, it *indirectly* expresses the thought that every man is a mammal; alternatively, its secondary usage is to express the thought that every man is a mammal. Now, it may be the case that there are several, conflicting conventions associated with a different term, resulting in inconsistent usages. That does not mean that the subjective experience associated with the term is inconsistent, but it does mean, if one assumes that language is in part constituted by those conventions, that colloquial language is inconsistent.

If, on the one hand, such linguistic conventions can be an inconvenience, by tacitly invoking unwanted associations in the mind of one's interlocutors, nevertheless they can also work to one's advantage. In particular, if one explicitly states a set of conventions governing a symbol's usage, one can override the tacit conventions operative in colloquial language. So, for instance, in order to avoid ambiguity, Kotarbiński establishes the convention that all propositions of the form "A is b" should be taken as singular propositions, and that propositions such as (the generic reading of) "man is mammal" will always be expressed as "every man is a mammal", thus avoiding the ambiguity; this strategy is quoted with approval by Leśniewski (1992c, p. 379), who proceeds to extend it to other cases. Moreover, and this is particularly relevant to the next section, in setting up an artificial language, one can employ certain explicit conventions in order to constrain one's interlocutor's reading of certain expressions; that is, one can set up explicit conventions which, in a sense, close the gap between direct and indirect usage, so that the conventions make each sentence of the system say exactly what the system's deviser wants it to say. Since conventions determine the indirect meaning of an expression in virtue of its shape or form, this means that in such systems the syntactic form of an expression will exactly match its intended meaning.[21] That Leśniewski had something like this in mind is clear from the following quotation:

> I have more than once pointed out that a system of linguistic symbols, just as

[21] I emphasize this here because this will be precisely the point of Convention (T).

> any other system of symbols, e.g., the system of railway signals, requires the existence of certain rules for constructing the symbols and keys for reading them. (...) Taking into account the need so specified for a precise language, I established, in my previous papers, various linguistic conventions indicating on what rules the system of linguistic symbols is based and how to understand statements about some constructions which I used in analysis. (Leśniewski 1992d, p. 56)

There follows a reference to another passage from a previous essay, in which, after remarking that it is "inevitable to appeal to linguistic conventions when some doubts arise as to the way in which an object can be symbolized or as to the way in which an expression can be understood" (Leśniewski 1992a, p. 36), Leśniewski lists four conventions, of which he observes:

> Conventions II, III, and IV indirectly determine the role of the word 'not' in the system of linguistic symbols. If the role of the word 'not' were not determined, it would not be possible to decipher the system of linguistic symbols which employs this word. (Leśniewski 1992a, p. 37)

So the point of those conventions is precisely to avoid misunderstandings by laying out precise rules governing the use of certain signs. These quotations appear in Leśniewski's early work, so some care must be taken with them, given his later rejection of such juvenilia.[22] Indeed, as far as I could determine, there's no explicit mention of such conventions in Leśniewski's later works; however, that doesn't mean that they are not doing any work there. As Patterson remarks, it seems that their role has been taken by the "directives" of Leśniewski's systems:

> On Leśniewski's mature view conventions that determine the intuitive thoughts expressed by sentences are conceived of as the "directives" governing a system, rules for adding new theses to it. The determination of sub-sentential meanings expressed is then an indirect matter of the role of a sign as established by the theorems in which it appears, and it is this determination that comes to be the central issue, as with the example of "ε" above. (Patterson 2012, p. 29)

In other words, the conventions from Leśniewski's early work eventually become the axioms and rules of inference ("directives") of his formal systems. This ties in with our comments above regarding the role played by theorems in constraining a term's associated intuition, which results in the fact that, for Leśniewski, the meaning of a term is given by the whole theory in which it appears, i.e. his is a holistic, non-compositional theory of

[22] Cf. Betti (2004) for a more precise account of the relationship between early and late Leśniewski.

meaning. The overall strategy is clear: by the use of conventions, such as in the form of axioms and "directives", Leśniewski intends to close the gap between his own intuitions and his readers'; since these directives constrain the intuitions which we associate with a sentence's shape, as Patterson remarks, the idea is to make the sentences' *syntactic forms* "go proxy for their meanings" (Patterson 2012, p. 32).

There are a couple of lessons to be drawn from the above discussion. First, Leśniewski adopted what, in connection with Kotarbiński, Gawroński (1990) calls "strong psychologism in semantics", namely the thesis that the main function of any language is to express the subjective experience of the speaker. Given the difficulty of communicating such subjective experiences to another, it's not surprising that one of Leśniewski's main worries was to avoid misunderstandings; hence his interest in setting up formal systems devised to avoid all ambiguity. This also explains his obsession with reducing the number of primitives of his systems: the less primitives, the less expressions are there to be misunderstood, the more chances he has of communicating effectively. Second, we have the importance of "conventions" in establishing the meaning of a term, where a convention is a (not necessarily explicit) rule which determines the way a term or a symbol is to be understood. Finally, although Leśniewski paid lip-service to the importance of compositionality in setting up a deductive system, his preferred account of how terms or symbols are endowed with meaning is actually holistic and inferential: a term or symbol's interpretation is given not by its (implicit or explicit) definition alone, but by certain theorems in which it figures. In other words, a term or symbol's meaning is constrained by the theorems in which it appears, in such a way that it is actually such theorems that bear the weight of determining the term or symbol's intuitive interpretation. This allows us to give more substance to the Leśniewskian variety of intuitionistic formalism:

Leśniewskian intuitionistic formalism: The *Leśniewskian* intuitionistic formalist is committed to two theses: (i) strong psychologism about semantics, which entails that non-interpreted "languages" (if they can be considered as such) do not serve any philosophically interesting purpose and (ii) that the best way to communicate one's thoughts is to set up *conventions* governing formal languages that work in such a way that the theorems of such languages constrain the meaning of their terms.

This is much more substantial than the minimal intuitionistic formalism delineated earlier. In particular, (i) alone is sufficient to distinguish Leśniewski from Frege: although the latter's terminology is somewhat misleading, it's clear that, unlike Leśniewski's intuitions, Fregean thoughts are objective entities, and so that Frege rejects strong semantic psychologism.[23] One important feature of Leśniewskian intuitionistic formalism that will

[23]Indeed, Haddock (2012, chap. 1) lambasts a number of Frege scholars precisely for failing to recognize

be important in the sequel is that the Leśniewskian has very different priorities from the Carnapian pragmatist, which is reflected in the marked difference on how they conceive the role of formal tools: whereas Carnap wants to use formal tools in order to attain greater precision in his conceptual engineering task, Leśniewski wants to use formal tools in order to better *express* his own concepts and propositions, which he deems as precise enough. In the one case, one of the most important philosophical tasks is to *replace* vague concepts by precise counterparts (explication), in the other case, one of the most important philosophical tasks is to non-ambiguously *express* concepts which are already precise.

1.2.2 Tarski as an Intuitionistic Formalist

In the last section, I distinguished *minimal* intuitionistic formalism from *Leśniewskian* intuitionistic formalism. The latter was associated with three important doctrines: (i) strong psychologism about semantics, (ii) the importance of conventions, and (iii) the holistic picture of meaning thus entailed. Now, given the quotation from "Fundamental Concepts of the Methodology of Sciences", it's clear that at the time Tarski endorsed at least the minimal version of intuitionistic formalism. Unfortunately, since the passage cited by Tarski can only be adduced in support of this minimal version, it's not entirely clear whether he also endorsed Leśniewski's version. In this section, I want to consider some circumstantial evidence from Tarski's early work that indicated that he also accepted (i)-(iii) above, and hence Leśniewskian intuitionistic formalism. This is relevant for our assessment of whether Tarski is engaged in explication in his early work: if Tarski was indeed a Leśniewskian intuitionistic formalist, then it's likely that he would also consider as one of his main philosophical tasks the expression of certain concepts in a formal language, and that (proto-)explication would thus receive a minor role in his work, if any.

There are four papers that can be taken to be in line with the basic intuitionistic formalism project: "Fundamental Concepts of the Methodology of the Deductive Sciences" (Tarski 1930/1983), "On Definable Sets of Real Numbers" (Tarski 1931/1983), "On the Concept of Truth in Formalized Languages" (Tarski 1933/1983) and "Some Methodological Investigations on the Definability of Concepts". Given the limited scope of this chapter, I won't be able to analyze these four papers (they receive a detailed treatment in Patterson's book), instead concentrating just on the points that may help to illustrate Tarski's commitment to the above doctrines.

Let's start with semantic psychologism. Some clues as to Tarski's position in this regard can be gathered from the following passage, though Woodger's translation, following Blaunstein, unfortunately obscures this by dropping the key adjective:

---

or in some sense hide this point.

> It is perhaps unnecessary to add that we are not interested here in languages and sciences which are 'formal' in a certain specific sense of this term, namely such sciences that one attaches no intuitive meaning to the signs and expressions occurring in them; in regard to such sciences the issue raised here ceases to apply and it becomes no longer intelligible. To the signs occurring in the languages considered here we shall always ascribe quite concrete and, for us, intelligible meaning. (Tarski 1933/1983, pp. 166-167, translation by Gruber slightly amended)

The reading here proposed is straightforward: it parallels Leśniewski's point made in the passage quoted in the last section, namely that such formal languages are *not* to be considered as devoid of meaning, but, on the contrary, are supposed to facilitate the expression of an intuitive content, taken in the sense of the previous passage, namely as associated subjective experiences. In this sense, it's unfortunate that the adjective "intuitive" was dropped in the translation, for it occasioned the confusion that Tarski was here talking about model-theoretic interpreted languages—i.e. formal languages coupled with a structure and an interpretation function—, when that's clearly not the point.[24]

Similar appeals to intuitive content or meaning can be found in numerous instances throughout Tarski's writings of the period, e.g., Tarski (1933/1983, pp 153, 157, 160, 161, 166, among many others just in the paper on truth). However, this can be obscured because, as noted by Patterson (2012, p. 47), like in the previous passage, Tarski doctored such appeals to intuition already in the German translation of his paper, changing it to an appeal to "linguistic usage" or simply dropping it. The reader should therefore check the relevant sections in Gruber (2016) in order to find such appeals.[25] Especially relevant here is also the section on "Intuition" in the introduction to Gruber (2016) for comment on these changes: according to Gruber, it seems that Tarski was guided here by caution over the logical positivists distrust of that concept (the "linguistic usage" locution was apparently a suggestion from Ajdukiewicz). So, although the evidence is tentative, it seems that Tarski intentionally doctored such passages in order not to cause trouble with the positivists, which indicates that he did consider the expression to be both philosophically significant and moreover contentious.

A more important piece of evidence that Tarski was a Leśniewskian comes from the following passage:

[24]For an example of this kind of mistake, cf. the remarks in Hodges (1986, p. 147). Even without the key adjective in mind, however, Raatikainen (2003, p. 46n7) had already warned against this reading. Another reading of the passage is furnished by Fernández-Moreno (1994), who proposes that we should explain "meaning" in terms of truth-conditions. Again, this does not seem to cohere well with the "intuitive" qualifier.

[25]Incidentally, considering the complexity of the issues, translational or otherwise, involved in that paper, it seems to me that Gruber's commentary is invaluable for a historical reconstruction of Tarski's views.

> We are interested in a term of which we have a more or less precise account in relation to its intuitive content, but whose significance hasn't been (at least in the mathematical domain) until now rigorously established. We will then attempt to construct a definition of this term which, while satisfying the postulates of methodological rigor, will at the same time capture, justly and precisely, its "found" meaning. (Tarski 1931, p. 212, my translation)

There are two important features of the above passage. First, there is the appeal to "intuitive content". It may be possible to read the above allusion to an "intuitive content" associated with a term in a more neutral manner, perhaps by taking it to mean whatever content the term has in colloquial language. However, the continuation of the passage makes clear that Tarski takes "intuition" here to mean something subjective: "[In the case of geometry] the idea is to capture spatial intuitions, acquired empirically during the course of one's life, and which are, by the very nature of things, vague and confused (...)" (Tarski 1931, p. 212, my translation). Second, the definition Tarski seeks to construct is not a *replacement* of a vague concept, but rather one that will attempt to "capture, justly and precisely, its 'found' meaning". So the focus here is on *expressing* an intuitive content in a precisely constructed definition: exactly what we would expect of a Leśniewskian, but not of a Carnapian pragmatist. As we saw in the first section, both Carnapian explication and classical conceptual analysis aim at clarifying or, at the limit, replacing a vague, ordinary term with a more precisely specified concept. In other words, these projects presuppose that the target ordinary concept is more or less vague, and aim at improving this vagueness through conceptual analysis. But, as we have seen, the Leśniewskian intuitionistic formalist conceives her task in a very different way: we start with a "quite clear and intelligible" "intuitive meaning"[26], and the aim is to *express* this "intutive meaning" unambiguously in a deductive system.

In fact, in "On the Concept of Truth in Formalized Languages", Tarski is explicit that he is not interested in *analyzing* or *replacing* the concept of truth[27]:

> A thorough analysis of the meaning current in everyday life of the term 'true' is not intended here. Every reader possesses in greater or lesser degree an intuitive knowledge of the concept of truth and he can find detailed discussions of it in works on the theory of knowledge. (Tarski 1933/1983, p. 153)

However, at this point a natural question arises: if Tarski was not interested in analyzing the concept of truth (or definable set), then why offer *definitions* of these concepts? Why

[26]This is how Tarski generally characterizes the targets of his conceptual analyses in his intuitionistic formalist phase: cf., e.g., Tarski (1931/1983, p. 112), Tarski (1933/1983, p. 152).

[27]Indeed, the context of the passage also suggests that such analysis *had already been accomplished* by Kotarbiński, as remarked by Patterson (2012, p. 12).

not proceed like Leśniewski himself, that is, by introducing a primitive term (say, "Tr") axiomatically and then using the theorems of the system to constrain its intended meaning?[28] Notice that this procedure is not alien to Tarski: in "Fundamental Concepts of the Methodology of the Deductive Sciences", he proceeded exactly in this way, introducing the term "Cn" (for syntactic consequence) axiomatically and then constraining its meaning through a series of theorems. So, again, why not proceed in this way in the other papers? The problem is that both the concepts of *definable* and *truth* give rise to antinomies (Richard's and the liar, respectively, both mentioned by Tarski in the introduction to each paper). As a result, Tarski must first show that such concepts are "safe", that is, free from contradictions, and he does this by introducing *formally correct* definitions. After this preliminary step has been taken care of, he can then proceed in the way of the Leśniewskian intuitionistic formalist and show how the theorems derived on the basis of the definition make sure that it *captures* or *expresses* the target intuitive notion. This procedure is described in the following passage:

> Now, the question arises if *the definition which has been constructed* and whose formal rigor is not up for objection is equally *just from the material point of view*; in other words, *does it indeed capture the ordinary and intuitively known sense of the notion*? This question does not contain, let it be understood, any problem of a purely mathematical nature, but is nevertheless of capital importance for our considerations. (Tarski 1931, p. 229, my translation, original emphasis)

This brings us to point (ii) above, since, like in Leśniewski and Kotarbiński, an important role is played by conventions in capturing an intuitive meaning in a deductive system. We saw in the last section how Kotarbiński considered a convention to be a (perhaps implicit) rule which indirectly determined the meaning of a symbolic expression, and that Leśniewski put this to use by devising his axioms and rules of inference to constrain the intuitive meaning of a given expression. The idea was to close the gap between a person's subjective associations with a given expression and the expression's indirect meaning by making a sentence's syntactic form "go proxy" for its meaning. Tarski also inherited this Leśniewskian strategy of making the meaning of a term be constrained by the whole theory in which it appears, as can be seen by his use of Convention (T)[29] to constrain the meaning of the truth predicate:

> A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an *adequate definition of truth* if it has the following consequences:

[28] Incidentally, something like this has been the focus of much work recently. Cf. Horsten (2011) and Halbach (2011) for good summaries.

[29] Tarski doesn't use the Polish for "convention", but rather an expression that could be best translated as "agreement". I concur with Gruber (2016, p. 49) that nothing important hinges on this, however.

> (α) all sentences which are obtained from the expression '$x \in Tr$ if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage;
>
> (β) the sentence 'for any x, if $x \in Tr$, then $x \in S$' (in other words '$Tr \subseteq S$'). (Tarski 1933/1983, pp. 187–188, original emphasis)

Condition (β) basically says that every truth is a sentence, i.e. sentences are the only truth-bearers.[30] Together with (α), it works much like Leśniewski's directives for 'ε', namely it forces the reader to associate the correct intuition with the symbol 'Tr'.[31] If this reading is correct, then Tarski is not defining a new concept, say "formal truth", which he then claims captures the essential core of ordinary *truth*, with Convention (T) providing a kind of proto-Carnapian similarity test.[32] Rather, Convention (T) is a convention governing the use of a *symbol*, making sure that this symbol is correctly interpreted.[33] Moreover, this corroborates Tarski's acceptance of point (iii) above, namely meaning holism, since the meaning of the symbol "Tr" is to be given not only by its defining clause, but rather by the *consequences* of the defining clause, something that is emphasized in Convention (T) itself.

The above thus makes clear that Tarski did employ conventions in the positive role assigned to them by Leśniewski. But, again like Leśniewski, he also emphasized their *negative* role in generating inconsistencies in ordinary languages:

> (...) the very possibility of a *consistent use* of the *expression* 'true sentence' which is in harmony with the laws of logic and the spirit of everyday language seems to be very questionable, and consequently the same doubt attaches to the possibility of constructing a correct definition of this expression. (Tarski 1933/1983, p. 165, my emphasis; the whole passage was in italics in the original)

Notice the parallels with Leśniewski's remarks about "is": both emphasize that it is the *use* of a certain *expression* which is inconsistent, not the concept associated with the expression. This explains Tarski's remarks to the effect that ordinary language is inconsistent (Tarski

[30] Tarski goes on to remark, however, that this condition is inessential, something that has puzzled some commentators. Cf. Corcoran and Weber (2015) for discussion.

[31] Early interpreters, such as Field (1972/2001) and Corcoran (1999), took the point of Convention (T) to be a kind of guarantee of extensional adequacy, conflating material adequacy with extensional adequacy. It's clear from our discussion, however, that Tarski is after intuitive adequacy, not extensional adequacy; Corcoran later corrected himself in this regard, cf. Corcoran and Weber (2015, p. 10).

[32] As claimed, e.g., by Dutilh Novaes and Reck (2015).

[33] Incidentally, this strongly suggests that, for Tarski, there is only one concept of truth, against more standard interpretations which take Tarski to hold that, for each language L, there is a concept "truth-in-L". Cf. Smid (2014) for discussion.

1933/1983, pp. 164–165): if a language is partially constituted by the (tacit) conventions underlying it, then it follows that, if these conventions are in conflict, then the language will be inconsistent. This doesn't have the seemingly absurd implication that it is "a condition of my speaking English that I be willing to assert things that are not true" (Soames 1999, p. 64), for, unlike the conventions in Lewis (1983), the conventions appealed to by Tarski need not be "willingly assented to", i.e. I may be party to a linguistic convention (in Tarski's sense) without knowing that I am.

It should be clear by now that the Tarski from "The Concept of Truth in Formalized Languages" is not a proto-Carnapian philosopher, but rather a Leśniewskian one.[34] In this connection, it's possible to make a broader point regarding Tarski's overall philosophy of language in this period. In a number of writings, Robert Brandom has drawn a suggestive distinction between two types of semantic theories, based on whether they privilege the notion of *representation* or that of *expression*.[35] Basically, whereas representational semantics understands language as being primarily a tool for representing the world, the expressivist paradigm understands language as being primarily a tool for expressing one's particular thoughts. As Patterson (2012, p. 2) remarks, within this division, Tarski work is traditionally thought to lie squarely in the representational side.[36]

However, and this is again noted by Patterson, the findings of this section should make us suspicious of this traditional picture. In fact, given that he shared with Leśniewski the picture according to which the basic function of a formal language is to express intuitive meanings construed as intuitive, private experiences, it seems that Tarski was actually more aligned with the *expressivist* side of the above divide by the time of "The Concept of Truth in Formalized Languages". This is all the more surprising, because Tarski's semantic techniques could be readily used to develop a non-psychologist theory of meaning, thus neatly aligning his formal developments with a representational picture of language, to which they seem to be much more congenial. In other words, Tarski's semantics, however paradoxical this might sound, was completely dissociated from his theory of meaning.[37] It was only

[34]That's not to say he was in agreement with Leśniewski about everything, or even that they shared an overall project, but merely to point out that many of their key commitments are the same. Cf. Betti (2008) for a clear picture of their differences.

[35]Cf. the introduction to Brandom (2000), esp. section 4, for a clear—by Brandom's standards—statement of the distinction. Note that although Brandom portrays the contrast in rather dramatic terms in that section, in his more sober moments he tends to think of these not as opposed, but as complementary, as can be seen in Brandom (2008, p. 8).

[36]For instance, by Brandom himself, in fact. Cf. Brandom (2000, p. 7).

[37]In this sense, it is interesting to note that, when Tarski (1933/1983, p. 252) lists a series of semantic concepts, the concept of *meaning* is conspicuous by its absence. I owe this observation to Burgess and Burgess (2011, p. 18). Compare this absence to the explicit mention of *meaning* and other related concepts in Tarski (1944, p. 354).

when he realized that the semantics techniques he had developed could be used to supply this theory of meaning that he left intuitionistic formalism behind.[38]

1.3 Tarskian Explication

Tarski's abandonment of intuitionistic formalism had as a consequence that he also abandoned the three doctrines expounded in the previous section, namely (i) strong semantic psychologism, (ii) the importance of conventions and (iii) semantic holism. This move seems to follow his realization that the semantic apparatus he had developed in the serve of intuitionistic formalism could actually stand on its own. This is particularly evident in the case of logical consequence. As Patterson (2012, p. 178) remarks, whereas in his intuitionistic formalist phase, Tarski used the concept of (proof-theoretic) consequence to constrain the interpretation of his defined expressions (most notably, the expression "Tr" gets its intuitive meaning in part by its set of consequences), one of the first uses to which Tarski will put his newly defined semantic apparatus is precisely to define the notion of consequence. Not surprisingly, the way he approached this task is markedly different from his intuitionistic formalist approach.

Such change is apparently the result of his replacing strong semantic psychologism by his own semantic techniques, as expounded in "The Establishment of Scientific Semantics". In other words, instead of buying into a theory that considered the main function of a language to be the expression of thoughts, he now considered this function to be mainly representational, in line with the now rehabilitated semantic concepts of *denotation*, *satisfaction*, etc. This allows him to give a *compositional* account of the meaning of well formed compound expressions by exploiting the recursive clauses of the relevant semantic concepts, which thus leads him to abandon both the role of conventions and semantic holism by a bottom-up approach to meaning. This will lead to a more pessimistic outlook regarding the ordinary concepts to be explained: previously, he could attribute the problems with an ordinary concept to the inconsistent conventions governing its usage. Now, such problems are located in the concepts themselves. Indeed, already in the opening paragraphs of "On the Concept of Logical Consequence" we find the following remarks:

> With respect to the clarity of its content the common concept of consequence *is in no way superior to other concepts of everyday language* [my emphasis—D.N.]. Its extension is not sharply bounded and its usage fluctuates. Any attempt to bring

[38] For more on this development, which I won't treat here, cf., among others, Coffa (1991, chap. 15 and 16) and Patterson (2012, chap. 6 and 7), both of which emphasize how the crucial step for Tarski was his reading of Carnap's *Logical Syntax of Language*, especially its definition of consequence. For a detailed analysis of this definition, comparing it with Tarski's, cf. de Rouilhan (2009).

> into harmony all possible vague, sometimes contradictory, tendencies which are connected with the use of this concept, is certainly doomed to failure. (Tarski 1983, p. 409)

Noticeable here is the change of attitude concerning the "ordinary" concept to be defined. In "On Definable Sets of Real Numbers" and "On the Concept of Truth in Formalized Languages" Tarski would say about the ordinary concept to be analyzed, respectively, that "the arbitrariness of establishing the content of the term is reduced almost to zero" (Tarski 1931/1983, p. 112) and that its meaning in colloquial language "seems to be quite clear and intelligible" (Tarski 1933/1983, p. 152). In the above passage, on the contrary, not only is the concept associated with contradictory tendencies, but also the concepts of everyday language are declared vague and unstable. This difference is significant, as it points to an abandonment of the semantic psychologism characteristic of intuitionistic formalism.

Another noteworthy difference between this paper and the previous ones concerns the role of conventions. First, Tarski does formulate a condition which he considers to be necessary in order for a definition of logical consequence to be considered adequate, but he names it "Condition (F)", not "Convention (F)". This may seem like a mere cosmetic change (though it's not a mere slip: in "On the Semantic Conception of Truth and the Foundations of Semantics", Convention (T) is now renamed as "equivalence of the form (T)"); nevertheless, there does seem to be a change, which is also indicated by the fact that, whereas in "On the Concept of Truth in Formalized Languages" the conventions of a language rendered it inconsistent, without, however, thereby impugning the concept of truth itself, here it is the concept itself which is associated with "contradictory tendencies".

This marks a significant change from his previous work and helps to explain why, in "The Semantic Conception of Truth and the Foundations of Semantics", Tarski writes that "the problem of consistency has no exact meaning with respect to [natural] language" (Tarski 1944, p. 349). As we saw in the previous section, in "On the Concept of Truth in Formalized Languages", Tarski appealed to the constitutive role of conventions in a language in order to claim that ordinary language was inconsistent. Hence, his rejection now of this conclusion is further evidence that he also rejected this premise (that conventions are constitutive of a language), thus making clear his further distance from intuitionistic formalism.[39]

In any case, it seems clear that this pessimistic outlook on ordinary language concepts would become his considered view from then on. Another striking example concerns Tarski's remarks about the classical definition of truth associated with Aristotle: "a true sen-

[39]I thus agree with Ray (2003) (contra Patterson (2012, p. 248n2)) that the difference between "On the Concept of Truth in Formalized Languages" and "The Semantic Conception of Truth and the Foundations of Semantics" on this point is very significant, though we disagree on the reasons for the change.

tence is one which says that the state of affairs is so and so, and the state of affairs is so and so". In "On the Concept of Truth in Formalized Languages", this is deemed an intuitively adequate definition (though not a formally correct one), and moreover "seem[s] to be quite clear and intelligible", echoing Tarski's characterization of our intuition associated with this concept (Tarski 1933/1983, p. 155). On the other hand, in "The Semantic Conception of Truth and the Foundations of Semantics", this same formulation is said *not* to be "precise and clear" (Tarski 1944, pp. 343, 359).[40]

This new pessimistic outlook has obvious relevance for the way Tarski conceives his task. Before, the idea was to express an intuitive concept in a deductive system. Now, however, this task is considered as "doomed to failure". Unfortunately, Tarski isn't terribly clear about how he conceives of his new task, something that has generated a fair amount of debate;[41] it seems that his views actually evolved over the years.

In "On the Concept of Logical Consequence", although he recognizes that any "precise definition of this concept will show arbitrary features to a greater or lesser degree" (Tarski 1983, p. 409), he still seems reluctant to admit that his proposal departs in any significant way from the "common concept".[42] Perhaps Tarski's point was that he had captured the essential core of the concept, as suggested by the following quotation from his logic textbook:

> If a scientist wants to introduce a concept from everyday life into a science and to establish general laws concerning this concept, he must always make its content clearer, more precise and simpler, and free it from inessential attributes (...). (Tarski 1941, p. 27ff)

In other words, Tarski may have thought that the "common concept" of logical consequence was a mongrel, yet that it had an essential core that was best captured by the concept Jané and Betti attribute to formal axiomatics.[43] Be that as it may be, there's in any case an obvious tension between Tarski's declaration in the opening paragraph that it's impossible

[40] Barnard and Ulatowski (2016) also note this difference between the two texts, attributing this change to a possible influence of Naess. However, as I hope to show here, this change is actually a reflection of a broader change in Tarski's philosophy which can already be seen in the consequence paper, thus earlier than his awareness of Naess's work.

[41] Cf. especially Jané (2006) and Betti (2008) for those who defend that Tarski was only interested in capturing a very specific concept of consequence, namely the one related to formal axiomatics, and Patterson (2012, chap. 7) for the opposite view, according to which Tarski was interested in the "ordinary" concept of consequence.

[42] Whatever that is. I'm not convinced by Etchemendy (1999, 2008) that there is something like *the* "pretheoretic" notion of logical consequence. Rather, the notion of consequence appears to be a decidely *theoretical* notion, one that has evolved over the years to accomplish multiple tasks, so that it's perhaps dubious to expect it to have any unifying core.

[43] Another possibility is that he doesn't even realize that the "common concept" and the concept that comes from the tradition of formal axiomatics are different.

to capture the "common concept" of logical consequence and his repeated insistence that his analysis conforms to the "common concept". This tension is reminiscent of the one identified by Dutilh Novaes and Reck (2015) with regards to Carnap's account of explication: the more Tarski pushed towards precision, the less like the ordinary concept his own definition would become.

This tension begins to be resolved in "The Semantic Conception of Truth and the Foundation of Semantics", where Tarski begins to lean towards a more pragmatic viewpoint. The decisive section here is section 14. There, he first distances himself from the standpoint according to which there could be an essential core to the concept of "truth", by disparaging the idea that there might be a "right conception" of truth as some kind of mysticism:

> In fact, it seems to me that the sense in which the phrase "the right conception" is used has never been made clear. In most cases one gets the impression that the phrase is used in an almost mystical sense based upon the belief that every word has only one "real" meaning (a kind of Platonic or Aristotelian idea), and that all the competing conceptions really attempt to catch hold of this one meaning; since, however, they contradict each other, only one attempt can be successful, and hence only one conception the "right" one. (Tarski 1944, p. 355)

So we are at one further remove from intuitionistic formalism. In "On the Concept of Truth in Formalized Languages", it was a question of expressing or capturing the intuitive conception of truth which every person possess to a greater or lesser degree. In "On the Concept of Logical Consequence", this intuitive conception is gone, and in its place we have a kind of mongrel of contradictory tendencies, of which Tarski will attempt to capture the common core. In "The Semantic Conception of Truth and the Foundations of Semantics", even this idea has receded into the background. Whether there is a common core or not, that's largely irrelevant. The above passage continues:

> It seems to me obvious that the only rational approach to such problems would be the following: We should reconcile ourselves with the fact that we are confronted, not with one concept, but with several different concepts which are denoted by one word; we should try to make these concepts as clear as possible (by means of definition, or of an axiomatic procedure, or in some other way); to avoid further confusions, we should agree to use different terms for different concepts; and then we may proceed to a quiet and systematic study of all concepts involved, which will exhibit their main properties and mutual relations. (Tarski 1944, p. 355)

Indeed, in spite of the aggressive tone of his remarks, Tarski agrees with Carnap that the correct procedure is simply to put forward the different proposals for explication and

see where they lead. It's actually quite likely that the influence here is direct. As Mancosu (2010d, sec. 15.8) relates, Carnap, Neurath, Tarski, and Kokoszyńska met in 1937 to discuss Neurath's objections to Tarski's and Kokoszyńska's account of truth. In this meeting, Carnap presented a paper which basically ended with the suggestion that each party to the dispute should be free to pursue their own systematic projects, without engaging in prolonged polemics one against the other. It's highly possible that Tarski had those remarks in mind when he wrote the above passage.

This Carnapian pragmatism is expressed also in section 17, where Tarski says that, although he still holds the belief that his account is in agreement with the "common usage" of the word "truth", he "readily admit[s] [he] may be mistaken" (Tarski 1944, p. 360). That he does not remark further on the consequences of this concession to his own account is indicative that he does not consider the point very relevant.[44]

This pragmatist element is somewhat inchoate in this paper, however, becoming only fully explicit in the papers from the 60's, namely "Truth and Proof" (Tarski 1969) and "What are Logical Notions?"(Tarski 1966/1986). In those papers, Tarski is much clearer on what he considers to be a successful conceptual analysis:

> Whenever one explains the meaning of any term drawn from everyday language, he should bear in mind that the goal and the logical status of such an explanation may vary from one case to another. For instance, the explanation may be intended as an account of the actual use of the term involved, and is thus subject to questioning whether the account is indeed correct. At some other time an explanation may be of a normative nature, that is, it may be offered as a suggestion that the term be used in some definite way, without claiming that the suggestion conforms to the way in which the term is actually used; such an explanation can be evaluated, for instance, from the point of view of its usefulness but not of its correctness. Some further alternatives could also be listed.
>
> The explanation we wish to give in the present case is, to an extent, of mixed character. What will be offered can be treated in principle as a suggestion for a definite way of using the term "true", but the offering will be accompanied by the belief that it is in agreement with the prevailing usage of this term in everyday language. (Tarski 1969, p. 63)

[44] Against Patterson (2012, p. 231), who thinks that there is a "plain inconsistency" between sections 14 and 17. There only seems to be an inconsistency if you think that the "common usage" is the "right" one, but it seems clear that Tarski himself is not of that opinion. Similarly for the supposed inconsistency between section 14 and 18. Although Tarski pokes fun at conceptions that deny the (T) sentences, accepting the (T) sentences is not tantamount to accepting Tarski's conception, and moreover to remark on the apparent paradox of accepting the negation of an instance of the (T) schema is not to say that any such a view is inherently absurd, as Tarski himself makes clear.

> Let me tell you in advance that in answering the question 'What are logical notions?' what I shall do is make a suggestion or proposal about a possible use of the term 'logical notion'. This suggestion seems to me to be in agreement, if not with all prevailing usage of the term 'logical notion', at least with one usage which actually is encountered in practice. I think the term is used in several different senses and that my suggestion gives an account of one of them. (Tarski 1966/1986, p. 145)

Even though Tarski still retains the belief that his analyses conform to some usage of the term he is analyzing (the "prevailing" one in the case of truth, an unspecified one in the case of logical notions), it's clear that the pragmatic element prevails. In both cases, the agreement with a pre-existing usage is mentioned as an aside, almost like a "bonus feature" of his account, instead of occupying center stage. Thus, in both cases he makes clear that what he is offering is a suggestion or proposal (which should thus be evaluated in terms of its *usefulness*), and that there is an accompanying "belief" that this proposal happens to agree with a pre-existing usage, but he doesn't pursue the topic further. Capturing *the* "intuitive meaning" of a concept has become, as Stein would put it, a secondary matter.

1.4 Conclusion

In this chapter, we have traced the evolution of Tarski's philosophy as a background to his famous articles on truth and consequence. In particular, we have seen how Tarski progressed from a position closely tied to views he inherited from Leśniewski (or what he called "intuitionistic formalism"), according to which one of his main tasks is to capture an intuitive meaning into a deductive system, to a much more Carnapian view, according to which his task was to *propose* or suggest how to use a given term from then on. In one case, success should be judged by how close his definition comes to capturing the target intuitive content, something that can be measured in part by the *consequences* of the proposed definition. In the other case, success should be measured by the usefulness of the proposal: what kind of question it elicits, what new areas it opens to exploration, what theorems it allows one to formulate and prove. The latter is also a more *tolerant* attitude: proposals should be put forward and explored, to see where they lead.[45] In this respect, it's not without importance to highlight, as does Betti, that Leśniewski was the least tolerant of the Lvov-Warsaw school:

> Leśniewski's uncompromising stance was rather the exception in the Lvov-Warsaw School. For, generally speaking, the Lvov-Warsaw School at its zenith

[45] This Tarskian pragmatism is especially emphasized by Sinaceur (2009), whose conclusions are much congenial to mine.

> was marked by a liberal attitude towards the use of all admissible mathematical methods, non-constructive ones included, and it was not committed to any particular philosophical position (Betti 2008, p. 50)

Tarski's influence on Carnap is well-documented; Carnap's influence on Tarski, less so, though we are definitely beginning to understand this relation better.[46] It's a curious relationship: just like Tarski's semantics opened Carnap's eyes to techniques which were already latent, so to speak, in his work, Carnap's influence on Tarski also seemed to be of this liberatory sort, allowing him to shed off the Leśniewskian baggage and adopt a philosophical attitude more in tune with his mathematical practice.

[46] In particular, the works of Coffa (1991), de Rouilhan (2009), Patterson (2012) have done much to better understand how Carnap's *Logical Syntax of Language*, for instance, was pivotal for Tarski's development.

Chapter 2

Tarski's Nominalism

In the first chapter, I argued that Tarski moved from a broadly Leśniewskian position to a position much more congenial to Carnapian pragmatism. In doing this, I showed how this perspective helped us to assess Tarski's puzzling remarks concerning his own proposed definitions, as well as furnishing us with his own criterion for a successful definition. I also discussed how this should be read not necessarily as a question of (reciprocal) influence, but perhaps of *confluence* between their respective pragmatist temperaments.

In spite of this confluence, however, there is a stark contrast between the two philosophers in at least one respect. Tarski, unlike Carnap, consistently defended a particular metaphysical outlook, namely physicalism, which sometimes led him in the direction of finitism.[1] This is surprising, since Tarski actively pursued research involving large cardinals, which is plainly at odds with this finitistic tendency. Moreover, he routinely employed in his investigations, when it suited him, higher-order logic, and even infinitary logic, which again went against his finitistic strictures, particularly as laid out in his conversations with Carnap. Especially relevant for us is his condemnation of higher-order logic as embodying some kind of platonism: this is puzzling, for his proposed definition of logical notions is more naturally interpreted against a type-theoretical background. So, in this chapter, I will examine his nominalistic tendencies more closely, focusing on how they might harmonize with his reliance on type-theory. If in the first chapter Carnap was the main contrasting figure, here I want to draw attention to certain similarities between *Quine* and Tarski, especially considering the way the first philosopher introduced his own notion of explication as a way of reducing the ontological commitments of the targeted theories.

Accordingly, the first part of the chapter discusses Quine's well-known explication doctrine, emphasizing its differences from Carnap's own project. This is done by first sketching a quick review of their extended polemic, using this as a fodder for discussing Quine's own

[1] Even ultrafinitism, as Givant told Rodríguez-Consuegra, and as appears in Tarski's conversations with Carnap. Cf. Rodríguez-Consuegra (2005, p. 255) and Frost-Arnold (2013, p. 153).

version of explication in detail in the next section. Finally, in the second part of the chapter, I analyze Tarski's nominalism, in particular how he developed a strategy strikingly similar to Quine already in the early 1940's. The technical appendix discusses a result connected with the nominalist strategy pursued by Tarski.

2.1 Quine and Carnap on explication

2.1.1 Quine's Polemic with Carnap

Quine's extended polemic with Carnap has been the subject of numerous studies.[2] It's not our purposes here to review this extensive literature; rather, I just want to make salient some points that will help to clarify Quine's own notion of explication and how it differs from Carnap. Particularly important here, given its overall relevance to Tarski's nominalist strategy, will be their differing views on the importance of ontological reduction.

Recall from the last chapter that Carnap (1950/1956) distinguished between two types of questions, questions *internal* to a given linguistic framework and questions *external* to a given linguistic framework. Very briefly, questions internal to a given framework should be answered according to the framework's rules, whereas questions external to a given framework, especially questions about which framework to adopt for a given purpose, are a matter of pragmatic evaluation. There, I focused on how this distinction impacted his conception of explication; specifically, I argued, following Stein (1992), that the measure of a successful explication was an external question, to be solved on pragmatic grounds. Here, however, I want to focus more narrowly on the impact of this distinction for ontological questions, which were the context of its original introduction in "Empiricism, Semantics, and Ontology". In particular, I want to focus on two theses advanced by Carnap:

Linguistic thesis: General ontological questions should be replaced by questions regarding the utility of linguistic frameworks;

Tolerance: Different linguistic frameworks may be adequate for different purposes.

Both theses are familiar enough. The linguistic thesis expresses Carnap's well-known attempt to deflate ontological questions by (in a sense) trivializing them. Roughly put, Carnap's argument is the following. We are given certain linguistic frameworks which are

[2]The classical essays from this debate are Quine (1935/1966, 1952/2004, 1963) and Carnap (1955, 1963). For some important studies, cf., among many others, Stein (1992), Richardson (1997), Ricketts (2004, 2009), Creath (1990, 1991, 2003, 2007, 2017), Friedman (2006, 2012b), Hylton (2007, esp. chaps. 2 and 9), Soames (2014), Ebbs (2014), Gustafsson (2014). My own approach here has been heavily influenced by Stein, Richardson, Friedman, and, on the specific issue of explication, by Gustafsson.

somewhat like interpreted languages, plus some methodological guidelines (in the form of rules of inference, for instance). Questions about the existence of certain types of entity can be of two sorts: in one case, the question is not trivial (e.g. about the existence of the Higgs boson, or about certain functions in a given Banach space), but are then supposed to be settled by the methods allowed for by the framework in which they are raised (e.g. by the methods, whatever they are, which guide particle physics or functional analysis). These are internal questions. On the other hand, we have certain general existence questions which apparently cannot be settled by employing the methodological canons of the framework in which they are raised (e.g. about the existence of abstract entities). Such questions, Carnap maintains, are ambiguous: they are either trivial, following in a straightforward way from the rules of logic, in particular existential generalization (e.g. five is a number, so there is something which is a number, i.e. there are numbers), or else they are nontrivial, that is, they are not questions raised inside a framework, but rather questions about the adequacy of the framework itself (e.g. "Is a framework which implies the existence of abstract entities an adequate one?"). Since, interpreted in the latter way, general existence questions, being external to any framework, are not amenable to treatment by any methodological canon (as these are tied to this or that framework), there is in general no systematic method for settling them. It follows that they are largely pragmatic. As Carnap famously puts it:

> We may still speak (and have done so) of "the acceptance of the new entities" since this form of speech is customary; but one must keep in mind that this phrase does not mean for us anything more than acceptance of the new framework, i.e., of the new linguistic forms. Above all, it must not be interpreted as referring to an assumption, belief, or assertion of "the reality of the entities". There is no such assertion. An alleged statement of the reality of the system of entities is a pseudo-statement without cognitive content. To be sure, we have to face at this point an important question; but it is a practical, not a theoretical question; it is the question of whether or not to accept new linguistic forms. The acceptance cannot be judged as being either true or false because it is not an assertion. It can only be judged as being more or less expedient, fruitful, conducive to the aim for which the language is intended. (Carnap 1950/1956, p. 214)

The last sentence takes us into the tolerance thesis. Since there may be different purposes for which different languages may be more or less adequate, it follows that we should not exclude certain linguistic forms merely on the basis of philosophical prejudice. On the contrary, the basic task of the philosopher, according to Carnap, is precisely the invention and exploration of different linguistic frameworks, supplying suitable linguistic forms for the scientist. It is the ideal of the philosopher as an engineer, creating useful tools for the

progress of science. Again, it is in light of this ideal that Carnap's notion of explication should be seen: explication, insofar as it is a piece of conceptual engineering, is one of the most valuable contributions that the Carnapian philosopher can give to the scientist.

Notice that this is *not* to say that such external questions are unimportant or not amenable to *rational* treatment. On the contrary, relative to certain values, some choices may appear as entirely rational or irrational, so we are very far from a crude relativism that could be thought of as the upshot of tolerance.[3] Indeed, as Stein (1992, p. 279) makes clear, Carnap thought of his own theses not as assertions, but as proposals, so that his controversy with Quine is also subject to rational evaluation, if they could be given enough time to verify which one enjoyed the most pragmatic success.

Turning now to Quine, he famously rejects the tolerance thesis. According to Quine, there is just one overarching goal of science: "As an empiricist I continue to think of the conceptual scheme of science as a tool ultimately, for predicting future experience in the light of past experience" (Quine 1952/2004, pp. 51–2). Since this is the main underlying purpose of our scientific theories, such theories should be judged according to this single standard, that is, how well they predict future experience from past experience. There is thus correspondingly no tolerance for those theories which do not fare well in this test, or fare worse than other competing theories.

As for Carnap's linguistic thesis, Quine accepts it, but transformed by his rejection of tolerance. Like Carnap, Quine agrees that ontological questions are in a sense relative to a given framework—indeed, they even agree that a good indication of a theory's ontological commitment is the values taken by the existentially quantified variables of the theory.[4] Since, however, different frameworks can all be judged according to the same standard, the best framework gives us the best ontology. Hence, ontological questions, even general ones, are not devoid of cognitive content, as Carnap wanted. Rather, they are settled as any other scientific matter, namely by analyzing our best scientific theory. That's why Quine says that "[o]ntological questions then end up on a par with questions of natural science" (Quine 1951/2004, p. 256). Quine thus replaces Carnap's engineering task with a, let us say, tidying up task:

[3]For more on this important theme that I will not be able to develop here, cf. Carnap (1958/2015) and the commentary by Carus (2017).

[4]In fact, this agreement may also indicate a tension inside both Carnap's and Quine's views. In the late 1950's, Carnap devised an ingenious device for separating a theory's factual content from its analytical content: the factual content is given by the Ramsey sentence of the theory and its analytical content by the Carnap sentence, which is a conditional which takes the Ramsey sentence of the theory as its antecedent and the theory itself as its consequent. This introduces tensions because, as Quine (1984/2008, pp. 124–5) himself recognizes, this does seem to give Carnap a working definition of analyticity. On the other hand, this may also push Carnap in the direction of realism, thus closer to Quine's position. For discussion of this last point, cf. Psillos (1999, chap. 3), Friedman (2012a), and Demopoulos (2013).

> Science, though it seeks traits of reality independent of language, can neither get on without language nor aspire linguistic neutrality. To some degree, nevertheless, the scientist can enhance objectivity and diminish the interference of language, by his very choice of language. And we, concerned to distill the essence of scientific discourse, can profitably purify the language of science beyond what might reasonably be urged upon the practicing scientist. (Quine 1954/2001, p. 199)

To reiterate, as this quotation makes clear, Quine agrees with Carnap that existence questions are framework-relative. However, he thinks that, insofar as we are constrained to choose one best framework, this by itself endows ontological questions with cognitive content. The philosopher's task is then to "distill the essence of scientific discourse" by tidying up, so to speak, our current best scientific theory in order to make explicit its ontological commitments. In a sense, then, Quine replaces the two Carnapian theses outlined above with:

Quinean Linguistic Thesis: The ontological commitments of a theory are to be "read off" the range of its existentially quantified variables.

Intolerance: There is a single set of standards against which to judge our theories, whence we can (ideally) select a single best one.

Notice that the Quinean linguistic thesis is in a sense very similar to Carnap's own linguistic thesis. Both agree that the entities to which a theory is committed should be read off the range of its existentially quantified variables. The difference is that Quine takes such commitments seriously (in part due to his adherence to intolerance), whereas Carnap deflates them with his own brand of tolerance. Quine's own version of explication should therefore be understood against this background. It is to this version that I turn to in the next section.

2.1.2 Quinean Explication

As mentioned in the last section, Quine believes that it is the philosopher's job to distill the essence of a scientific theory. This is rather vague, but Quine actually has some precise indications about how to proceed. The first thing the philosopher should do in order to reveal the structural underpinnings of a scientific theory is to *regiment it into a canonical language*, typically the first-order predicate calculus. Quine's idea is simple: he, like Carnap, is aware of the many ambiguities of natural languages. It is therefore desirable to have a special language which is as unambiguous as possible, especially if such a language wears

its structure in its sleeve, so to speak. In any case, Quine has quite some grandiose things to say about this seemingly pedestrian activity of paraphrasing certain theories into a more perspicuous language:

> The same motives that impel scientists to seek ever simpler and clearer theories adequate to the subject matter of their special sciences are motives for simplification and clarification of the broader framework shared by all the sciences. Here the objective is called philosophical, because of the breadth of the framework concerned; but the motivation is the same. The quest of a simplest, clearest overall pattern of canonical notation is not to be distinguished from a quest of ultimate categories, a limning of the most general traits of reality. Nor let it be retorted that such constructions are conventional affairs not dictated by reality; for may not the same be said of a physical theory? True, such is the nature of reality that one physical theory will get us around better than another; but similarly for canonical notations. (Quine 2013, p. 147)

Quine's thought seems to be this: sometimes, during our scientific investigations, we arrive at certain formulations that, while they may get us going for the moment, are less than ideally clear. One of the philosopher's job is then to paraphrase such formulations into something more acceptable. In particular, some paraphrases may replace a formulation containing an obscure notion by one free of such obscurity. As Quine himself says, "the simplification and clarification of logical theory to which a canonical logical notation contributes is not only algorithmic; it is also conceptual" (Quine 2013, p. 146). That is, paraphrase does not help only in achieving clarity merely for the sake of communication, such as in the elimination of ambiguity; in showing how we can make do with a simpler formulation than the original one, paraphrasing may also make our theories simpler, be it by reducing our stock of primitive predicates (say, by showing how every sentence containing a certain predicate may be paraphrased into one that does not contain it) or even in reducing our ontology (say, by showing how we can avoid certain ontological commitments by paraphrasing them away). That's why Quine takes this activity to be a "limning of the most general traits of reality".

This takes us into his own account of explication, which, not surprisingly given the preceding, is most fully elaborated in the chapter about "Ontic Decision" of *Word and Object*. There, after offering the ordered pair as a paradigm of explication, Quine defines it by saying that:

> We do not claim synonymy. We do not claim to make clear and explicit what the users of the unclear expression had unconsciously in mind all along. We do not expose hidden meanings, as the words 'analysis' and 'explication' would

> suggest; we supply lacks. We fix on the particular functions of the unclear expression that make it worth troubling about, and then devise a substitute, clear and couched in terms to our liking, that fills those functions. Beyond those conditions of partial agreement, dictated by our interests and purposes, any traits of the explicans come under the head of "don't-cares" (Quine 2013, p. 238)

Superficially, this may seem like the Carnapian concept, and indeed Quine references Carnap's *Meaning and Necessity* at this juncture. However, as should be clear from the previous section, I will argue that Quinean explication is importantly different from Carnap's. Of course, as mentioned in the previous chapter, there is a certain minimal version of the concept of explication which is sufficiently neutral to cover the Quinean and the Carnapian variety. Let's recall its definition:

Minimal explication: Given a concept C as an *explicandum*, a *successful minimal explication* for C is another concept C' which fulfills the similarity, exactness, fruitfulness, and simplicity criteria.

As we saw in the previous section, Carnap takes this to be essentially an engineering task. Not surprisingly, he thinks this task typically involves the creation of new linguistic frameworks that in some sense contain a formalized version of the explicandum—for instance, his own work in the development of an inductive logic. As we also discussed there, Quine has no place for this engineering activity. For him, "explication" is basically another form of paraphrase, as becomes clear from his analysis of the ordered pair as a paradigm.

In this analysis, Quine points out that the ordered pair is a very useful mathematical notion, which allows us to reduce, e.g., properties, relations, and functions to sets of ordered pairs. Since ordered pairs function in these reductions as elements of sets, they are to be treated on a par with other objects. But what type of object is an ordered pair?

One thought would be to take the ordered pair as a sui generis object, subject only to the following constrain (below, $\langle x, y\rangle$ is the traditional notation for the ordered pair of x, y):

$$\langle x, y\rangle = \langle w, z\rangle \text{ iff } x = w \text{ and } y = z.$$

Unfortunately, however, as clear as this postulate is (in particular, it furnishes us with a clear cut identity criterion for ordered pairs), it is bound to generate some perplexities; Quine cites here Peirce's convoluted definition as an example of such confusion, but he could have cited, e.g., Frege and Russell as well for some complicated attempts at a definition.[5] If we could therefore find a suitable substitute, couched in a more innocent lan-

[5] Both Frege and Russell took the ordered pair to be *derived* from relations, and not vice versa. It took some considerable time and clarity over set-theoretical foundational issues—such as the distinction between set-

guage, that would perhaps eliminate any lingering anxiety over these peculiar entities. Fortunately, there is not just one, but two such substitutes: Wiener's definition of $\langle x, y\rangle$ as $\{\{\{x\}, \varnothing\}, \{\{y\}\}\}$ and Kuratowski's definition of the same object as $\{\{x\}, \{x, y\}\}$. This may seem one definition too many: how should we choose between them? Which one is the *real* definition, the one which captures the essence of the ordered pair?

As I hope is clear by now, Quine evidently rejects these questions. That is not to say that one definition is as good as any other: there may be pragmatic criteria for deciding in favor of one over the other (indeed, Kuratowski's definition is slightly simpler, since it (a) requires sets of a lower rank and (b) does not employ the empty set, but only sets constructed using only x, y). But that is not the crux of the matter. The real issue is that whatever perplexities we had about ordered pairs are dissolved by such a definition, which is couched entirely in terms of sets. Before we discovered such definitions, we needed to make use of at least two types of entities, namely ordered pairs and sets, the former being the source of some misguided philosophical doctrines. The definitions shows us that we can proceed with only one type of entity, by eliminating every mention of ordered pairs in favor of sets. *This* is the main goal of explication for Quine: to reduce philosophical perplexities by *reducing* or *eliminating* certain entities by offering innocent paraphrases of every context in which the suspect entities occur:

> A similar view can be taken of every case of explication: *explication is elimination*. We have, to begin with, an expression or form of expression that is somehow troublesome. It behaves partly like a term but not enough so, or it is vague in ways that bother us, or it puts kinks in a theory or encourages one or another confusion. But also it serves certain purposes that are not to be abandoned. Then we find a way of accomplishing those same purposes through other channels, using other and less troublesome forms of expression. The old perplexities are resolved. (Quine 2013, p. 240)[6]

Summing up, we arrive at the following definition:

Quinean explication: Given a target concept C and a community of researchers interested in C, but still somehow perplexed by C, a concept C' is an explication for C iff:

(i) C' fulfills the same relevant purposes $O_1, \ldots, O_n$ as C;

membership and set inclusion—for the current definition to emerge. Cf. Kanamori (2003) for an account of this development; incidentally, Kanamori mentions, but does not analyze, the passages from Quine in which he uses the ordered pair as a prime example of explication.

[6]This statement is typical of Quine. Here is another clear statement of this idea: "To define is to eliminate, to excuse, to exonerate. Statements containing the defined term import no content, no risk of error, not already present in statements lacking the term" (Quine 1984/2008, p. 124).

(ii) C' is stated in a canonical notation, typically the first-order predicate language;

(iii) C' contains in its definition only notions which the researchers agree are ontologically innocent, or at least more ontologically innocent than the notions contained in C.

As observed by Carus (2007, p. 266), requirements (ii) and (iii) above introduce new, philosophical requirements that were not to be found in the Carnapian definition.[7] It is precisely those features that explain the nominalist strategy of paraphrasing away ontologically inflated theories into more innocent counterparts. From our point of view, these strategies, such as the one famously pursued by Field (2016), are thus best viewed as attempts at Quinean explication. In the next section, we will see how Tarski develops one such strategy, thus making him a forerunner of this type of project.

2.2 Tarski's Nominalism

2.2.1 Tarski's nominalist and physicalist tendencies

Tarski's nominalist and physicalist leanings are well documented.[8] Tarski himself was quite forthright about it, even jocosely so:

> I happen to be, you know, a much more extreme anti-platonist. (...) So, you see, I am much more extreme; I would not accept the challenge of Platonism. You agree that continuum hypothesis has good sense; it is understandable. No, I would say, it's not understandable to me at all. (...) I represent this very rude kind of anti-Platonism, one thing which I could describe as materialism, or nominalism with some materialistic taint, and it is very difficult for a man to live his whole life with this philosophical attitude, especially if he is a mathematician, especially if, for some reasons, he has a hobby which is called set-theory, and worse—very difficult. (Tarski 2007, pp. 259-260)[9]

[7] As also observed by Carus in the same place, this also brings Quine closer to the Carnap of the *Aufbau* and introduces a certain tension within Quine's naturalism, insofar as such requirements appear to be demands imposed by the *first philosopher* on the activities of the scientist. But I will not dwell on such matters here.

[8] Cf., in particular, Mancosu (2010b,c) and Frost-Arnold (2013).

[9] It's not entirely clear what are Tarski's criteria for something to be "understandable". This issue also appears in the conversations with Carnap and Quine, when Tarski says that he only "understands" certain types of finitistic, nominalist language; the intended contrast is obviously that between an uninterpreted calculus, on the one hand, and a language to which we can attach concrete meanings, but that's not an explanation of what constitutes understanding. Cf. Frost-Arnold (2013), p. 153 for the Tarski quotation and pp. 27-37 for an analysis. Cf. also the recent discussion around Frost-Arnold's book, some of which revolves around this specific point: Creath (2015) and the reply by Frost-Arnold (2015).

How should we interpret such nominalism? Some clues can be found in Tarski's conversations with Carnap during his stay at Harvard in the early 1940's. According to Carnap, Tarski says that he can only understand a language which (1) talks only about a finite number of individuals,[10] (2) in which the individuals are assumed to be physical things, and (3) does not refer to universals or classes (Frost-Arnold 2013, p. 153). Accordingly, Tarski's nominalism is one that does not assume an infinite number of individuals (and actually leans towards a kind of finitism), assumes a certain physicalism (Tarski refers here to Kotarbiński's "reism"), and does not accept universals such as sets,[11] properties, etc. It is important to note how Tarski links his nominalism with the type of language which he can understand; I will come back to this linguistic thesis in the next section; for now, let us examine some other ideas that he associated with nominalism.

The above picture has clear implications for how we interpret its key term "the world" in the 1966 lecture. Given the above, it seems safe to conclude that, for Tarski, "the world" is meant to refer to the universe of physical objects, in agreement with his professed nominalism. Thus, when, in the lecture, he speaks of the "world", he intends it literally. This can be obscured by the fact that several times when mentioning the term "the world" (such as in the quotation above), Tarski mentions it together with "universe of discourse", which may give the impression that he is talking metaphorically, as if "the world" was merely a non-empty set whose elements are considered as basic from the point of view of the hierarchy of types.[12] Considered this way, it's natural to ask what happens when we consider different "universes of discourse", that is, when we consider different sets as possible domains for a typed hierarchy. But that doesn't seem to be Tarski's point of view. In a revealing parenthetical remark, he says:

> When we speak of transformations of the 'world' onto itself we mean only transformations of the basic universe of discourse, of the universe of individuals (which we may interpret as the universe of physical objects, although there is nothing in *Principia Mathematica* which compels us to accept such an interpretation). (Tarski 1966/1986, p. 152)[13]

[10]This was later relaxed to (1') does not refer to an infinite number of individuals; cf. Frost-Arnold (2013, p. 156). Although the informal tone of the remarks may suggest that this is merely a stylistic variant, I would say that (1) implies that they are explicitly adding to the meta-theory a statement to the effect that there are only finitely many individuals, whereas with (1') they merely not adding an axiom of infinity.

[11]Incidentally, this last point already casts doubt on Rodríguez-Consuegra's interpretation according to which Tarski thought we could somehow physically perceive (finite) sets; cf. Rodríguez-Consuegra (2005, esp. p. 256). I'll come back to this when I discuss Tarski's preference for a type-theoretical formulation of his proposal.

[12]This seems to be view taken, e.g., by Hitchcock in his introduction to (Tarski 1936/2002); cf. his gloss on the term "the world" on p. 169 of the aforementioned introduction.

[13]The hedge about *Principia Mathematica* is curious—obviously, there's nothing there that *forces* this inter-

Interestingly, this connects with another important controversy surrounding Tarski's work, especially his account of logical consequence; in particular, there is a debate as to whether he employs a fixed domain or a variable domain approach.[14] From the available evidence, it's a tentative conclusion that he employed a fixed domain approach, where the domain in question is the class of all (physical) individuals. This will thus point, for better or worse, to a remarkable coherence throughout Tarski's career: although during the 50's he started using a multiple domains approach, as it's technically more convenient, philosophically he seems to have always been inclined towards a fixed domain approach. As we have seen, there's a philosophical reason for that: Tarski has always been inclined towards both physicalism and nominalism, and a fixed domain approach can model those ideas better than a variable domain one.[15]

2.2.2 Nominalism and Type-Theory

We have seen in the preceding section how Tarski endorsed a rather strong anti-platonism with a characteristic commitment to nominalism. Specifically, we have seen how Tarski tied his own nominalism with certain linguistic considerations, mentioning that he could not understand languages which quantified over sets, properties, etc.. Indeed, in his conversations with Carnap, he even expresses the "wish" for the future disappearance of set theory: "It would be a wish and a guess that the entirety of *general set theory*, as beautiful as it is, will *disappear in the future. With the higher types, Platonism begins.*" (Frost-Arnold 2013, p. 141, original emphasis) Further evidence can be found in his rejection of Corcoran's work on concatenation theory as "meaningless", precisely because it relied on second-order logic.[16] This rejection tale is told by Corcoran and Sagüilo (2011, p. 372) as part of a developmental story according to which there is a radical discontinuity between the Tarski from Warsaw (mid-1930's) and the Tarski from Berkeley (mid-1940's onward). It is thus clear that Tarski accepted the Quinean linguistic thesis: a theory's ontological commitment should be read off the range of its existentially quantified variables.

This creates a puzzle. As Bellotti (2003) argues, the most natural environment in which

pretation, though it may be of interest to remark that Russell and Whitehead also worked with a fixed domain approach, the difference being that their domain didn't consist just of *physical* objects.

[14]Cf. Mancosu (2010a) for a summary of the debate, including the most important references.

[15]This seems to be in agreement with Bellotti (2003), who also comments on this coherence throughout Tarski's career.

[16]The work in question is Corcoran et al. 1974. On the other hand, Tarski also made some remarks in Tarski (2007, p. 263) that may be relevant here. Basically, he says there that, due to the uncertainty in the foundations of set theory, any categoricity result which relied on "artificial assumptions" (e.g. the non-existence of large cardinals) is some "kind of deception". I haven't examined Corcoran et al's result in depth (it apparently shows that string theory is bi-interpretable with second-order Peano Arithmetic), but maybe Tarski had something like this in mind?

to couch Tarski's proposal is type theory. Although he suggests towards the end of the lecture that the proposal could be recast in terms of first-order set-theory, there's considerable evidence that he considered the type-theoretical approach more natural. Firstly, as we have already seen, Tarski defines the term "notion" by reference to type theory:

> I use the term 'notion' in a rather loose and general sense, to mean, roughly speaking, objects of all possible types in some hierarchy of types like that in *Principia Mathematica*. Thus notions include individuals (points in the present context), classes of individuals, relations of individuals, classes of classes of individuals, and so on. (Tarski 1966/1986, p. 147)

Secondly, most of the examples he gives are most naturally interpreted inside a hierarchy of types (not surprisingly, that's exactly how most commentators have interpreted them), especially given the way he talks about "levels". Thirdly, in the only other place in which he discusses the matter (Tarski and Givant 1988, §3.5), the discussion is explicitly couched in type-theoretical terms. Does that mean we should reject Tarski's thesis as "meaningless" as well, like he himself rejected Corcoran et al. 1974? And, if Tarski preferred working in a first-order framework, why didn't he do so? The last question is particularly pressing, given that he himself indicates at the end of the article that it's possible to do so. So why did he opt for type theory?

The answer is a bit speculative, but I believe that the reason is related to his nominalism, as analyzed in the last section. We saw there that the problem with type theory was that it quantified over classes, and that it was this quantification that brought with it a commitment to abstract entities. If *this* is the problem, merely shifting to a first-order theory such as ZFC would not solve the problem, because then, as Tarski puts it, "individuals and sets are considered as belonging to the same universe of discourse"(Tarski 1966/1986, p. 153), so that they are considered to be on a par. Therefore, if Tarski had adopted a first-order approach, then the platonistic commitment to sets would have been unavoidable, since they would be in the range of existentially quantified variables. This conclusion is probably the reason why he didn't opt for this approach. On the other hand, adopting type theory has, as an attractive feature, that "only the basic universe, the universe of individuals, is fundamental" (Tarski 1966/1986, p. 152). Hence, if it were possible to construe type theory in such a way to make it nominalistically acceptable, this would probably explain why Tarski adopted a type-theoretical approach.

Of course, another possibility is just that, in the 1966 lecture, Tarski was merely presenting a working hypothesis, so to speak. That is, he would be operating under a tacit "if-thenism": if sets are real and if it makes sense to quantify over them, then the logical operations are the invariant under all permutations of the base domain. Type theory would

then be chosen as the basic framework in which to develop this suggestion simply because it was more convenient to work within it, especially since Tarski had already proven, in an article with Lindenbaum (Tarski and Lindenbaum 1935/1983), a number of the type theoretic results he appeals to in the article, including that every notion definable in the simple theory of types is invariant under all permutations of the domain. Indeed, some parts of the lecture are extracted almost *verbatim* from that article. From a historical point of view, this points to a further hypothesis: perhaps this lecture is similar to Tarski (1969), which Patterson describes as "clearly a set of mothballed remarks from the 1930's" which show that "to the extent he had given the topics [of the article—D.N.] any thought at all, his views had not changed" (Patterson 2012, p. 229). That Tarski thought of the two lectures as companion pieces could be taken as indicative of this similarity; more than anything, such repetition would just confirm that, after Tarski's experiences with the positivists in the late 1930's, he thought it better to remain silent over philosophical matters.

We have already seen in the first chapter, however, how Tarski's views on definitions changed from the 1930's to the 1960's, so these pieces are not just "mothballed remarks" from the 1930's. So the historical picture sketched in the last paragraph is implausible. As for the proposal that Tarski was only using "working hypothesis" in the 1966 lecture, it does seem plausible, and it would corroborate the interpretation of those like Sinaceur (2009), who see a distinction between Tarski's philosophical inclinations as revealed by his pronouncements and his philosophical inclinations as revealed by his practice, the latter being much more open-ended and exploratory than the former. That is, Tarski's mathematical practice would be much more driven by pragmatic considerations than his philosophical inclinations would show.[17] Nevertheless, if it were possible to square the 1966 lecture's commitment to type theory and Tarski's nominalism, that would surely be an interesting investigation on its own, particularly if it followed a procedure suggested by Tarski himself.

And, in fact, Tarski does suggest at least two procedures for interpreting type theory in a nominalistic way, which could perhaps be used in this context. In the next section, I will explore these strategies and their viability.

2.2.3 Two Nominalist Strategies

In his conversations with Carnap, right after the above quoted remark according to which platonism begins with the higher types, Tarski makes the following observation:

[17]Something that would make Corcoran suggest, half-jestingly, that maybe Sinaceur got things reversed and that "perhaps Tarski's most basic philosophical temperament was a Platonism that led him into fields requiring Platonist premises while his avowed materialism was pragmatically motivated to make his work more palatable to positivists and to preserve at least an appearance of loyalty to humanism and to his mentor [Kotarbiński—D.N.]." (Corcoran 2011). If that were true, then the platonic considerations in the next chapter would actually be congenial to Tarski, and the 1966 lecture would be much easier to interpret.

> The tendencies of Chwistek and others ("Nominalism") to talk only about designatable things are healthy. The only problem is finding a good implementation. Perhaps roughly of this kind: in the first language numbers as individuals, as in language I, but perhaps with unrestricted operators; in the second language individuals that are identical with or correspond to the sentential functions in the first language, so properties of natural numbers expressible in the first language; in the third language, as individuals those properties expressible in the second language, and so forth. Then one has in each language only individual variables, albeit dealing with entities of different types. (Frost-Arnold 2013, p. 141)

The procedure described in the above quotation seems to be an instance of what Burgess and Rosen (1997) call a *substitutional strategy* for defending nominalism: the idea is to try to avoid certain ontological commitments by employing substitutional quantification (hence the name).[18] This strategy exploits the fact that truth has already been defined in a given language L_0 to construct a truth definition for a new language L_1 that does not employ the notion of satisfaction, but instead defines truth directly using as a basis for the definition the set of true sentences from L_0. I describe this procedure more formally in the appendix to this chapter, building on Kripke (1976); here, I will just sketch the overall idea.

According to this strategy, our Tarskian nominalist starts with a nominalistically acceptable language L_0 and a nominalistcally acceptable theory T_0 (closed under logical consequence) stated in L_0.[19] For simplicity, suppose the logical vocabulary of L_0 consists of the standard first-order quantifier $\exists$, and two propositional connectives, say $\wedge$ and $\neg$, with the other symbols defined in their usual way. We then introduce a new language, L_1, with a new quantifier, say $\exists_1$, and new propositional connectives, say $\wedge_1, \neg_1$.[20], along with infinitely many variables not in L_0, such as $x_0^1, x_1^1, \ldots$. These variables will range over the formulas of L_0. The intuitive idea is that, e.g., $\exists_1 x_i^1 x_i^1(t^0)$ is true iff there is a formula ϕ from L_0 such that $\phi(t^0)$ is true (t^0 is a term from L_0); Kripke (1976) shows that this intuitive idea can be rigorously formulated so that truth for L_1 is well defined (again, for an outline of his procedure, cf. the appendix).

Apparently, then, Tarski's idea is to extend this construction further, so that L_2 has its

[18] Interestingly, they connect this strategy with Tarski's teacher, Leśniewski. For some support for this attribution, Tarski himself mentions Leśniewski in this connection in a 1953 lecture on nominalism, if we are to trust Beth's report. Cf. the quotation from Beth by Mancosu (2010c, p. 406).

[19] In their conversations, Carnap, Quine, and Tarski apparently opted for a very weak form of arithmetic as a "toy" theory, preferably one that couldn't decide even whether the domain was finite or infinite. This means that the theory would be weaker than the theory R studied by Tarski et al. (1953/2010).

[20] Generally, the new quantifier is denoted by Σ or some other symbol; it's important that the chosen symbol be different from the symbol chosen in the original language. Since we will be working with a hierarchy of languages, I thought it simpler to add a subscript.

own existential quantifier $\exists_2$ ranging over formulas of L_1, L_3 has $\exists_3$ ranging over formulas from L_2, etc. Presumably, he wanted to consider L_ω in which one would have existential quantifiers of all types, resulting in a substitutional version of the finite simple theory of types. The basic idea is to trade an ontology of sets for the linguistic device of substitutional quantification. How plausible is this strategy?

As Burgess and Rosen (1997) mention, one problem is that the above approach is committed to there being infinitely many expressions, perhaps even infinitely many expression types. This is, of course, a general problem for any nominalist strategy: since most specifications of formal languages allow for the concatenation of any expression with any expression, and this operation can also be iterated, it's clear that in general formal languages will contain infinitely many expressions. Indeed, in light of this obvious fact, Quine and Goodman (1947, p. 106) adopt the rather eccentric strategy of considering as expressions "all appropriately shaped spatio-temporal regions even though they be indistinguishable from their surroundings in color, sound, textture, etc.", and, similarly, Tarski (1933/1983, p. 174n2) suggests that "we could consider all physical bodies of a particular form and size as expressions". As Wetzel (2009, p. 101) notes, this has the bizarre consequence that, if the Goldbach conjecture is provable, then there is a proof of it written *somewhere* (perhaps as an arrangement of certain subatomic particles), although nobody has written it yet. Be that as it may, it's not entirely clear that this strategy will produce infinitely many expressions; as Quine and Goodman (and also Tarski) note, this essentially depends on whether the universe is itself infinite or not. Moreover, even leaving this problem to the side, and despite Quine and Goodman's efforts, it's not obvious that this construction is able to avoid commitment to expression *types*—for instance, in specifying the syntax of the language, we employed locutions such as "*the* quantifier, *the* propositional connective", etc. (the obvious idea of using a relation of "likeness of shape" isn't useful here, since *shape* is presumably an abstract type). Perhaps this commitment can be avoided by adopting some form of *resemblance nominalism*, i.e. for two individuals to be of the same type is for them to *resemble* each other in some way, with resemblance being a primitive relation.[21]

In any case, whether or not the Tarskian nominalist can eliminate types may be considered beside the point. After all, we have traded an allegedly obscure ontology of sets for one of types, which may be taken to be a good deal, provided that types are less obscure than sets. Moreover, some of the problems may be mitigated by some suggestions from Tarski's second nominalization strategy.

A sketch of this second Tarskian strategy can be found in the writings of Evert Beth. As told in Mancosu (2010c), in the summer of 1953 Beth organized a meeting in Amersfoort

[21] An important recent defense of resemblance nominalism is Rodriguez-Pereyra (2002). For discussion, cf. Guigon and Rodriguez-Pereyra (2015).

to discuss "Nominalism and Platonism in Contemporary Logic", whose main speakers were Tarski and Quine. Although apparently no extant typescript of Tarski's lecture survives,[22] it's possible to extract the content of such lecture from Beth's writings; in particular, Beth (1970, p. 94–6) proposes the following program for developing a nominalistic acceptable reconstruction of mathematics, which he claims follows "in the main lines the exposition given by Tarski in Amersfoort" (Beth 1970, p. 100n16).

The program proceeds in three basic steps. In the first step, we identify the domain of objects with the individuals, that is, the basic objects which we will admit in our ontology; this domain can be called, for convenience, S_1, whose elements will be identified with *material bodies*. In the second step, we build a (finite) type hierarchy, but instead of taking the full power set of the previous domain, at each level we add only the *definable* subsets of the previous set. The construction seems to be very similar to Gödel's procedure and indeed Beth mentions his papers in this connection.[23] Now, in order to obtain the whole of mathematics, we need to appeal to a "cosmological hypothesis" which asserts that there are infinitely (countably) many material bodies. Since this hypothesis is highly contentious, the final step consists of stating every theorem that depends on it in a conditional form; i.e., we use the deduction theorem to "eliminate" the assumption of such hypothesis and replace it by the respective conditional (a kind of "if-thenism"). Almost as an aside, Beth mentions the objection that, since this hierarchy will produce at most a countable set (as, assuming choice, the countable union of countable sets is itself countable), and since mathematics requires uncountable sets, this won't suffice as a nominalist reconstruction of mathematics. Beth's answer is an appeal to the Löwenheim–Skolem theorem: the needed uncountable sets are uncountable *inside* the model, but the fact that a model is countable can only be assessed from *outside* the model. Thus, there could be relatively uncountable objects living inside the type hierarchy, and these are all we need in order to get mathematics going.[24]

It is not entirely clear how successful such strategies are. First, there are the worries raised by Burgess and Rosen (1997), for instance that these strategies privilege the simplicity virtue to the detriment of other dimensions that may be more relevant in theory assessment, such as fruitfulness, technical expediency, etc. Indeed, as we have seen, Tarski himself is sensitive to this fact: not surprisingly, when it came to his own mathematical practice, he eventually

[22] As Mancosu (2010c, p. 559n4) notes, it's likely that Tarski never wrote the text of his lecture. Cf. the letter from Tarski to Quine quoted by Mancosu, in which Tarski states that "it would be too late" for him to "prepare any formal talk". Interestingly, the excerpt quoted by Mancosu of this letter ends with Tarski saying to Quine that he wants to examine the "possibility of a semantic interpretation of quantifiers with variables of higher orders".

[23] For an in-depth exposition of the constructible universe, cf. Devlin (1984).

[24] Actually, a stronger result is possible: there are models of ZFC in which every set is definable without parameters. For a discussion with this result in relation to a similar argument, viz. that there are undefinable reals because there are only countably many formulas, cf. Hamkins et al. (2012).

came to adopt ZFC as his preferred environment, in large part because doing so proved to be technically much more expedient than working even with the simple theory of types. Second, as we have seen, it seems that commitment to some abstract entities is unavoidable (especially to types). But if you are willing to admit types, why not go all the way and embrace a technically more expedient ontology of sets?

I won't explore here these questions further, however.[25] My purpose in this section was merely to indicate a possible rationale for Tarski's preference for type-theory in his 1966 lecture, when it seems so much at odds with the rest of his philosophical inclinations. It may very well be that Tarski himself was not entirely convinced of such strategies—which would also explain why he never bothered to put them into more definitive form, relegating them to private conversations and unpublished remarks.

2.3 Conclusion

In an intriguing article, McGee (2004) explores what he considers to be "Tarski's staggering existential assumptions". In the article, McGee concludes that Tarski's proposal requires that, if it's possible for there to be a model for a given theory, then there is a model for that theory. In other words, consistency implies existence, so that, if "there is a supercompact cardinal" is consistent relative to ZFC, then there really is a supercompact cardinal (McGee's example).

In this chapter, we have seen that Tarski may have indeed "staggering existential assumptions", but in the opposite direction: while McGee credits Tarski with the assumption of too many abstract entities, I have argued that, in fact, at least in his more "philosophical" moments, he shares Quine's taste for desert landscapes, even outlining his own strategy for a nominalist reduction of mathematics. It would be interesting to see the consequences of this assumption for his account both of logical consequence and logical notions, which are the main concern of McGee's paper. In any case, what is striking here is precisely this contrast between the staggering existential assumptions of his mathematical work and his nominalist inclinations. It is not surprising, therefore, that Tarski's preferred description of his own philosophical views was as a "tortured platonist".

Another striking feature of Tarski's inchoate remarks about nominalism which I tried to highlight here was how in tune he was with two leading giants of the analytic tradition, namely Carnap and Quine, and how his own tortured outlook seemed to be in a sense torn between those two. In the previous chapter, we saw how close his own conception of his conceptual analysis was to Carnap's, in particular after 1937. Similarly, in this chapter we have seen how Tarski shared with Quine the latter's perspective on ontological questions,

[25] Beth himself seems to think they are answerable, as he comments that: "That the nominalistic interpretation of the different forms of contemporary set theory is *tenable* can hardly be disputed from the above." (Beth 1970, p. 96, original emphasis).

both in taking them seriously and in considering that the best way to deal with them is by investigating the language in which our best theories are couched. In this respect, we have also seen how Tarski anticipated Quine and Goodman's paraphrasing strategy for defending nominalism, a strategy that would latter become ubiquitous in the nominalist literature.[26]

A Appendix: Kripke on substitutional quantification

The account here is basically the one presented in Kripke (1976) (indeed, the following is basically a paraphrase of Kripke's account).

By way of introduction, let me note that the essential idea behind substitutional quantification seems to be that, instead of using *satisfaction* conditions in order to recursively establish a truth definition for the language in question, these languages instead exploit directly the notion of truth. The idea is that, certain sentences of a base language L_0 being given as true, it's possible to expand this language into a language L whose variables will range over those sentences. Let's see how to work this out more formally.

In the larger language L_1, the sentences from L_0 will be the *atomic* sentences. Now separate some subset of expressions from L_0 as the subset C. This subset will be called the *substitution class* of L; the elements of C will be the *terms* of L. In our case, C will be the set of all formulas from L_0.

Now let $x_0^1, x_1^1, \ldots$ be an infinite list of variables not in L_0. Let ϕ be a sentence from L_0. We call an expression ϕ' obtained from ϕ by replacing zero or more terms in ϕ by variables an (atomic) *preformula*, or simply *preform*. If the result of replacing variables by arbitrary terms in an atomic preform is always itself a sentence of L_0, the preform is an (atomic) *form*.

In order to expand L_0 into L_1, specify some given set of forms as *atomic formulae*. These will include all the sentences of L_0, and, if $\phi(x_{i_1}^1, \ldots, x_{i_n}^1)$ is an atomic formula, so is the result of replacing the listed variables with others. Given this definition of atomic formulae for L^1, it's now possible to recursively specify the set of *well-formed formulas* (wffs) of L^1:

1. an atomic formula is a formula;
2. if ϕ and ψ are formula, then so is $\phi \wedge \psi$;
3. if ϕ is a formula, so is $\neg\phi$;
4. if ϕ is a formula, so is $(\exists_1 x_i^1)\phi$ for any $i \in \omega$.

Wffs without free variables are called *sentences* of L. In contrast to the case where the language has objectual quantifiers, in which wffs which are not sentences can be 'satisfied' by

[26] Though it may now be receding. Cf., e.g., Field's second thoughts on the matter in the "Preface" to the second edition of *Science Without Numbers* (Field 2016).

sequences, here, wffs which are not sentences won't be assigned any semantic interpretation, playing thus a merely auxiliary role.

Finally, we characterize truth for L_1, *assuming that truth has already been characterized* for L_0. So let S_0 be the set of true sentences of L_0, we define the set S_1 of true sentences of L_1 as the set which obeys the following conditions:

1. $S_0 \subseteq S_1$;
2. $(\neg\phi) \in S_1$ iff $\phi \notin S_1$;
3. $(\phi \wedge \psi) \in S_1$ iff $\phi, \psi \in S_1$;
4. $(\exists_1 x_i^1)\phi \in S_1$ iff there is a term t such that $\phi' \in S_1$, where ϕ' comes from ϕ by replacing all free occurrences of x_i^1 by t.

Given these conditions, Kripke proves the following theorem, which shows that the above is well-defined:

Theorem A.1. *Given a set S_0 of true sentences from L_0, there is a unique extension S_1 in L_1 that obeys the above conditions.*

Proof. There are two parts to the theorem: one involves uniqueness, the other existence.

Let's start with the uniqueness part, which we will prove by induction on the complexity of the sentences of L. Suppose S' and S'' are two extensions of S in L that obey the above conditions. The base case is trivial: since the true atomic sentences of L are taken from L_0, an atomic formula $\phi \in S' \iff \phi \in S \iff \phi \in S''$. Now suppose (the induction hypothesis) that $\theta, \psi \in S' \iff \theta, \psi \in S''$ and let $\phi = \theta \wedge \psi$. Given these equivalences and condition (3), it follows that $\phi \in S' \iff \phi \in S''$; similarly for $\phi = \neg\psi$. Finally, let $\phi = (\exists_1 x_i^1)\psi$ and suppose, without loss of generality, that $\phi \in S'$. This means that, for some formula ψ', $\psi' \in S'$. But, by the induction hypothesis, $\psi' \in S''$. Thus, $\phi \in S''$ as well (the converse is established in a similar way). From the preceding, we can conclude that a sentence $\phi \in S' \iff \phi \in S''$, whence, by extensionality, $S' = S''$.

The existence claim is a bit more complicated. We need to recursively specify a way of expanding the given set S into S' (this will be similar to constructing a Hintikka set). First, arrange all the wffs of L in a list $\phi_0, \phi_1, \ldots$ (for simplicity, I will assume here that L is countable; the uncountable case is anyway similar). Set $S_0 = S$ and, assuming as given S_n, define S_{n+1} depending on the wff ϕ_n:

- Case 1: ϕ_n is an atomic formula. Then this is already covered by S_0;
- Case 2: $\phi_n = \psi \wedge \theta$. If $\psi, \theta \in S_n$, set $S_{n+1} = S_n \cup \{\phi_n\}$. Otherwise, set $S_{n+1} = S_n$;

- Case 3: $\phi_n = \neg\psi$. If $\psi \in S_n$, then set $S_{n+1} = S_n$. Otherwise, set $S_{n+1} = S_n \cup \{\phi_n\}$;
- Case 4: $\phi_n = (\exists_1 x_i^1)\psi$. If there is a $\psi' \in S_n$ such that ψ' is the result of substituting a term t for all free occurrences of x_i in ψ, set $S_{n+1} = S_n \cup \{\phi_n\}$. Otherwise, set $S_{n+1} = S_n$.

Now set $S' = \bigcup_{n\in\omega} S_n$. It's clear that S' is the desired set: $S \subseteq S'$ and every true sentence ϕ from L will be in S. This last claim is provable by induction on the complexity of ϕ: if ϕ is atomic, then $\phi \in S$, whence $\phi \in S'$. If $\phi = \psi \wedge \theta$, since our enumeration listed all the wffs of L, it will eventually fall under Case 2 above; similarly for the negation and existential case. ■

Note that, as Kripke (1976, pp. 331-2) observes, the proof above crucially uses the assumption that L_0 does not contain the same connectives and quantifiers as L. Otherwise, t could be a term containing one of those, and the substitution of t for x_i in ψ could result in a formula ψ' of greater complexity than ψ or even $(\exists_1 x_i^1)\psi$.[27]

Finally, since we are only interested here in its use in elucidating Tarski's nominalistic strategy, we will not further develop the semantics for this language. The interested reader should consult Kripke's insightful article.

[27] As noted in Kripke (1976, p. 332, esp. the digression), that's not to say that our theorem would then be impossible to prove; rather, it's just to say that the above proof, which relies on an induction on the complexity of the sentences of L, would not work. There could be other ways of proving the theorem, and Kripke himself sketches some conditions for alternative proofs in the passage mentioned.

PART II

Tarski's Proposal

Chapter 3

The Proposal

In this chapter, I will explore some metaphysical themes that emerge out of Tarski's 1966 lecture, in particular his extension of Klein's Erlangen Program. This development will be distinctly non-Tarskian, in that it will involve both an elaborate metaphysical picture (against the Carnapian strand in Tarski's thought) and a defense of robust platonism to boot (against his Quinean tendencies). Nevertheless, there is a sense in which Tarski's work actually encourages this picture, both because of his focus on *notions* and his recourse to Klein's ideas. Let's start by discussing the first topic.

A natural question to ask regarding Tarski's lecture is: what are these *notions* mentioned in the lecture's title? Are they concepts, expressions, or objects? One important piece of data here is the fact that the use of this term "notion" is remarkably stable throughout Tarski's career. Besides figuring in the 1966 lecture, it also figures in the title of Tarski and Lindenbaum (1926), it is at work at least implicitly in Tarski and Lindenbaum (1935/1983), and also in Tarski's paper "On a General Theorem Concerning Primitive Notions of Euclidean Geometry" (Tarski 1956).[1] In all those papers, the usage is ambiguous between two closely connected readings: on the one hand, a notion is *an object in a type-theoretical hierarchy*; call this the *object* reading. On the other hand, the term can also refer to *an interpreted expression whose meaning is an object in a type-theoretical hierarchy*; call this the *concept* reading.[2]

This double meaning is most clear in the last of the previously mentioned papers. In the course of the paper, Tarski repeatedly makes use of the following proposition:

> (I) *If the notion N is definable in terms of the notions $N_1, \ldots, N_m$ within a given theory, then N is invariant under every one-to-one transformation of the universe of discourse*

[1] Interestingly, all of these papers deal with geometrical notions in one way or another.

[2] Villegas-Forero and Maciaszek (1997) argue that there is a distinction between logical *notions*, as what they call "Fregean concepts" (which are objective entities), and logical *entities*, which would be akin to a Fregean object. They don't develop the distinction further, however, so it's unclear how to relate their discussion to mine. In any case, the analogy with Fregean concepts is suggestive, so that's why I called this the *concept* reading.

> *of this theory onto itself under which all the notions $N_1, \ldots, N_m$ are invariant.* (Tarski 1956, p. 469, original italics)

Appended to this passage is the following footnote:

> In this formulation (I) applies to interpreted theories, i.e., to theories which are provided with definite interpretations of symbols occurring in them; moreover, each such theory is assumed to have a well determined universe of discourse—a set U such that all the notions of the theories are intrinsic with respect to U (i.e., they are subsets of U or relations between elements of U or relations between these subsets and relations, etc.). The formulation of (I) must be modified if (I) is to apply to non-interpreted theories. In this case we speak of the definability of a constant C in terms of other constants $C_1, \ldots, C_m$; we consider all the models of a given theory—each such model $\mathfrak{M}$ is formed by a set $U(\mathfrak{M})$ and by certain *notions* [emphasis mine—D. N.] $N(\mathfrak{M}), N_1(\mathfrak{M}), \ldots, N_m(\mathfrak{M}), \ldots$ which are intrinsic with respect to $U(\mathfrak{M})$ and which (because of their logical structure) can serve as interpretations of the constants $C, C_1, \ldots, C_m$. (Tarski 1956, p. 469n)

It's thus clear that "notion" can't mean the same thing in (I) and in the footnote. In (I), N is an expression from an interpreted theory (the concept reading), whereas in the footnote "notion" is used to refer to objects in a type theoretic hierarchy (the object reading). As mentioned, all papers above alternate between the two readings of "notion", and sometimes the expression is rather ambiguous between them. In any case, it's clear that the object reading is the fundamental one: an expression N is considered as a "notion" in the concept reading *because* it refers to a "notion" $N(\mathfrak{M})$ in the object reading, as the above footnote makes clear. This makes the ambiguity pointed out here mostly harmless; nevertheless, drawing attention to it helps to bring to the forefront an apparently neglected aspect of the 1966 lecture, which is precisely that the target of Tarski's explication is the logical *notions*.

In his lecture, Tarski again makes clear that he privileges the object reading of "notion": "I use the term 'notion' in a rather loose and general sense, to mean, roughly speaking, objects of all possible types in some hierarchy of types like that in *Principia mathematica*" (Tarski 1966/1986, p. 147). Here Tarski's previous discussion of Klein's project is illuminating. As we will see when analyzing Klein's proposal in more detail, there is a kind of abstraction principle implicitly at work in Klein's analysis, which allows him both to move from object tokens (e.g. a specific triangle) to object types (e.g. triangle), as well as to distinguish between which properties of an object belong to that object *qua* object of a certain type and which properties hold of an object merely accidentally, so to speak. Given the relevance of this Kleinian element for my reading of Tarski's proposal, the first sections of this chapter will be dedicated to a detailed analysis of these ideas.

3.1 Klein's Strategy and the Nature of Types

In order to best appreciate the power of Klein's approach, I will first sketch a geometrical problem about reasoning with diagrams that can be solved by his group-theoretic methods. This will allow me to introduce the concept of a *projectible property*, which will play a relevant role later on.

3.1.1 A Kantian Predicament

Kant infamously remarked that all mathematics rests on intuition;[3] in the case of geometry, Kant thought that geometrical proofs essentially involved spatial intuition, mostly in the forms of diagrams, that were either imagined or drawn in a piece of paper (cf. the first chapter of the Transcendental Doctrine of Method).[4] Take, for instance, his discussion (A 716/B 744) of the Euclidean proof that the angles of a triangle sum to two right angles. As Kant describes it, the geometer's demonstration amounts to constructing a triangle and some auxiliary figures (such as auxiliary lines), and then "reading off" the fact that its angles sum to two right angles from the resulting figure. This creates a puzzle: since the drawing or imagined figure is always a *singular* representation, how can one read off general conclusions from such representations? Note that Kant clearly recognizes that (Euclidean) geometry deals not with properties of this or that particular representation (or token) of a triangle, but rather with properties that belong to the triangle figure as a *type*. It's then possible to restate the above puzzle as: how is it possible to derive properties of the type from a given token? As Coffa (1991, p. 46) notes, the puzzle puts Kant in a dilemma: either we can reason directly about the type, thus making the reference to its tokens (or intuitions) otiose, or else we need some criterion to select the properties to which we can be indifferent, and then it's no use saying that we must be indifferent to the properties that don't make a difference when reasoning about triangles. Let's call this puzzle *the Kantian predicament*.

Of course, current mathematical practice makes it easy to solve the Kantian predicament: we attend precisely to those properties that follow strictly from the axioms and definitions, so that reference to intuitions is indeed otiose, and this is the end of the matter. But I want here to explore this issue a bit further, since I believe it reveals interesting features of the type-token relationship, features that will be relevant in our examination of Klein. Let's start with a definition:

[3] Coffa (1991) is a good account of how much of post-Kantian philosophy arose as an attempt to show that that claim is false. In fact, the present account owes much to his observations in chapter 3 of that book.

[4] I'll cite Kant according to the standard practice of citing the pages of the first (A) and second (B) editions of the *Critique of Pure Reason*; the edition I'm using is (Kant 1998). Since my aim here is not Kant's exegesis, I'll not engage in the controversy surrounding Kant's account of geometrical reasoning. For an up-to-date account and defense of the Kantian line, cf. Vinci (2015).

Projectible properties: Consider a property F and a type T. The property F is said to be *projectible across the type* T if, and only if, for every token a of T there is some kind of principle connected with T that necessitates that Fa.[5]

This is a provisional definition, of course, since it is extremely vague. What does it mean for a principle to be connected with a type? And what kind of necessity is involved in the definition? In fact, the next sections will be devoted precisely to the task of spelling out these details, beginning in the next section, in which I argue that this problem has important connections to neo-Fregean abstractionism.

3.1.2 Bromberger's account

The Kantian predicament, as analyzed in the last section, bears a striking similarity to the problem that Bromberger (1992, chap. 8) deals with in "Types and Tokens in Linguistics".[6] In that essay, Bromberger (1992, p. 176) calls the *Platonic Relationship Principle* the principle that allows one to infer properties of types (which are "platonic" entities) from properties of tokens (which are empirical). This problem is a version of the Kantian predicament, as it can be rephrased as the question of how to detect which properties of linguistic tokens are projectible across their types.[7] Bromberger's own answer to this problem is too complicated to be rehearsed here (especially given its reliance on his peculiar account of *questions*), but nevertheless I want to retain two of its aspects that may be relevant for what follows.

The first is his emphasis on quasi-natural kinds, one of his technical terms that I will parse here very loosely as a set of entities which share some characteristics as a matter of nomological necessity;[8] notice that, in spite of the moniker, this notion does not *exclude* natural kinds from its purview—natural kinds are a *subset* of quasi-natural kinds, and in fact Bromberger's favored example of a quasi-natural kind, samples of mercury, is a natural kind. In any case, the important point to note here is Bromberger's appeal to *nomological necessity* in order to account for natural kinds: it is this component that specifies which properties are projectible across the type, namely those that somehow follow from the nomological connections among them.[9] Indeed, this points to the explanatory role of nomological ne-

[5]This definition is inspired by Heck (2017). I'm grateful to prof. Heck for clarifying some of the issues surrounding it via electronic correspondence.

[6]I am grateful to Richard Heck for calling my attention to Bromberger's work.

[7]As the vocabulary employed here makes clear, this problem also has connections to Goodman (1983), especially to what he calls "the new riddle of induction". I won't be able to explore these connections here, however. Indeed, if Hirsch (1993, chap. 2) is right that Goodman is worried more with projectible terms than properties, it may be that my problem here is slightly different than his.

[8]Cf. Bromberger (1992, p. 183) for the more technical definition.

[9]Hence, Bromberger's strategy is actually an inversion of Goodman's: whereas Goodman wanted to explain laws in terms of counterfactuals and these in terms of projectible qualities, Bromberger wants to appeal

cessity in Bromberger's account: it is *because* these properties follow from certain laws that we may take them to be projectible. Thus, going back to the problem of the last section, it is tempting to say that Bromberger would modify our provisional definition of projectible property into something like this:

Proto-Brombergerian definition of projectible properties: For any type T and property F, we say that F is *projectible across* T if, and only if, for every token a of T, it is nomologically necessary that Fa.[10]

This would be a bit premature, though. To see why, remember that Bromberger is interested in the Platonic Relationship Principle, namely how to determine properties of the types from properties of the tokens. But not every property projectible across a type according to the above Proto-Brombergian definition is a property of the type itself. Consider a word type, say *Arial*. Every token of this type necessarily has a certain length, if we assume that complex[11] spatial objects all have some length; does it follow that the word type *Arial* has a certain length?

In order to deal with this kind of problem, Bromberger introduces a distinction between projectible properties, w-projectible properties, and individuating properties. The former are properties which not only are projectible in the above "Proto-Brombergerian" sense, but also which assume a specific, nomologically *determined* value for every token, e.g. the boiling point of a specific substance. On the other hand, w-projectible properties are *determinable* properties, which all tokens share, but whose values may vary, even though this variation is also nomologically determined, e.g. the length or temperature of a sample of a substance at a given time. Finally, individuating properties are those properties which vary for each token, this variation not being nomologically determined, e.g. their spatio-temporal location.

Using this new distinction, Bromberger then proposes that the Platonic Relationship Principle holds only for projectible properties, not w-projectible properties. Thus, since *having a certain length* is a w-projectible property, it follows that the type *Arial* does not have a certain length. There is still a problem, however, with determinate properties that are too general, such as *being concrete*: they are clearly shared by all tokens of *Arial*, but the type itself, being abstract, is not concrete! This generality itself may be a clue to solving the problem: projectible properties are in a sense characteristic of a type, so they are shared only by tokens of that type. *Being concrete* is not a characteristic property in this sense, so

to laws in order to explain projectible qualities.

[10]Notice that Brombeger himself does not define projectible properties; this is an extrapolation based on his account of quasi-natural kinds.

[11]The adjective "complex" is there in order not to rule out the possibility that points are spatial objects with no length.

we can rule it out. Thus, we can state the following definition of projectible property along Brombergerian lines:

Brombergerian projectible property: A property F is *Brombergerian projectible across a type* T if, and only if, the following holds: (i) for every token a of T, it is nomologically necessary that Fa; (ii) if F is a determinable property, then the specific value of F is nomologically determined to be the same for every token a of T; (iii) if b is not a token of F, then F does not hold of b.

Recall that, in the last section, I provisionally defined a property to be projectible across a type iff there was some kind of principle connected with the type and which necessitated the property. It's now possible to replace this vague idea with something more precise: the necessitation in question is nomological necessity, and this is connected to the type insofar as there is a law which *explains* why the tokens of the type, and only tokens of that type, have this specific property.

Unfortunately, however, this means that at least this part of his strategy cannot be straightforwardly adapted to our context.[12] The reason is that most accounts of nomological connections rely on counterfactuals to spell out such connections, but it's generally accepted (and I endorse this view) that mathematical objects exist and have their properties out of metaphysical necessity. So they either satisfy or fail to satisfy counterfactuals trivially, which is why this type of reasoning does not reveal much about them. Nevertheless, Bromberger's account has provided us with some important clues as to the shape of the desired definition. What we need is to replace nomological necessity in the above account for some other kind of principle that, like the relevant nomological claims, *explain* why these properties are projectible. In the next two sections, I will argue that two important ingredients in answering these questions are equivalence relations and natural properties.

3.1.3 Types, Equivalence Relations, Abstraction

In the last section, I argued that the identification of the projectible properties should be tied to some kind of explanatory principle which gave us an account of why these properties, and no others, belong to a particular type. In the case of empirical properties, Bromberger's appeal to nomological necessity seemed to do the job, but when it came to mathematical properties, some other principle was called for. In order to uncover such a principle, I want

[12] Tappenden (2008a,b) gestures at such an adaptation, suggesting that there may be some unified account of causal or ontological dependence that applies to both physical and mathematical entities. He does not provide for one, however, and I myself am skeptical that such an account can be provided.

to take a closer look at the nature of mathematical types.[13] It is here that I believe the neo-Fregean tradition has some important insights.

As should be clear from our emphasis on projectible properties, one of the key features of a type is the fact that its tokens share certain properties *qua* tokens of the type (precisely the properties we are here calling projectible) and do not share certain other properties, which can thus be safely ignored. For instance, tokens of the type *Euclidean triangle* all share the property of having their internal angles sum two right angles, but they do not share their coordinates with respect to a given coordinate system. In particular, given that the properties we are here calling projectible are shared exactly by tokens of a given type, it is clear that the domain of objects is *partitioned* into those which share the properties and those which do not. To put it another way, there is an *equivalence relation* which holds precisely among objects of a given type.[14]

It should not be surprising that there is a strict connection between types and equivalence relations. Historically, equivalence relations were introduced precisely in order to abstract away from irrelevant features and identify distinct objects which are nevertheless indistinguishable from the point of view of some salient characteristic.[15] That is, equivalence relations allowed mathematicians to treat as *one object* distinct objects which belonged to the same equivalence class. Of course, some care must be taken here in order not to end in confusion. As Frege (1884/1960, §34) himself remarks, distinct objects are, well, distinct, and therefore we would incur in error if we treated them literally as identical; moreover, ignoring an object's properties does not make them go away, so something more has to be said about what it means to treat distinct objects as being somehow identical. As much as Frege deplores this imprecise way of thinking, however, I prefer to view it as groping towards something significant, namely that mathematicians use equivalence relations to reason about the *type* instead of the *tokens*. In order to clarify the situation, let's consider the matter a little more formally.

The basic idea is this: let $\sim$ be an equivalence relation on a given domain. Then it's possible to define a function f on the domain such that the following holds:

(AP): For any a, b in the domain, $f(a) = f(b)$ if, and only if, $a \sim b$.

Suppose for a moment that this best captures the mathematician's practice of treating distinct objects as somehow indistinguishable from the point of view of a salient property,

[13] Perhaps the argument here could be generalized to other kinds of types, but for simplicity I will restrict myself here to the mathematical context.

[14] That is because every equivalence relation gives rise to a partition and, conversely, every partition is associated with an equivalence relation.

[15] For a detailed historical account of the emergence of definitions by abstraction and their connection to equivalence relations, cf. Mancosu (2016, chaps. 1 and 2).

i.e. that in doing this they are actually reasoning about $f(a)$ instead of a.[16] The question then arises about how to interpret the range of this function. Historically, there were three options:

Option 1: The first option, especially prominent in number theory, is to consider the value $f(a)$ to be a *canonical representative* of the equivalence class of a. Ideally, one would want the function f to be explicitly given, so that one can obtain the canonical representative from any member of its equivalence class by some sort of algorithm (e.g. Kronecker's treatment of binary quadratic forms, as emphasized by Mancosu (2016, chap. 1)).

Option 2: The second option is to take the value of $f(a)$ to be simply the equivalence class of a under $\sim$, i.e. the *canonical or natural projection* of a set to its quotient set by the equivalence relation. Historically, this has been less favored than the first option, since one may lose computational information by taking this route; e.g. the function f defined in the first option gives an explicit procedure for computing the greatest common divisor of two numbers, whereas the canonical projection does not encode this information. Still, especially when dealing with more abstract, algebraic objects (e.g. quotients of a group by a normal subgroup), this is the only sensible option.

Option 3: The last option is to consider the range of f to be a set of *new* entities, abstracted from the given equivalence relation. Thus, to use Frege's famous example, one may define a relation $\sim$ between two lines a, b as holding if, an only if, they are parallel. From this relation, one then abstracts a new object, called the *direction* of a given line, whose identity conditions are given by $\mathsf{dir}(a) = \mathsf{dir}(b)$ if, and only if, $a \sim b$. More interestingly, although traditional set theoretical treatments generally treat the real numbers as, say, equivalence classes of Cauchy sequences or as Dedekind cuts, Dedekind himself took them to be new entities associated with such classes or cuts.[17]

Notice that there is a certain asymmetry between the first two options and the third one. Whereas the first two options arguably do not introduce any genuinely new objects

[16] Some historical support for this suggestion is again to be found in Mancosu (2016, chaps. 1 and 2).

[17] "Whenever, then, we have to do with a a cut (A_1, A_2) produced by no rational number, we *create* a new, an *irrational* number a, which we regard as completely defined by this cut (A_1, A_2)." (Dedekind 1872/1963, p. 15, first emphasis added). Although this may be taken to be mere rhetorical flourish, it's clear from other texts by Dedekind that this is not so: responding to a similar suggestion by Weber to simply identify the natural numbers with (what we call today) Frege-Russell cardinals, he flatly rejects it, insisting that they are new objects: "If one wishes to pursue your approach I should advise not to take the class itself (the system of mutually similar systems) as the number, but rather something new (corresponding to this class), something the mind creates." (Dedekind 1932, p. 489 apud Reck [2003, p. 385]) Thus Dedekind's approach is an instance of option 3, even though contemporary practice generally uses his techniques in the service of option 2.

(if we assume that the equivalence class can be eliminated in favor of its defining relation), the third establishes a new range of entities as the function's values. Given this, one may reasonably ask: when is one justified in introducing new abstract entities? Why not rest content with just canonical representatives and equivalence classes? In order to answer these questions, I will first make a quick excursus through the theory of concept settings and domain extensions as developed by Manders (1987, 1989). This will in turn give us a deeper insight into the nature of types.

3.1.4 Carving Nature at Its Joints

In two important papers, Manders (1987, 1989) considers a question that is in many respects similar to our own, namely the fruitfulness of *domain extensions*: in general, why is it fruitful to introduce new entities, extending the domain of objects? It is well-known, for instance, that the introduction of points at infinity not only greatly simplifies many geometrical theorems, but also gives deeper insight into the reasons why such theorems hold.[18] Likewise, the introduction of complex numbers also allows for deeper insight into why, e.g., certain series converge, by giving uniform conditions for convergence. In fact, this last example is illuminating of the general features of the situation, so we might as well develop it a little further in order to bring out such features.

Consider the way the complex numbers were introduced as a way of obtaining solutions to equations with coefficients in the real numbers. For instance, $x^2 + 1 = 0$ has no solutions in the real numbers, but it does have solutions in the complex numbers, namely $x = i$. Manders highlights three noteworthy features of this situation, which he thinks is typical: first, the old elements are preserved, that is, we have an *extension*, not merely a change of subject. Second, some properties of the original situation will be preserved, while others will be dropped. So, for instance, in this case, just as there was no $r \neq 0$ in the real numbers such that $r \cdot 1 = 0$, there will be no complex number $c \neq 0$ such that $c \cdot 1 = 0$ (i.e. the characteristic of the field will remain invariant). On the other hand, the real numbers are an ordered field, whereas the complex numbers are not. The preserved properties are called by Manders *invariant conditions*. Finally, given an invariant condition $\phi(x)$, one may also wish to preserve the universal closure of this condition, e.g. the associativity laws for addition and multiplication.

Next, he calls a *solvability condition* a formula $\exists y \phi(x, y)$, where x may be a parameter from the original domain. If there is any y which makes this formula true, then y is a solution to the formula. If every such a formula of a given type has a solution in a given structure, this structure is called *existentially closed*.[19] The domain extensions that interest Manders are

[18] For a particularly enlightening case study, cf. Lange (2015), which deals with Desargues's theorem.

[19] A more formal, textbook treatment may be found in Hodges (2004, chap. 8).

then the *existential closures* of an initial structure, that is, an extension which makes every solvability condition of a given type (for instance, every quadratic equation) satisfiable in the new domain, while at the same time satisfying the three above desiderata.

After this preliminary analysis, Manders proceeds to give some technical constraints on the logical features of the formulas expressing solvability conditions, constraints which will give a criterion as to when a given extension will be fruitful. The technical details of his discussion are unimportant here.[20] The upshot of his discussion is that two things may occur after the extension: either the complexity of the structure will greatly diminish, resulting in an overall simplification of the theory, or else it will increase to the point of making it very difficult to handle.[21]

What is interesting about the first case is that the simplification results in greater *conceptual unity*, by erasing certain complications of the original setting. To go back to the extension of the real numbers to the complex numbers, in the original setting the existence of solutions to the equation $ax^2 + bx + c = 0$ depended on whether $b^2 - 4ac \geq 0$ (in which case it has at least one solution) or $b^2 - 4ac < 0$ (in which case it has no solutions). In the new setting, however, order is not expressible anymore, so these case distinctions disappear and, in fact, every such equation has now exactly two solutions. Indeed, whereas in the original setting, the existence of solutions (and their number) for a given polynomial equation had to be treated on a case by case basis, now they all follow from the Fundamental Theorem of Algebra, which asserts that every polynomial equation with complex coefficients and degree n has exactly n solutions in the complex field. In other words, the domain extension unifies all the diverse cases of the original setting into one fundamental theorem that encompasses all such cases, resulting in greater unity.

This last feature is the key to the whole procedure. It shows that there are certain domain extensions which shed more light on certain phenomena, or bring conceptual unity to disparate areas, such as the extension of the real numbers to the complex numbers. That is, such extensions erase distinctions that are irrelevant and classify together cases that naturally belong together, sharing important salient features. Dramatically speaking, we can say, to employ the old Platonic metaphor, that those extensions which posses these virtues *carve nature at its joints*, that is, are in some sense *natural*. Of course, the criteria for naturalness may vary from situation to situation: in the case of domain extensions which are existentially closed, Manders's criteria may apply. But there is no reason to expect that the same criteria will work in every situation (nor, for the matter, does Manders suggest this—quite the

[20] They basically amount to a set of criteria that will ensure that, after the extension, the resulting structure admits quantifier elimination, among other things.

[21] In the second case, Manders remarks that the solvability condition will express a condition that was previously expressible by an infinitary condition.

contrary).[22]

But why should the naturalness of a certain object be a criterion for its ontological respectability? Here, Mark Lange's (2017, p. 334) is revealing. Lange's argument, recast using the language of naturalness, is:

P1: Natural objects or properties explain certain facts about the objects in the original domain (e.g. the complex number field's structure explains certain facts about the real number field);

P2: What explains a fact about some entities must be on an ontological par with those entities (as he puts it, "only facts about what exists can explain facts about what exists");

C: Natural objects or properties are ontologically on a par with objects on the original domain.

In other words, it is *in virtue of*[23] the structure of the new objects that certain properties hold of the original objects. Hence, if we are willing to accept the objects in the original domain, we should also be willing to accept the objects in the extension.

Let's take stock. I started this chapter with a discussion of what I called the Kantian predicament, namely how to identify which properties are projectible across a given type. I then argued that the answer to this problem should come with an account as to what makes such properties projectible, taking my cue from Bromberger's analysis of linguistic types. In the last section, I proposed that an important role in this explanation was played by equivalence relations, since typically types are abstracted from such equivalence relations. Since, however, not every equivalence relation gives rise to an associated abstract object, the question arose as to which equivalence relations do so. My proposal in this section, then, has been to consider *natural* equivalence relations, in the sense of relations which somehow bring more conceptual unity to a given subject matter. In the next section, I will argue that Klein's central insight in the Erlangen Program allows us to articulate precisely why (for instance) geometrical types are natural, thus finally answering the Kantian predicament.

[22]Importantly, however, there is a sense in which, on my view, such criteria are *intrinsic* to the theory or structure being analyzed. My proposal is thus distinct from Lange's (2015; 2017), according to which the criteria for naturalness are tied to the way a given property figures in mathematical explanations. Although I agree with Lange that this is generally a useful heuristic for establishing *that* a property is natural, I don't think it is sufficient to establish *why* it is natural.

[23]This is suggestive of the language of *grounding*, at least as developed, e.g., by Fine (2012).

3.1.5 Klein's Insight

I will not give a full account of Klein's celebrated paper here, but, instead, I will consider only points that are directly relevant to our discussion.[24] As is well-known, one of the main innovations of Klein's paper was his use of the group concept as tool to classify and unify the disparate geometries of his time. My main interest here is in how this use of groups can help us to see types as natural objects.

It's important to remember that, at the time of Klein's address, not only were there many different styles (e.g. synthetic and analytic), but also many apparent disparate types of geometry, each with its own distinct character, so to speak. Especially troublesome was the status of projective geometry, which had come to dominate the scene by then; as Mark Wilson notes, "the world of the so-called 'projective geometer' is considerably more bizarre than non-Euclidean geometry per se" (Wilson 1995, p. 113).[25] Part of the problem was the seeming incompatibility of projective geometry with a metric, as Klein himself highlights in the opening paragraph of his essay.[26] One of the main goals of his paper, then, was to organize these different strands into a coherent whole.

Famously, Klein's ingenious solution was to use the group of transformations behind the geometry as the main tool of his classification. Specifically, given a space and a set of primitive notions, Klein looked to the group of transformations that preserved these notions. So, for instance, given the Euclidean plane and the primitive notions of angle and length, we obtain the group of transformations that preserves these notions, namely the group of isometric transformations. On the other hand, if we're given the projective plane and the primitive notion of the cross-ratio, we obtain the group of projective transformations. I will call this central idea *Klein's insight*:

Klein's Insight: In order to identify the natural notions of a given geometry, one should look at the group of transformations behind the geometry. Or, as he himself puts it, "*geometric properties are characterized by their remaining invariant under the transformations of their*

[24] I am here greatly indebted to the discussion byYaglom (1988, chap. 7) and especially Marquis (2009, chap. 1). Indeed, the initial spark for reading Klein as providing an answer for the Kantian predicament was occasioned by my reading of Marquis. For further discussion of Klein's Erlangen Program, cf., among others, Wussing (2007, Part III), Gray (1992), Rowe (1992), Hawkins (1984), and Birkhoff and Bennett (1988). Hawkins (2000), although more focused on Lie, provides a rich analysis of Klein's ideas and background in the first chapter. Curiously, as can be gathered from Hawkins (1984), it seems that Klein's Erlangen Program itself was not very influential, even though retrospectively many mathematicians (e.g. Cartan) would find in it key ideas that oriented their work.

[25] The first sections of Wilson's article furnish a nice picture of the development of geometry in the first half of the 19th century, with special emphasis on the work of Poncelet and von Staudt.

[26] This may be especially important in connection with Riemann's views. According to Gray, for Riemann, "all geometry is based on specific metrical considerations" (Gray 2011, p. 201).

principal group" (Klein 1892–1893, p. 218), where the "principal group" is simply the group of transformations associated with a given geometry.

This may not sound very enlightening, especially since it still relies on the some undefined way of identifying those primitive notions.[27] So Klein actually turns the idea sketched on the above paragraph on its head: instead of identifying the group of the transformations by the properties it preserves, he had the idea of identifying the primitive notions as being the properties preserved by the group of transformations. That is, one would not say that the isometric transformations are the transformations which preserve the Euclidean notions, but rather that the Euclidean notions are the notions preserved by the isometric group, thus reversing the direction of explanation.

This move looks like a mere trick or sleight of hand, but it has profound consequences. Klein's shift to groups acting on a space emphasizes that the *orbits* of an element of space under a group action is precisely an equivalence class, which can then be identified with the *type* of that element. The idea is simple: from the point of view of Euclidean geometry, two line segments are indistinguishable as long as they have the same length; they can therefore be considered as tokens of the same object type. But the isometries of the space will precisely carry a line segment of a given length to another one of the same length, and will not carry a line segment of a given length to another one of different length. That is, the orbit of a line segment under the group of isometries will be precisely the congruent line segments. Therefore, one can treat the equivalence class of a given line segment generated by the action of the group of transformations on the space as being precisely the *type* of the line segment. If necessary, this process can be iterated: if lines are sets of points, then a figure can be considered as a set of lines. Again, the type of figure will be the equivalence class of certain lines by the induced group action along the type-hierarchy.

The notions of metric Euclidean geometry (lines, triangles, etc.) are therefore precisely the equivalence classes induced by the group action of the isometric transformations.[28] This is a completely general criterion: given a space S and a group of transformation G, the notions corresponding to the geometry (S, G) are precisely the objects of the corresponding type-hierarchy $U(S)$[29] left invariant under the action of G on $U(S)$. in other words, the notions will correspond rather precisely to the object types obtained by abstracting from the equivalence classes generated by the group action.

[27] Not surprisingly, according to Gray (2011, p. 236), many mathematicians, including Cayley, thought that Klein's argument was viciously circular.

[28] The idea of using the group structure of the isometric transformations as a criterion for object identity comes from Yaglom (1988); the helpful use of the type-token distinction in this context comes from Marquis (2009). Also relevant in this connection is Wilson (1995, 2005), whose papers, although focused on Frege, deal with abstraction principles in the context of projective geometry.

[29] This is the type-hierarchy having S as its domain.

As should be clear, this key idea will also allows us to define rather precisely which properties are projectible across the type: the projectible properties are exactly the properties left invariant by the group of transformations associated with the geometry. So, for instance, in the previous examples, *being of a certain length* is invariant under isometric transformations, whence it is a property projectible across the type of a given line segment. Similarly, given that length is invariant, *being equilateral* is also invariant, so this is a property projectible across the triangle type. Hence, focusing on the group of transformations allows us to formulate a precise criterion for when a given object or property is natural: an object a or property P is natural if, and only if, it is invariant under the relevant group of transformations.

This criterion of naturalness agrees with the considerations put forward in the last section. As I mentioned in the beginning of this subsection, Klein developed his framework, among other reasons, to *conceptually clarify* the notions involved in the geometry of his time. The first step was to find a common background which could reveal what all these different subjects (Euclidean, affine, projective geometry) had in common. These he found in the group concept: these different areas are all concerned with the study of properties left invariant by transformations of the space. This allowed him to take the second step, namely to show how this *conceptual unity* revealed the precise relations each geometry entertained with each other. It is this step that I want to analyze now.

The idea that geometry is essentially the study of invariants under certain transformations allowed Klein to organize and classify the different geometries in a neat hierarchy, using two basic principles: (i) it's possible to identify what are apparently different geometries by identifying their respective transformation groups. For instance, the real projective line and the complex upper half-plane may look like completely different geometrical objects, yet they can give rise to isomorphic transformation groups (the transformations which preserve the cross-ratio in the case of the real projective line, the Möbius transformations in the case of the complex upper half-plane), so they are identified as being essentially the same. This allows one to work with one object and then transfer this work via a canonical map from one object to the other, which greatly facilitates some proofs.[30] Equally importantly, (ii) if a geometry's transformation group is a subgroup of another's, then the first geometry can be considered a sub-geometry of the second. This means that every theorem valid in the second setting is also valid in the first setting. Moreover, it allows one to build a nice hierarchy of geometries, which fulfills the classification goal.[31] This hierarchy also makes possible to classify one geometry as *more general* than another, an idea that will be

[30] It also makes possible to endow the real projective line with a non-Euclidean metric, as Klein himself realized.

[31] A picture of which can be found in Kline (1972, p. 919). Although Kline includes affine geometry in the picture, which was not known to Klein at the time, it fits so nicely into the scheme that this anachronism is justified. Another, different picture can be found in Marquis (2009, p. 32).

important for Tarski later on: a geometry A is said to be more general than geometry B iff the group of transformations of B is a subgroup of the group of transformations of A. So, for instance, affine geometries are less general than projective geometries, since the group of affine transformations is a subgroup of the projective transformations.

Using these ideas, we can finally solve the Kantian predicament. To recapitulate, we started with the problem of identifying the properties which are projectible across a given type. Bromberger's analysis gave us a clue, namely to look for explanatory principles that not only identified those properties, but at the same time explained why they are projectible. Neo-Fregean abstractionism provided a further ingredient: types are associated with equivalence relations, so if we could somehow sort the "good" equivalence relations from the "bad" ones, that could go some way towards solving our problem. A way of sorting out the "good" equivalence relations came from Manders's work on domain extensions: the "good" equivalence relations were those that cut nature at its joints, that are somehow natural. Now, Klein's insight gave us a criterion for when an equivalence relation is natural in the context of (homogeneous)[32] spaces. An equivalence relation is natural in the context of homogeneous spaces if, and only if, it is invariant under the transformation group associated with the space. So a *type* in this context will be an object *abstracted* from these equivalence relations, and the properties which are projectible across a type will be exactly those invariant under the group of transformations. In the next section, I'll show how these ideas can be put to use when considering Tarski's logicality proposal.

3.2 Tarski's Extension of Klein's Erlangen Program

In the last section, I argued that Klein's insight allowed us to identify the geometrical notions in a very precise way. In particular, I argued that a notion, for Klein, should be (anachronistically!) understood as an object in a type-theoretical hierarchy, and that the type of notion (Euclidean, affine, projective) was to be specified by considering the transformation group associated with the space. In this section, I want to show how Tarski incorporates and extends this viewpoint in order to answer his question, "What are logical notions?". Again, I'll tackle this in two separate sections, first explaining Tarski's strategy and then what he understood by "notions".

3.2.1 Logical notions

One point that is important to note about Klein's project is that it only works for homogeneous spaces, that is, spaces in which no point is distinguishable from any other. While some may consider this as a defect, because it implies that his project only works for spaces of

[32] I will get to this qualification in the next section.

constant curvature,[33] it can also be seen as an asset, since it characterizes *precisely* such spaces. If that is so, then the first step towards Tarski's proposal is the idea that, from a logical point of view, the world itself is a homogeneous space, that is, no individual is distinguishable by logical means. This allows Tarski to apply Klein's methods to the situation at hand.

Tarski's idea is rather simple. Recall from the last section that Klein used the fact that the subgroups of a given group form a lattice in order to establish a unified, hierarchical picture of the different geometries, with projective geometry at the top. He also tentatively mentioned higher groups in the hierarchy, such as the group of all homeomorphisms (continuous transformations), but didn't develop this idea in his paper. Of course, extending this idea further, the result is the group of all transformations from the space onto itself. This is precisely the group of all permutations of the space. This give us *Tarski's proposal*:

Tarski's proposal: *The logical notions are exactly the notions invariant under all permutations of the world onto itself.*

We can reconstruct Tarski's reasoning as follows. First, notice that the relation "is a subgroup of" is a partial order in the set of all subgroups of a given group G, so we can say that a group H is smaller than H' iff H is a subgroup of H' (of course, there will be incomparable groups according to this relation). Moreover, this partial order has a maximum element, namely G itself. Given this, Tarski's reasoning can be roughly summarized like this:[34]

(P1): The logical notions are the most general notions.

(P2): The most general notions are those invariant under the largest transformation group.

(P3): The largest transformation group is the full permutation group.

(C): The logical notions are those invariant under the full permutation group.

The argument above is clearly valid. Is it sound? Tarski apparently considers (P1) as given, since he doesn't argue for it. (P3) is a mathematical fact: the largest *group* of mappings from a set to itself is the full symmetric group of the set.[35] So the whole argument seems to rest on (P2). Let's take a closer look at this premise.

[33] For some comments on the situation, including a discussion of the Helmholtz-Lie theorem, cf. Stein (1977, p.36n29) and Friedman (2002, pp. 196ff).

[34] Bonnay (2008, p. 5) contains a similar reconstruction. It will be clear in the sequence that my reading is somewhat different from Bonnay's, however.

[35] There are groups of mappings over a set which are not subgroups of the symmetric group of the set. However, if a group of mappings over a set contains one injective mapping, then it is a subgroup of the symmetric group of the set.

Tarski seems to have in mind the following line of argument for supporting (P2). Let's say that a notion *is relevant for* a given science (or the science *is about* that notion, in a rather weak sense of aboutness) if, and only if, the associated group of transformations distinguishes the notion (i.e. the notion is invariant under the group action). It's now possible to define a partial order among notions in the following way: a notion N is *more general* than another notion N' if, and only if, N is relevant for more sciences than N'. Hence, the most general notions are the notions invariant under the largest group of transformations, and (P2) is vindicated—if one accepts this definition of "more general", that is, and always relative to a give space, a point that I will come back to below.[36]

Note how the concept of *group* is essential to the above line of thought. It's essential for defining "notion" in an appropriate way, as we have seen in the last section, a definition on which the subsequent arguments depend. In this regard, there's a historical curiosity that should be mentioned. The idea of using Klein's classification as a criterion for logicality apparently comes from the Polish mathematician Alexander Wundheiler, himself inspired by a previous paper by Tarski and Lindenbaum (1935/1983). If we are to trust Carnap's journal, Wundheiler mentioned this idea to Carnap, Tarski, and Quine in a conversation of the "logic group" at Harvard on January 10th, 1941. Here's Frost-Arnold's translation of the relevant transcript by Carnap, which contains a surprising remark by Tarski:

> *Wundheiler*: Can we perhaps characterize the difference between logic, mathematics, and physics through transformation groups, just as we characterize projective, affine, and metrical geometry through transformation groups?
>
> *Tarski*: It is doubtful whether the concept of *group* helps much in this context. (Frost-Arnold 2013, pp. 152-3, original emphasis).

The above exchange is definitely surprising, given how essential the group concept has proven to be in foregoing discussion. Maybe Tarski only had in mind that the group concept is not of much help in *distinguishing logic from mathematics*. Tarski was consistently skeptical of this distinction;[37] even in Tarski (1966/1986), Tarski closes the text by noting that his proposal does *not* imply an absolute distinction between logic and mathematics. So this may explain Tarski's skepticism as registered by Carnap. I will comment further on this point in the next section, so let me now concentrate on another point, namely whether Tarski's proposal covers only *relative* logicality or if it can also be used to define *absolute* logicality.

[36]This conception of relevance is too coarse, of course. Notice that it's not necessary to take that as a definition of "more general"; the right-to-left implication would suffice.

[37]Cf., among others, the closing paragraphs of Tarski (1983), as well as Tarski (1944). Also, in the conversations registered by Carnap and presented by Frost-Arnold (2013), there are scattered many skeptical remarks by Tarski about this distinction.

To understand this issue better, note that a group of transformations is quite obviously the group of transformations *of some space*, or, alternatively, that the invariant objects of the corresponding type-hierarchy are all defined relative to some initial domain. The natural conclusion here is that *being logical* is always thus domain relative, and then one may ask for an *absolute* conception of logicality, one that would be domain independent.[38] There are two ways of resisting this conclusion, however. One way would be to treat *being logical* in a way that resembled Kaplan's treatment of indexical expressions, such as "I". According to this view, *being logical* would always have the same *content*, namely *being invariant under all transformations of the domain*, but the specification of its extension would depend on a further parameter, that is, *which* domain we are taking into account. On this reading, *being logical* would be absolute in the sense that the content of the concept would not be relative, but *which notions are logical* would be relative. This would accord with some readings of his definition of truth for formalized languages, according to which Tarski is authorized in claiming that he defined *the* concept of truth precisely because he was able to capture the content of this concept, even though its extension shifted from language to language.[39]

Nonetheless, this seems unsatisfactory, since then which notions are logical is still relative, so we may inquire whether it's possible to read the proposal in such a way to avoid this problem. And here, I believe that some of the comments I made on Chapter 2 regarding Tarski's metaphysical views are relevant. In particular, in that chapter I stressed how Tarski always favored a single domain approach, one that took as its domain *the world itself*. Of course, as pointed out in that chapter, Tarski unfortunately coupled this idea with a dubious attempt at defending nominalism by way of a paraphrasing strategy. In contrast, here I want to suggest that the best way to advance his proposal may be to embrace platonism and develop one's ontology accordingly. Indeed, we can directly apply the ideas developed in this chapter to Tarski's proposal. Here is a rough sketch of such an application.

We start with our world, a (logically)[40] homogeneous space populated by concrete individuals and abstract objects. Among these, a certain class is rather special, namely the logical ones. To identify these, we turn to the equivalence relations that ground them in some way. Since the world is homogeneous, Klein's ideas find a natural extension in Tarski's proposal, which identifies the relevant equivalence relations with those that are invariant under the action of the group of all transformations of the world onto itself. The objects associated with those equivalence relations would be "absolutely logical", that is, they would be *the* logical objects, those that inhabit our world.

There are two ways that I think can be developed further, though the differences be-

[38] Some of Feferman's criticisms may be interpreted in this direction, though I believe he has something more specific in mind. I discuss Feferman's position more fully in the next chapter.

[39] For this interpretation of Tarski's definition of truth, cf. Patterson (2012, pp. 60-1).

[40] That is, from a logical point of view. From a physical point of view, the world is not homogeneous!

tween them may turn out to be merely cosmetic. One is to consider as given only the domain of individuals and build a hierarchy of abstract objects using only a certain class of permissible abstraction principles.[41] This would generate a hierarchy similar to the one outlined by Hale (1987, Appendix 1 to Chap. 3). Another approach would be to start with the whole type-theoretical hierarchy and then to expand the initial domain by introducing new objects by way of abstraction principles, in a way reminiscent to the domain extensions analyzed by Manders. The first way seems more promising to me, but regardless, the overall idea is that the logical objects would be those introduced by abstraction principles whose equivalence relations would be precisely those invariant under the action of the full group of transformations of the domain. In other words, instead of taking Tarski's proposal to "select" among the objects in a type-theoretical hierarchy which ones are logical, we would use the corresponding abstraction principles to introduce new, *sui generis* objects, namely the logical objects.

3.3 Consequences of the proposal

In the last section, I analyzed in some detail Tarski's proposal and the metaphysical picture which I find most congenial to it. In this section, I want to discuss a couple of consequences of his proposal. There are three topics I want to highlight: (i) Tarski's remarks on cardinality properties; (ii) the consequences of the proposal for the distinction between logic and mathematics; (iii) the light the proposal may shed on which are the logical constants.

3.3.1 Cardinality properties

These consequences are interesting froma historical point of view, given that Tarski's observations here point to a a surprising continuity in his thought about these matters. First, note that it's not difficult to see that, if we construe the cardinality of a given class as the class of all classes equinumerous to it, this property will come out as logical under Tarski's proposal, as he himself remarks in his lecture. From this, Tarski draws a striking conclusion:

> This result seems to me rather interesting because in the nineteenth century there were discussions about whether our logic is the logic of extensions or the logic of intensions. It was said many times, especially by mathematical logicians, that our logic is really a logic of extension. This means that two notions cannot be logically distinguished if they have the same extension, even if their intensions

[41] The qualification is necessary because some abstraction principles give rise to contradictions, as Frege's infamous Basic Law V. Actually the problem is worse, since there may be consistent abstraction principles which are nonetheless inconsistent together. For a detailed exploration of this type of problem, cf. Fine (2002).

> are different. As it is usually put, we cannot logically distinguish properties from classes. Now in the light of our suggestion it turns out that our logic is even less than a logic of extension, it is a logic of number, of numerical relations. We cannot logically distinguish two classes from each other if each of them has exactly two individuals, because if you have two classes, each of which consists of two individuals, you can always find a transformation of the universe under which one of these classes is transformed into the other. Every logical property which belongs to one class of two individuals belongs to every class containing exactly two individuals. (Tarski 1966/1986, p. 151)

Interestingly, *the same conclusion was already drawn in his 1935 article with Lindenbaum, in almost the same words*:

> It is customary to say that our logic is a logic of extensions and not of intensions, since two concepts with different intensions but identical extensions are logically indistinguishable. In the light of Th. 5 [which asserts that no two classes of individuals of the same cardinality are logically distinguishable—D. N.] this assertion can be sharpened: our logic is not even a logic of extensions, but merely a logic of cardinality, since two concepts with different extensions are still logically indistinguishable if only the cardinal number of their extensions and the cardinal number of the extensions of the complementary concepts are also equal. (Tarski and Lindenbaum 1935/1983, p. 388)

Incidentally, this shows that already in 1935 Tarski was willing to draw consequences about the nature of logic notions from their permutation invariance. I mention this because Patterson (2012, p. 212) argues that there is no continuity here between the two papers, since "the very fact" that Tarski announces as a theorem what he later took to be a definition "gives the game away". This is bizarre. Suppose that I were able to show that pairs (Γ, ϕ), where Γ is a set of sentences, and ϕ is a sentence, satisfy a certain property if, and only if, ϕ is a logical consequence of Γ. This is a theorem, which I take to be revealing about the nature of logical consequence. Later, I am so impressed by the naturalness of this property that I decided to make it into a definition of logical consequence, using the theorem to buttress my claim that the property should enjoy definitional status. Although there has obviously been *some* change in my attitude towards my theorem, it seems certain that there is a notable continuity between my attitude towards the property I have discovered, and that this continuity is much more important than the noted discontinuity. In this respect, it seems that the 1966 lecture is, like its companion article, "Truth and Proof" (Tarski 1969), an update of old papers, incorporating old views into the new perspective Tarski attained after his break with intuitionistic formalism.

In any case, the suggestion that "our logic is even less a logic of extension", but it is instead "a logic of number, of numerical relations" may seem to imply that Tarski is here assimilating logic to mathematics, which brings us to our next topic.

3.3.2 Mathematics as logic?

We saw in the last section how Wundheiler introduced the idea of using transformation groups to identify the logical notions, and thus establish a line between logic and mathematics. In that context, we also saw how Tarski surprisingly expressed skepticism regarding Wundheiler's proposal. So, has Tarski changed his mind in the intervening years? The answer seems to be negative, that is, he is still skeptical of a principled demarcation between logic and mathematics. That does not mean that he assimilates one to the other; rather, he thinks the distinction is *relative*.

A comparison with Carnap may again prove to be instructive.[42] As emphasized in the last chapter, Carnap had a rather radical pragmatist outlook, in which philosophical frameworks should be evaluated in terms of their scientific fruitfulness. There, we saw in particular how this pragmatism guided his discussion of explication, by providing the appropriate benchmark against which to judge a purported explication. Here, I want to emphasize a different aspect of this pragmatism, namely his famous *Principle of Tolerance*:

> *In logic, there are no morals.* Everyone is at liberty to build up his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments. (Carnap 1937/2001, p. 52, original emphasis)

Carnap's tolerance is tied with his explication project. Recall from the last chapter that sometimes different explications of a same concept are possible: thus, the concept *fish* can be explicated both in a more traditional Linnean taxonomy (which would result in sharks being classified as fish) or in a cladistic one (excluding shark from the fish), the choice between one or the other being an *external* question, that is, one to be decided on pragmatic grounds. In some cases, however, there could be doubts even about the logical framework in which we couch our explications; the principle of tolerance provides an answer to this worry. Instead of fixing one logical framework once and for all, Carnap insists that the adoption of a framework must be based on pragmatical considerations. Thus, for instance, the choice between intuitionism, predicativism, or simple type theory is deflated as a pragmatic choice,

[42] Again, a full analysis of Carnap's position is not intended here. For a detailed analysis, cf. Carus (2007, chap. 10), as well as many of the essays in Wagner (2009), in particular the ones by Richard Creath, Thomas Ricketts, and Michael Friedman. Creath's essay is particularly congenial to the viewpoint adopted here.

so that one can change logics based on the purpose at hand. This shows the remarkable extent to which Carnap became a pragmatist.[43]

This line of thought seems to be congenial to the late Tarski. In the previous chapter, I argued that Tarski moved from a traditional position associated with Leśniewski to a pragmatic position, more in the spirit of the Lvov-Warsaw school of mathematics. For instance, in a contribution to a Chicago meeting of the Association for Symbolic Logic and the American Philosophical Society in 1965, Tarski made the following remarks:

> (...) maybe the notion of truth is simply not proper, this classical notion of truth, for mathematical sentences. I think that this is the negative conclusion which one could draw from these developments starting with Gödel, and therefore I am quite interested in attempts of [at?] constructing set theory on the basis of some non-classical logics, simply as an experiment. We shall see to what it will lead. (Tarski 2007, p. 261)

So Tarski is at least not opposed to considering alternative logics as "an experiment"; presumably the last line indicates that such a logic could end up being preferable to classical logic, if it proved more fruitful, indicating his pragmatic attitude towards such matters. Such an attitude is in evidence also in his abandonment of type theory in favor of an untyped language for set theory, e.g. first-order ZFC. Here's how Carnap recalls his conversation from February 13th, 1941, with Tarski on this matter:

> The *Warsaw Logicians*, especially *Leśniewski and Kotarbiński*, considered a *system like PM* [[*Principia Mathematica*]] (but with a simple theory of types) completely self-evident as a formal system. This limitation worked strongly and suggestively on all the students, and on T. himself until "*Wahrheitsbegriff*" (where neither transfinite types nor a system without types is considered, and finitude of types is implicitly presupposed; they were first articulated in the appendix, added later). But then Tarski saw that an entirely different system-form is used in *set theory* with great success. So he finally came to consider this system-form without types as more natural and simpler.(Frost-Arnold 2013, p. 160, original emphasis, insertion in the double brackets by Frost-Arnold)

[43] In this connection, I cannot agree with Coffa (1991, pp. 320ff) that Carnap adopted a "second-level semantic factualism", i.e. that he thought his principle of tolerance was somehow "true". As remarked by Stein (1992), Carnap himself pointed out that the divergences between him and Quine could (only?) be solved on pragmatic grounds. Since these include his principle of tolerance, it seems that Carnap was thoroughly pragmatist about philosophy: the dispute between him and someone who denied tolerance is not an internal dispute, but an external one, also to be decided based on the fruitfulness of the respective projects. In other words, I think Coffa underestimates both the extent and the cogency of Carnap's pragmatism.

This passage is especially revealing given that, before (e.g. in a conversation with Carnap from October of the previous year), Tarski had remarked (i) that higher-order logic was implicitly committed to platonism and (ii) that he rejected platonism, so, if metaphysical considerations of this sort were driving him, this would be the perfect place to remark on it. Instead, Tarski gives as his main reason for adopting a system like first-order ZFC its "great success", once again emphasizing pragmatic considerations.[44]

Therefore, if the adoption of a logical framework is itself subjected to pragmatic considerations, it's no surprise that the question of whether we should distinguish logic from mathematics is also relative to this type of approach. So we find Tarski telling Morton White in a letter from 1944:

> (...) sometimes it seems to me convenient to include mathematical terms, like the $\in$-relation, in the class of logical ones, and sometimes I prefer to restrict myself to terms of 'elementary logic'. Is any problem involved here? (Tarski 1944, p. 29)

Since Tarski considered that all of mathematics could be reduced to the membership relation, the above quotation is basically saying that whether or not mathematics is reducible to logic is a matter of *convenience*. This is exactly the view we find also expressed in the 1966 lecture. After raising the question of the reducibility of mathematical notions to logical notions under his proposal, Tarski remarks:

> Are set-theoretical notions logical notions or not? Again, since it is known that all usual set-theoretical notions can be defined in terms of one, the notion of belonging, or the membership relation, the final form of our question is whether the membership relation is a logical one in the sense of my suggestion. The answer will seem disappointing. For we can develop set theory, the theory of the membership relation, in such a way that the answer to this question is affirmative, or we can proceed in such a way that the answer is negative. (Tarski 1966/1986, p. 152)

Or, as he himself glosses the above in the very next line, "So the answer is: 'As you wish'!". Of course, this jocose remark should not be taken too literally; as per the letter to Morton White, it's not like Tarski thinks the distinction between logic and mathematics is completely arbitrary or up to a person's whim. Rather, he thinks such a distinction is a matter of convenience or technical expediency.

[44]Indeed, Tarski himself would adopt a type-theoretical framework if that suited his purposed. Cf., for example, the discussion of logicality in Tarski and Givant (1988) and the comments on that work by Bellotti (2003).

There are nevertheless some problems with this answer. The first is that there is a certain incongruity between it and the Kleinian motivation behind Tarski's proposal. For the motivation to that proposal was precisely the realization that certain geometrical notions are less general than the logical notions. If, now, geometrical and logical notions are on a par, this seems to undercut that line of reasoning.[45] Of course, this only happens because we seem to be operating with two different sets of objects: the ones that correspond roughly to the orbits of the action of the geometric groups and the ones corresponding to their, let us say, set-theoretical surrogates constructed inside the type-theoretical hierarchy. Seem in this light, the problem dissolves itself: the geometrical objects generated by the geometrical groups are not logical, whereas their surrogates are. This is not surprising: the type-theoretical hierarchy was developed precisely to show that mathematics could be modeled inside it. This does not mean that mathematics is logic, anymore than the fact that we can model the real numbers in the cumulative hierarchy means that analysis is really ZFC in disguise.

3.3.3 Logical constants

As I argued in the beginning of this chapter, Tarski's proposal is essentially a proposal about *notions*, taken as objects, properties of or relations between objects, and not about logical constants, that is, linguistic items. Nevertheless, there may be some way of translating claims about logical to logical constants, some *bridge principles* that allows us to cross over the two domains. I want here to examine Tarski's proposal in light of some candidate bridge principles. The results will be largely negative, though I do try to point in the direction of further work in this area.

The first bridge principle was proposed by Tarski himself, together with Givant. Fix a domain $\mathcal{D}$ and suppose S is a symbol from a type-theoretical language. Tarski and Givant (1988, p. 57) provide the following definition for logical constants:

(BP1): *A symbol S from a type-theoretical language is said to be a logical constant iff for every interpretation $S^{\mathcal{D}}$ of S with domain $\mathcal{D}$, $S^{\mathcal{D}}$ is a logical notion or operation.*

Of course, given that the principle makes reference to *every interpretation*, it implicitly assumes a list of basic logical constants (otherwise no constant would come out as logical), so this defines at best relative logicality. One way to try to fix this is to work with interpreted

[45] This is related to an objection by Bonnay (2008, pp.8-10), according to which Tarski's proposal is faulty since any mathematical structure can be made into an invariant object by employing certain tricks. I don't find the line pursued by Bonnay convincing, however, since his point is that this undermines our "intuitive" picture according to which logic is more basic than mathematics; I personally don't find this "intuitive" at all, and I certainly don't see the relevance of such brute "intuitions" to the discussion.

languages. So supposing a type-theoretical language and its corresponding interpretation as given, one could propose:

(BP2): *An expression S of an interpreted type-theoretical language is a logical constant iff it denotes a logical notion.*

This still has problems, though. To see this, consider an arbitrary predicate P. Then, according to (BP2), the expression $Px \vee \neg Px$ will come out as logical, since it will denote the universal set. As intuitively non-logical predicates such as "is blue or isn't blue" count as logical according to this principle, it might be better to search for a better one. Here's a tentative suggestion.[46]

BP3: *An expression S of a given language L is a logical constant iff it* analytically *denotes a logical notion or operation.*

Evidently this depends on whether or not we can define "analytically" in a satisfactory way. It seems to me that this is possible, but, again, I will content myself here with only a very rough sketch. The idea is to employ the general semantics framework developed by Lewis (1970/1983) for this purpose.[47] Roughly put, we use a categorial grammar to sort our expressions into certain basic types and then show how certain semantic values are attributed to the expressions of the language based on their "construction tree". These values are basic functions which take as values items from a certain category, plus certain parameters, and give another semantic value as output. An expression then analytically denotes a value if, for every relevant parameter, it denotes that value.[48] If this rough sketch could be fleshed out, then we would have a working definition of "analytic" that we could use to work out our bridge principle. Without such a definition, I personally don't see how such a principle can be sensibly constructed. This is a task for a further work, however.[49]

3.4 Conclusion

In the first part of this chapter, I analyzed Klein's proposal and how it could be used as an answer to the problem of identifying projectible properties, a problem which I called the

[46] This is similar to the suggestion made by McGee (1996) at the end of his paper.

[47] A similar idea is developed by MacFarlane (2000, Chap. 6), though I would personally pursue another direction, one more directly tied with what Lewis calls "meaning"—note that MacFarlane (2000, p. 189) explicitly says he will not "venture into the theory of meaning".

[48] Compare with the discussion about meaning and being analytically true in Lewis (1970/1983, pp. 200-3).

[49] In a recent talk at our department, my colleague André Quirino remarked that a possible definition for "analytic" would be that it is *essential* for the word to have that particular semantic value. "Essential" here would be cashed out in terms of necessary qualities and supervenience. Again, however, these ideas are too inchoate for me to evaluate whether they can be made to work in the present context.

Kantian predicament. In summary, Klein's idea was to employ the concept of invariance to simultaneously define an object-type and identify which properties are projectible across the type. I then showed how this framework was extended by Tarski in his proposal for the logical notions, with some observations on how certain "metaphysical adjustments" could be made so that the proposal would be more attractive. Finally, I analyzed three specific consequences from the proposal, showing how some of the criticism leveled against it could be met.

A Appendix: Group Actions and Homogeneous Spaces

Given the central importance of group actions and homogeneous spaces in this chapter, I decided to include here in the Appendix the basic results and definitions related to these notions. They are, of course, not an introduction to the subject, but merely pointers for the curious reader who may be puzzled by my use of these notions in the text. The exposition is hence very informal.

Let's start with the definition of a group:

Definition A.1. A *group* is a set G together with a binary operation $\circ$, a unary function $^{-1}$, and a distinguished element e satisfying the following three axioms:

1. $\forall x \forall y \forall z (x \circ (y \circ z) = (x \circ y) \circ z)$ (Associativity);
2. $\forall x (x \circ e = x)$ (e is the identity);
3. $\forall x (x \circ x^{-1} = e)$ (x^{-1} is the inverse of x).

Example: If X is a set, let $\mathrm{Sym}(X)$ be the set of all bijections from X to itself. Then $\mathrm{Sym}(X)$, together with function composition $\circ$, the unary operation $^{-1}$ which takes every function to its inverse, and e as the identity function, is a group.

The reader who is wondering where this "load of easily-forgettable axioms" (Arnol'd 1998, p. 234) comes from may consult Hans Wussing (2007) interesting book on their origins; and the diligent reader will easily verify that any group satisfies the following propositions:

Proposition A.1. *If G is a group, then G satisfies both the left and the right cancellation laws: for any $a, b, c \in G$, $a \circ b = a \circ c$ and $b \circ a = c \circ a$ each implies that $b = c$.*

Proposition A.2. *If G is a group, then the identity element and the inverse of a given element are both unique, i.e. if there is $e' \in G$ such that $x \circ e' = x$, then $e' = e$ and, moreover, for any $x' \in G$ such that $x \circ x' = e$, $x' = x^{-1}$.*

Definition A.2. An *abelian* group G is a group such that, for any $a, b \in G$, $a \circ b = b \circ a$.

Example: The integers $\mathbb{Z}$ with the operation of addition $+$, the unary function $-$ which takes every integer to its negative, and 0, is an abelian group.

We will use this notational abbreviation: we write x^n for $\overbrace{x \circ x \circ \cdots \circ x}^{n \text{ times}}$, where n is a natural number; x^0 is defined to be the identity of the group for any x in the group. As another abbreviation, we will often write $x \circ y$ as simply xy.

Using those conventions, the reader may also show the following proposition:

Proposition A.3. *G is an abelian group if, and only if, either, for any natural number n and for any $x, y \in G$, $(x \circ y)^n = x^n \circ y^n$ or, for any $x, y \in G$, $(x \circ y)^{-1} = x^{-1} \circ y^{-1}$.*

Definition A.3. If G is a group and $H \subseteq G$, then H is a *subgroup* of G, in symbols $H \leq G$, if, and only if, it satisfies the following conditions:

1. For any $a, b \in H$, $a \circ b \in H$;
2. For any $a \in H$, $a^{-1} \in H$.

Note that these conditions imply that $e \in H$: take any $a \in H$. By 2, $a^{-1} \in H$, so, by 1, $a \circ a^{-1} = e \in H$.

Definition A.4. Let G be a group and $H \leq G$. The right and left *cosets* of H in G for a given $g \in G$ are defined, respectively, as $Hg = \{hg \mid h \in H\}$ and $gH = \{gh \mid h \in H\}$. Similarly, we also have $gHg' = \{ghg' \mid h \in H\}$.

Here is another easy exercise for the reader:

Proposition A.4. *Let G be a group and $H \leq G$. Define a relation $\sim$ on G by setting $a \sim b$ iff $b \in aH$. Then $\sim$ is an equivalence relation.*

The above proposition remains valid if we exchange left cosets for right cosets, as the reader may readily verify.

Definition A.5. Let G be a group and $H \leq G$. H is said to be a *normal* subgroup of G, in symbols $H \triangleleft G$, iff for every $a \in G$, $aH = Ha$.

The concept defined of normal subgroup, as we will see, is central to group theory. Here are a couple of equivalent characterizations:

Proposition A.5. *Let G be a group and $H \leq G$. The following are equivalent:*

1. *$H \triangleleft G$;*
2. *For any $g \in G$, $gHg^{-1} = H$;*

3. For any $g \in G$, $gHg^{-1} \subseteq H$.

Proof. Suppose 1. Then $g(Hg^{-1}) = g(g^{-1}h) = (gg^{-1})H = H$; the first equality follows by normality and the second by associativity. So 1 implies 2 and 2 obviously implies 3, hence it remains to be seen that 3 implies 1. So suppose 3 and let $a \in gH$ be arbitrary, i.e. $a = gh$ for some $h \in H$. We need to show that $a \in Hg$. By the hypothesis, $ag^{-1} = ghg^{-1} \in H$, so $ag^{-1}g = a \in Hg$. as required. A similar argument shows that, for arbitrary $a \in Hg$, $a \in gH$. Therefore, $gH = Hg$ for any $g \in G$, i.e. H is normal. So 3 implies 1. ■

Proposition A.6. *Let G be a group and $H \triangleleft G$. Then the operation $aH \circ bH = (ab)H$ is well defined.*

Proof. Let $a' \in aH$ and $b' \in bH$. I claim $a'b' \in (ab)H$. Since H is normal, by hypothesis $b' \in Hb$, so $a' = ah_1$ for some $h_1 \in H$ and $b' = h_2 b$ for some $h_2 \in H$. Hence, $a'b' = ah_1h_2b$. But $h_1, h_2 \in H$, so $h_1h_2 \in H$ as well, whence $h_1h_2b \in Hb$. By normality again, this means that $h_1h_2b \in bH$, i.e. there is $h \in H$ such that $h_1h_2b = bh$. Thus, $a'b' = ah_1h_2b = abh \in (ab)H$, as required. ■

Using this result, it's possible to define the important concept of a quotient group:

Definition A.6. Let G be a group and $H \triangleleft G$. Set $G/H = \{aH \mid a \in G\}$. Then it's clear that the operation $aH \circ bH = (ab)H$ inherits the properties of the group operation in G, so that the structure $(G/H, \circ)$ is also a group, with $eH = H$ as the identity element. This group is called the *quotient group of G by H*.

Finally, we can define a group homomorphism:

Definition A.7. Let G, H be groups and $f : G \to H$ a function. Then f is a *group homomorphism* if it takes the identity in G to the identity in H and, moreover, $f(gg') = f(g)f(g')$ for any $g, g' \in G$. If f is an injective homomorphism, it is an *embedding* and if it is a bijective homomorphism it is an *isomorphism*. The set $\ker(f) = \{g \in G \mid f(g) = e_H\}$, where e_H is the identity of H, is called the *kernel* of f.

This allows us to give another nice characterization of the normal subgroups of G.

Proposition A.7. *Let G be a group. Then $H \triangleleft G$ if, and only if, $H = \ker(f)$ for some homomorphism f.*

Proof. Suppose that $H \triangleleft G$. Consider G/H and the canonical projection of G onto G/H, i.e. the function $f : G \to G/H$ such that $f(g) = gH$. It's not difficult to see that f is a group homomorphism, and that $\ker(f) = H$.

Conversely, if $f : G \to H$ is a homomorphism, let $K = \ker(f)$. In order to show that $K \triangleleft G$, it suffices, by Proposition A.5, to show that $gKg^{-1} \subseteq K$ for any $g \in G$. So suppose $a \in gKg^{-1}$ for an arbitrary g. Then $a = gkg^{-1}$ for some $k \in K$, whence $f(a) = f(gkg^{-1}) = f(g)f(k)f(g^{-1}) = f(g)f(g)^{-1} = e_H$, that is, $a \in K$, as required. ■

We are now almost ready to define the concept of group action. We just need one more result about the set of all bijections from a set to itself.

Definition A.8. Let S be any set. Then ${}^{S}S$ is the set of all functions from S to itself.

Proposition A.8. *Let S be an arbitrary set. Then the set $\mathrm{Sym}(S) = \{f \in {}^{S}S \mid f$ is a bijection$\}$ is a group under function composition.*

Proof. Just take the identity function on S as the group identity and the inverse functions as the group inverses. ■

Definition A.9. The group $\mathrm{Sym}(S)$ is called the *symmetric* group of S.

We are now ready to define a group action:

Definition A.10. Let G be a group and S a set. A *group action* is a homomorphism $f : G \to \mathrm{Sym}(S)$. If there is such an action, we say that G *acts on* S and that S is a G-set.

For convenience, if there is no risk of confusion, given an action $f : G \to \mathrm{Sym}(S)$, I will write $f(g)(s)$ for $g \in G$ and $s \in S$ as simply $g \cdot s$ (corresponding to the *left* action).

The next three definitions will be important in the sequel:

Definition A.11. Let $f : G \to \mathrm{Sym}(S)$ be an action and $s \in S$ an arbitrary element. Then the *orbit* of s under this action is the set $\mathrm{orb}(s) = \{g \cdot s \mid g \in G\}$.

Definition A.12. Let $f : G \to \mathrm{S}$ be an action and $s \in S$ arbitrary. Then the *stabilizer* of s under the action is the set $\mathrm{Stab}(s) = \{g \in G \mid g \cdot s = s\}$.

Definition A.13. Let $f : G \to \mathrm{Sym}(S)$ be an action. This action is said to be *transitive* if, for some $s \in S$, $\mathrm{orb}(s) = S$. In this case, G is also said to *act transitively* on S.

First, note that a group action has three important properties:

Proposition A.9. *Let $f : G \to \mathrm{S}$ be an action, e the identity in G, and id the identity in $\mathrm{Sym}(S)$. Then:*

1. *$e \cdot s = s$ for any $s \in S$;*
2. *$g \cdot (g' \cdot s) = (gg') \cdot s$;*

3. $(g)^{-1} \cdot s = (g^{-}1) \cdot s$ *(the left side is the inverse of the image of* g *under the action, whereas the right side has the image of the inverse of* g *under the action).*

Proof. Since f is a homomorphism, it must take the identity to the identity. Thus, $e \cdot s = \mathrm{id}(s) = s$, as required. Moreover, $g \cdot (g' \cdot s) = f(g)(f(g')(s)) = f(g) \circ f(g')(s) = f(gg')(s) = (gg') \cdot s$. Finally, $(g)^{-1} \cdot s = (f(g))^{-1}(s) = f(g)^{-1}(s) = f(g^{-1})(s) = g^{-1} \cdot s$. ■

In the text, I mentioned that equivalence relations correspond to orbits of elements under group actions. It's time to prove this basic result.

Proposition A.10. *Let* $f : G \to \mathrm{Sym}(S)$ *be an action. Then the relation* $\sim$ *on* S *given by* $s \sim s'$ *if, and only if,* $s' \in \mathrm{orb}(s)$ *is an equivalence relation. Conversely, given an equivalence relation* $\sim$ *on* S*, there is a group action* $f : G \to \mathrm{Sym}(S)$ *such that the orbit of a given point is precisely its equivalence class.*

Proof. Suppose $f : G \to \mathrm{Sym}(S)$ is a group action and define the relation $\sim$ as in the statement of the proposition. Using Proposition A.9, then, for any $s \in S$, $s \sim s$, since $e \cdot s = s$, where e is the identity in G. So $\sim$ is reflexive. If $s \in \mathrm{orb}(s')$, then, by definition, there is $g \in G$ such that $g \cdot s' = f(g) = s$. But then, $g^{-1} \cdot s = s$. So $\sim$ is symmetric. Finally, if $s \sim t$ and $t \sim u$, then $g \cdot s = t$ and $g' \cdot t = u$, so $u = g' \cdot (g \cdot s) = g'g \cdot s$, so $s \sim u$ as well, that is, $\sim$ is transitive.

To show the converse, let $G \subseteq \mathrm{Sym}(S)$ be the subset of permutations of S which respect the given equivalence relation, i.e. such that for any $g \in G$, $s \sim s'$ iff $g(s) \sim g(s')$. It's not difficult to show that this is a subgroup of $\mathrm{Sym}(S)$. The action is then defined to be the inclusion map. ■

This also means that if G acts on S, then the action *partitions* S, i.e. divide it into disjoint blocks.

Definition A.14. Let G be a group and $H < G$, not necessarily normal. Then we can define the *coset space* of H in G to be the set G/H of left cosets of H in G.

Proposition A.11. *If* G/H *is a coset space, the function* $f : G \to \mathrm{Sym}(G/H)$ *given by left multiplication, so that* $f(g)(aH) = (ga)H$ *for* $aH \in G/H$*, is an action.*

Proof. Clearly f takes the identity to the identity. Next, suppose $g, h \in H$ and consider an arbitrary aH. Then $f(gh)(aH) = (gha)H = g(ha)H = f(g)(haH) = f(g)(f(h)(aH))$, as required. ■

Proposition A.12. *Let* $f : G \to \mathrm{Sym}(S)$ *be a transitive action. Then, for any* $s \in S$ *there is a bijection between* $G/\mathrm{Stab}(s)$ *and* S*.*

Proof. For each left coset choose a representative g and define a map $\pi_s : G/\mathrm{Stab}(s) \to S$ such that $\pi_s(g\mathrm{Stab}(s)) = g \cdot s$. This is well defined, since if $g' \in g\mathrm{Stab}(s)$, then $g' = gh$ for some $h \in \mathrm{Stab}(s)$. Thus $\pi_s(g') = g' \cdot s = (gh) \cdot s = g \cdot (h \cdot s) = g \cdot s$. Since the action is transitive, π_s is surjective. Moreover, $\pi_s(g\mathrm{Stab}(s)) = \pi_s(g'\mathrm{Stab}(s))$ implies that $g \cdot s = g' \cdot s$, so $(g^{-1}g') \cdot s) = g^{-1} \cdot (g' \cdot s) = g^{-1} \cdot (g \cdot s) = s$, so $g^{-1}g' \in \mathrm{Stab}(s)$. Thus, $g' = g(g^{-1}g') \in g\mathrm{Stab}(s)$, so $g\mathrm{Stab}(s) = g'\mathrm{Stab}(s)$, which is what we wanted to prove. ■

A rigorous definition of homogeneous spaces and, more importantly, Klein geometries, would require too much technical machinery for me to give a self-contained exposition here. So instead I'll just make some quick observations, mostly following the discussion in Reid and Szendrői (2005, chap. 9). Note that the conception according to which a space is homogeneous if every point is indistinguishable from every other point can be precisely captured using our framework: a space is *homogeneous* if, and only if, the group action associated with the space is transitive. Furthermore, by the last proposition proven above, it will in general be possible to recover the entire space from the group G and a subgroup $H < G$ such that H is the stabilizer of some point of the space. In fact, this allows us to *define* the space to be the coset space of G/H.

Chapter 4

Coda: Criticism of Tarski's Proposal

In the last chapter, I proposed a more metaphysical reading of Tarski's proposal than what is customary in the literature. In this chapter, I want to show how this reading allows one to answer a number of criticisms leveled against the proposal. In particular, I propose to deal with three famous critiques, by, respectively, Solomon Feferman, Denis Bonnay, and Catarina Dutilh Novaes. The reason why I focus on these critics is admittedly selfish: I believe that answering their criticism may help to bring to focus some of the advantages of my metaphysical reading. Additionally, I want also to deal with a criticism that could be raised specifically against my proposal, coming from the nominalist camp.

The fact, however, that I reject the philosophical basis for their criticism of Tarski's proposal should not lead one to think that I dismiss their results as unimportant or uninteresting. Indeed, quite apart from their philosophical merits, I believe their own positive proposals lead to some very interesting mathematical results. To illustrate this, I dedicate three lengthy appendices to this chapter to developing some lines of thought present in Feferman's approach. The reason for choosing Feferman as the main case study for this more technical part is simple: it requires less mathematical background from the reader than Bonnay's work, which makes use of Galois theory and other results from oligomorphic structures. On her turn, Dutilh Novaes offers as a tentative proposal the idea of using invariance under bisimulation as a criterion for logicality, bringing her somewhat close to Bonnay's own proposal, which uses partial isomorphisms; she does not, however, develop these ideas further, so I preferred to focus on the more developed proposal of Feferman.

4.1 Eliminativism

The eliminativist tendency I have in mind is related to the link I have established between projectible properties and properties of the type.[1] Suppose for a moment that we have a

[1] I am here following the general strategy pursued by Heck (2011).

plausible account of such properties. Then, one may ask, isn't it possible to *eliminate* reference to the type altogether, relying on whatever principle explains such properties to explain away our type-theoretical language? Take, for instance, Bromberger's quasi-natural kinds. Suppose we have a sentence $\phi(\tau)$ which predicates a certain property to a quasi-natural kind τ. By hypothesis, such a property is projectible across τ, and hence every token a of τ shares the property. Moreover, also by hypothesis, we have some kind of nomological principle, say $\psi(x)$, which accounts for this property. We can thus explain away references to τ by paraphrasing ϕ into a universally quantified sentence $\forall x(\psi(x) \rightarrow \phi(x))$. Given that, in the account I want to propose, there will be in general to every type τ an explanatory principle $\psi(x)$, this strategy should be applicable across the board.[2]

There is reason to resist this broad eliminativist strategy, however. [3] As a consequence of pursuing such a strategy, the resources employed in the elimination of such entities must be of the kind acceptable to the eliminativist, say, physical objects and properties in the case of the physicalist. But this means that the reduction will only work if the explanatory principle involved in the account given of projectible properties will not be able to have recourse to types or other undesirable entities. This means that the eliminativist will generally have to paraphrase the explanatory principles themselves using only sparse resources, a formidable task in most cases.[4]

To go back to an example from the last section, consider word types. How can we account for the unity of the word tokens without having recourse to types? In fact, it's not even clear that there are any nominalistically acceptable properties which are projectible across a given word type, since even being very stringent about what kind of inscriptions or sound patterns count as a token of a type will still result in a wide variety of tokens: consider a token of a word as a series of taps in Morse code and a token of a word as typical inscription in a blackboard. The type theoretician can have recourse to properties such as their character length, their shared characters, their function in a linguistic system, etc. But what about the eliminativist? What properties can he appeal to? Similarly, the type theoretician can explain that these properties are projectible because of, say, the role these words play in given linguistic systems, which themselves are formulated by appeal to types.

[2]Heck (2011) calls this position "syntactic reductionism".

[3]The reader should consult Heck (2011, 2017) for a more detailed case against these types of eliminativist strategies, as well as Wetzel (2009) for a sustained case against nominalism about types.The reader of those texts will notice how much my position owns to both author's cases against nominalism, even if my own positive proposals differ from theirs.

[4]This type of objection seems to me decisive against proposals such as Klement (2017), which attempt to use higher-order resources in the service of syntactic reductionism. Klement also exploits the fact that types are generally defined by way of equivalence relations, but he does not say how to account for these relations themselves without making recourse to types.

It's not clear how the eliminativist would proceed here.[5]

Of course, as the old cliché goes, absence of evidence is not necessarily evidence of absence, and the above considerations are a far cry from forming a decisive argument against such eliminativism. Nonetheless, they do suggest that there is a certain naturalness in admitting types, and that theories which make use of them will most likely turn out to be simpler and more elegant than those which avoid them, since they will avoid cumbersome paraphrases and implausible principles.

4.2 Feferman's criticism

Feferman (1999) raises three criticisms against Tarski's proposal:

(a) The thesis assimilates logic to mathematics, more specifically to set-theory.

(b) The set-theoretical notions involved in explaining the semantics of $\mathcal{L}_{\infty,\infty}$ are not robust.

(c) No natural explanation is given by it of what constitutes the *same* logical operation over arbitrary basic domains.

We have dealt with (a) in the previous chapter. Criticism (b) is more delicate; it obviously depends on what we mean by "robust". Feferman glosses this as being relatively independent of features of the surrounding set-theoretical universe. For instance, the property of being uncountable is not robust, since a set which is uncountable in a given set-theoretical universe may not be uncountable in a generic extension. More precisely, Feferman proposes to gloss "robust" as *absolute*; of course, since being absolute is relative to a given theory, one may ask which set theory Feferman has in mind. Feferman (2010) proposes to assume as background theory *Kripke-Platek* without the axiom of infinity and allowing for urelements($\mathsf{KPU} - \mathsf{Inf}$); the reasoning being that this is a theory which does not "encapsulate any problematic set-theoretical content" (Feferman 2010, p. 17). Or, as Feferman (1999, p. 38) also makes clear, the point is that considerations about logicality should ideally be independent of "what there is": the more robust a notion, the less ontologically committed it is. Assuming this definition of robustness, by a theorem of Manders we have that finitary first-order predicate logic is the only logic which is absolute with respect to $\mathsf{KPU} - \mathsf{Inf}$.[6] Thus, accepting Feferman's criticism is tantamount to accepting the notions definable in first-order logic as the only logical notions.

[5] Again, cf. Wetzel (2009) for consideration of some nominalist proposals and why they are unconvincing absent some recourse to types.

[6] Cf. Appendix A for the technical details, including a proof of the theorem due to Väänänen.

As for criticism (c), considered by him as the strongest reason to reject Tarski's proposal, Feferman's idea is the following. Define a connective $\curlyvee$ as the "wombat disjunction":[7] the sentence $P \curlyvee Q$ is true iff either there are wombats in the universe and one or both of P and Q are true, or else there are no wombats in the universe and P and Q are both true. In other words, $\curlyvee$ behaves like the typical disjunction in domains with wombats and like the typical conjunction in domains in which there are no wombats. Since in each domain $\curlyvee$ behaves like a logical constant, it's clear that, in each domain, $\curlyvee$ is invariant under any permutations of the universe, so it apparently counts as logical according to Tarski's criterion. But (a) the definition of $\curlyvee$ essentially involves a reference to a non-logical notion, namely wombats, so it shouldn't be counted as logical; (b) even disregarding (a), it's still the case that $\curlyvee$ has a very different behavior depending on the domain on which it's applied. According to Feferman, however, there's a sense in which the logical operations (and this is certainly true of the typical logical operations, such as the ones from the first-order predicate calculus) have the same meaning across domains, in such a way that any (successful) logicality criterion should be able to explain how different applications of the same operation "connect naturally" (the expression is Feferman's) with each other.[8]

Observe that "unnatural" operators such as McGee's wombat disjunction would still appear even if we adopted the generalization of Tarski's thesis proposed by Sher, i.e. if instead of permutations of a single domain we considered bijections between domains. Instead of defining wombat disjunction, one could still define, e.g. an operator that would behave as disjunction in domains with countable cardinality and as conjunction in domains with uncountable cardinality.

Why Feferman considers this the "strongest reason" for rejecting Tarski's criterion is a mystery to me. First, as Sher (2008, p. 333) points out, one could similarly define a quantifier Q such that Q behaved like $\forall$ in domains of cardinality < 101, as $\exists$ in domains of cardinality $101 - 745$ and as $\neg\exists$ in all other cardinalities. Why should this quantifier not count as logical? Or one could define a set theoretical operation, say $\cap\!\!\!\cap$, which behaved like $\cap$ in domains of countable cardinality and as $\cup$ in domains of uncountable cardinality. Clearly, the latter is still a set-theoretical operation, though, perhaps, not a very natural one. This seems to show, as Sher rightly concludes, that the issue of "naturalness" is entirely separate from the issue of "logicalness", so to speak. Moreover, considering our analysis of the proposal from the previous chapter, it rather misses the mark: as discussed then, Tarski's proposal is better read not as domain-relative, but rather as concerning which objects are logical in our world. Of course, this means that the proposal is best read as a metaphysical thesis, one that coheres

[7] The example is McGee's. Cf. McGee (1996, 2004).

[8] In the second Appendix, I examine Feferman's own proposal for meeting this requirement, whereas in the third Appendix I present Casanovas's analysis of why Feferman's proposal is problematic.

well with a platonist ontology, and such an ontology is anathema to Feferman, who is well-known for his predicativist and anti-platonist standpoint. This brings us back to criticism (b).

As he states quite clearly in the conclusion of his article, Feferman (1999, p. 51) believes that any proposal about logical notions should be related to the "more empirical study of the role of logic in the exercise of human rationality". Therefore, it's not surprising that he eschews metaphysical considerations in discussions about logicality; if logicality is intimately tied to human rationality, it would seem to depend more on epistemological or broadly formal considerations than on ontological matters. Briefly put, if there is a divide between mind and world, then logicality falls on the "mind" side of the divide. But this just means that Feferman's concerns are not the same concerns which motivate Tarski's proposal. On my reading, Tarski's proposal is not concerned with logic as somehow related to "human reasoning", but rather with which objects and properties can be considered "logical"—dramatically speaking, which objects and properties are in a sense a fundamental part (the logical part) of the structure of the world. This inquiry is obviously concerned with "what there is", and hence metaphysical considerations, instead of being some kind of stain which contaminates it, are precisely what constitutes it. Therefore, ontological assumptions are not inherently problematic, at least insofar as they cohere with the overall metaphysical picture being described. We are thus free to employ whatever mathematical theory adequately fits in with our scheme.

This reveals a theme that will see recur in our examination of certain criticisms of Tarski's proposal: they often seem to presuppose that the proposal has as its target something other than describing which objects or properties are logical (a metaphysical endeavor), and hence fail to meet it on its own ground. To use the terminology from Chapter 1, it is as if critics did not pay attention to the crucial clarification step of Tarski's explication, finding fault with his explicatum for not matching a concept which simply isn't his explicandum.

4.3 Bonnay's criticism

The main objection raised by Bonnay (2008) is that targeted at Tarski's generality argument. Recall from the last chapter that the argument rests on the premise that the most general notions are those invariant under the biggest transformation group, which, as we mentioned, is simply the group of all permutations of the domain. Bonnay raises two objections to this characterization: first, why restrict such notions to intra-domain notions? This first objection has already been answered in the previous chapter, where I noted that Tarski's proposal is best construed as a proposal about our world, not just any domain. But Bonnay also has a second objection: why restrict oneself to *permutations* of the domain? Why not

allow for other kinds of functions or even relations? As he himself puts it:

> Now there are a lot of other concepts of similarity between structures which are used in model theory and in algebra which are far less demanding [than permutation invariance—D.N.]. Instead of requiring the structure to be fully preserved, they lower the requirement to some kind of approximate preservation. Why should we refrain from resorting to these other concepts? To sum up, even if one grants that generality is a good way to approach logicality, there is no evidence that the class of all permutations is the best applicant for the job. (Bonnay 2008, p. 10)

If one approaches the question with a purely mathematical framework, then indeed the generality argument seems to beg the question against more general proposals, as Bonnay makes clear. But if one approaches the question from a metaphysical perspective, then the *group* concept becomes entirely relevant, and this will allow us to allay Bonnay's worries. As we saw in the previous chapter, one metaphysical picture that coheres well with Tarski's proposal is the neo-Fregean platonist picture, according to which we are able to introduce or describe certain abstract objects by considering their identity conditions, which are given by equivalence relations. Notice that, since we are working with equivalence relations, we are also implicitly working with groups, as every equivalence relation can be considered as generated by a group action. Since we are working with groups, it is inevitable then that the most general group will be the full symmetric group of the domain; aside from very artificial examples, for a set of functions to form a group under composition they must all be bijections.

So Bonnay has, like Feferman, misunderstood the motivation behind Tarski's proposal. He seems to think that the proposal is purely mathematical, so that it becomes a puzzle why Tarski does not contemplate more general similarity relations. On my reading, however, the proposal is not purely mathematical. It aims at describing the *identity conditions for the logical objects and properties*. It is clear, under this proposal, that such identity conditions will be connected to equivalence relations, so it's only natural that we should look at the groups which generate such relations in order to gain better insight into them. Once one adopts this perspective, the class of all permutations, endowed with a binary operation of composition, *is* the most general class—it is, in Bonnay's words, the "best applicant for the job".

4.4 Dutilh Novaes

In a recent paper, Dutilh Novaes (2014) argues that Tarski's proposal is inadequate on two fronts: it counts too much and too little as logical to be a good proposal. The former charge

is basically the charge leveled by Feferman and Bonnay against the proposal, namely that it assimilates logic to mathematics. The latter charge is novel: it basically amounts to the claim that certain obvious logical notions are not counted as such by the proposal. In particular, Dutilh Novaes argues that the criterion excludes modal notions from counting as logical; since she considers modal notions to be logical, she concludes that the proposal must then be rejected.

Why does Dutilh Novaes considers modal notions to be logical, however? The reason she gives is connected to her adoption of the "practices" point of view:

> Thus, I submit that the failure of the permutation invariance criterion to count these modal operators as logical should make us reconsider the whole idea of permutation invariance as a criterion for logicality. After all, modal logics and their descendants are currently among the most widely studied logical systems; they are highly influential both for the interface of logic with computer science and for philosophical discussions on modalities and related topics. If a criterion for logicality deems the corresponding modal operators to be non-logical, this seems to be a real case of undergeneration from the point of view of practices. (Dutilh Novaes 2014, p. 95)

Otherwise put, there is already a well-established subject called "logic", just as there is a well-established practice of mathematics. Just as it would be strange for the philosopher to propose a definition of mathematics that excluded from its extension already entrenched subjects such as, say, algebraic geometry or higher set-theory, it is equally strange for a philosopher to put forward a definition of "logic" that excluded certain entrenched subjects, such as modal logic. The philosopher should thus occupy the role of the *second philosopher*, who is only allowed to make explicit what is already implicit in the practices of the scientists, and who can never contradict this practice.

Indeed, Dutilh Novaes's language in certain phrases almost makes it sound as if such pretensions of *first philosophy* are akin to a kind of *hubris*: "I am here suggesting that, if we [do not count modal logic as logic], as philosophers we will be excluding a vibrant portion of logical practices from the realm of analysis, which I take not to be recommended" (Dutilh Novaes 2014, p. 94n14). That is, it's not in our position qua philosophers to question the status of these *vibrant* portions of logical practices.

It should not come as a surprise if I say that this seems to be a confusion. Tarski's proposal is not a proposal about how to best describe whatever is studied in the logic departments, or to make explicit what is implicit in the practice of professional logicians. Similarly, the proposal should not lead to practical decisions of excluding modal logic from the logic courses and textbooks, under the pretense that such notions do not pass out test. Rather, it is a proposal about the fundamental structure of the world: which objects and properties should

count as logical? It is not surprising then that some objects and properties typically studied by the logicians should not be found among these, as they may not even be fundamental objects and properties to begin with. So, for instance, logicians may be interested in the study of epistemic logic, which contains a modal epistemic operator, say "knows". Why should this notion be counted as logical, as opposed to epistemic? The mere fact that we are able to model an object in what we call a logical system is no reason to suppose that this object is logical, anymore than the fact that we can model a DJ's vinyl scratching using Fourier transforms means that such scratching is a mathematical object.[9] We thus need some independent argument as to why modal notions should be considered as logical; Dutilh Novaes, however, does not provide any.

4.5 Conclusion

In this chapter, I have analyzed the three different criticisms against Tarski's proposal by Feferman, Bonnay, and Dutilh Novaes. In all three cases, I argued that such authors were aiming at different targets than Tarski: Feferman for something close to logic as human reasoning, Bonnay at a purely mathematical criterion for logicality, and Dutilh Novaes at an account of our logical practice. Indeed, at some points one gets the impression that these authors start with their own "intuitions" about what *logic* is, and then try to argue that Tarski's proposal is "counterintuitive" for not matching these intuitions. There are two problems here: the first is that such reliance on intuitions is out of place in this kind of investigation. An old term such as "logical" is bound to elicit diverse intuitions in different people, many of them conflicting; why should we trust one over the other? Moreover, and this is the second point, on my reading, Tarski's proposal is not a proposal about *logic*, it is a proposal about *logical objects*.

How could such eminent critiques have missed the true target of Tarski's proposal? Leaving aside personal prejudice against metaphysical inquiries (which is acute in the case of Feferman), another possible (I dare not say plausible) explanation is that the technical trappings of the proposal may have served to obscure the matter. In particular, it is not difficult to connect logical objects to logical operators (see the previous chapter), and then to infer that Tarski was after a characterization of which *languages* are logical. Since the framework proposed in Tarski's lecture is inadequate for tackling this question in its full generality, one is then led to search for broader frameworks. This is exactly the program initiated by Feferman and developed, in a somewhat different direction, by Bonnay. Once one is amidst this wealth of mathematical material, it is easy to get lost in technical minutiae and miss the forest for the trees.

[9] The reader intrigued by this Fourier scratching should consult Amiot (2016, pp. 149–151).

A Appendix: On the Absoluteness of First-Order Logic

In this appendix, I would like to focus on one of Feferman's objections, namely, that the logic $\mathcal{L}_{\infty\infty}$ is not "robust" in a certain demanding sense. Feferman (1999, p. 38) explains that a notion is "robust" if it has the "same meaning independent of the exact extent of the set-theoretical universe". This is rather vague, as no particular set theory is specified with which to cash out this specification. Later, Feferman (2010) gave a more precise criterion: a notion is "robust" iff it's absolute relative to KPU-Inf.[10] In particular, a logic is itself robust iff its syntax and satisfaction relation are robust. One result that is specially important, as it characterizes these "robust" notions, is the following theorem by Väänänen (1985), which I will call the "Main Theorem":

Theorem A.1 (Main Theorem). *$\mathcal{L}_{\omega\omega}$ is the only logic which is represented in HF and is absolute relative to* KPU-Inf.

I will proceed as follows: in the Section 1, I will provide the necessary set-theoretical background within which the theorem is proven; in particular, I'll define KPU-Inf and prove a series of useful theorems, as well as providing a general characterization of absolute formulas. In Section 2, I provide further background for the theorem, this time analyzing some model-theoretic notions that will play a key role in the argument. Specifically, Feferman's notion of *adequate to truth* is presented and explained. Finally, in section 3, the Main Theorem is proved. The proof itself is divided into two parts: the first part shows that $\mathcal{L}_{\omega\omega}$ is indeed absolute (by showing that its syntax set and its satisfaction relation are defined by Δ_0 formulas), while the second part shows that, if a logic is represented in HF and is absolute relative to KPU-Inf, then it's weaker than $\mathcal{L}_{\omega\omega}$ in a sense to be defined (since among regular logics $\mathcal{L}_{\omega\omega}$ is the weakest logic, it follows that it is the only logic to satisfy the conditions of the theorem).

A.1 Set-Theoretical Background

In this section, I'll present the theory KPU-Inf and also prove some results which will be needed later; most of this section is based on the material found in Barwise (1975). Let L be a first order language; the theory KPU-Inf will be formulated in a language $L^* = L \cup \{\in, \dots\}$. A structure $\mathfrak{A}_{\mathfrak{M}} = \langle \mathfrak{M}; A, E, \dots \rangle$ consists in:

(i) a structure $\mathfrak{M} = \langle M, \dots \rangle$ for the language L, with possibly $M = \emptyset$; the elements of M are called *urelements*;

(ii) a set $A \neq \emptyset$ and such that $A \cap M = \emptyset$; the elements of A are called *sets*;

[10] These notions will be explained in the following sections.

(iii) a relation $E \subseteq (M \cup A) \times A$, which will interpret the *membership* symbol $\in$;

(iv) possibly other relations, functions, and constants on $M \cup A$, as needed to interpret L^*.

Next, I define the collection of Δ_0 formulas of L^* as the smallest collection Y containing the atomic formulas and also closed under:

(i) if $\phi \in Y$, then $\neg\phi \in Y$;

(ii) if $\phi, \psi \in Y$, then $\phi \wedge \psi \in Y$ and $\phi \vee \psi \in Y$;

(iii) if $\phi \in Y$, then $\forall u \in v\phi$ and $\exists u \in v\phi$ are also in Y for any variables u and v.

It's now possible to state the KPU-Inf axioms. They consist in the universal closure of the following formulas:

Extensionality: $\forall x(x \in a \leftrightarrow x \in b) \rightarrow a = b$;

Foundation: $\exists x\phi(x) \rightarrow \exists x[\phi(x) \wedge \forall y \in x\neg\phi(y)]$ for all formulas $\phi(x)$ in which y does not occur free;

Pair: $\exists a(x \in a \wedge y \in a)$;

Union: $\exists b\forall y \in a\forall x \in y(x \in b)$;

Δ_0 **Separation:** $\exists b\forall x(x \in b \leftrightarrow x \in a \wedge \phi(x))$ for all Δ_0 formulas in which b does not occur free;

Δ_0 **Collection:** $\forall x \in a\exists y\phi(x, y) \rightarrow \exists b\forall x \in a\exists y \in b\phi(x, y)$ for all Δ_0 formulas in which b does not occur free.

Characterization of Absolute Formulas

Let $\mathfrak{A}_\mathfrak{M}$ be a structure for L^*. For $a \in A$, a_E is defined as: $a_E = \{y \in M \mid yEa\}$.

Definition A.1. Given another L^*-structure $\mathfrak{B}_\mathfrak{N}$, we say that $\mathfrak{B}_\mathfrak{N}$ is an *extension* of $\mathfrak{A}_\mathfrak{M}$, in symbols $\mathfrak{A}_\mathfrak{M} \subseteq \mathfrak{B}_\mathfrak{N}$, if $\mathfrak{M} \subseteq \mathfrak{N}$ (as L structures), $A \subseteq B$, and if the interpretations of $E, \ldots$ are just the restrictions to $M \cup A$ of the corresponding relations in $\mathfrak{B}_\mathfrak{N}$.

Definition A.2. $\mathfrak{B}_\mathfrak{N}$ is an *end extension* of $\mathfrak{A}_\mathfrak{M}$, in symbols, $\mathfrak{A}_\mathfrak{M} \subseteq_{\text{end}} \mathfrak{B}_\mathfrak{N}$, if $\mathfrak{A}_\mathfrak{M} \subseteq \mathfrak{B}_\mathfrak{N}$ and, for each $a \in A$, $a_E = a'_E$, where E' is relation corresponding to E in $\mathfrak{B}_\mathfrak{N}$.

Given the above, it's possible now to define the notion of *persistence* and *absoluteness* for a formula. If $\phi(u_1, \ldots, u_n)$ is a formula of L^*, we say that $\phi(u_1, \ldots, u_n)$ is persistent relative to a theory T of L^* if for all models $\mathfrak{A}_\mathfrak{M}, \mathfrak{B}_\mathfrak{N}$ of T such that $\mathfrak{A} \subseteq_{\text{end}} \mathfrak{B}_\mathfrak{N}$ and $x_1, \ldots, x_n$ in $\mathfrak{A}_\mathfrak{M}$, $\mathfrak{A}_\mathfrak{M} \models \phi[x_1, \ldots, x_n]$ implies $\mathfrak{B}_\mathfrak{N} \models \phi[x_1, \ldots, x_n]$; if $\mathfrak{B}_\mathfrak{N} \models \phi[x_1, \ldots, x_n]$ implies $\mathfrak{A}_\mathfrak{M} \models \phi[x_1, \ldots, x_n]$, then $\phi(u_1, \ldots, u_n)$ is *downward persistent*, and if $\mathfrak{A}_\mathfrak{M} \models \phi[x_1, \ldots, x_n]$ iff $\mathfrak{B}_\mathfrak{N} \models \phi[x_1, \ldots, x_n]$, then $\phi(u_1, \ldots, u_n)$ is said to be absolute.

Recall that a formula ϕ is Σ_1 if it is equivalent to a formula ψ such that ψ is $\exists x\theta$ such that θ is Δ_0. Analogously, ϕ is Π_1 if it is equivalent to a formula ψ such that ψ is $\forall x\theta$ and θ is Δ_0. Finally, a formula ϕ is Δ_1 if it is equivalent to formulas ψ, θ such that ψ is Σ_1 and θ is Π_1.

We can now prove a theorem that will later be useful. Let T be an arbitrary theory and ϕ a formula in the language of this theory. Then:

Theorem A.2. *ϕ is absolute relative to T iff ϕ is Δ_1.*

The proof here will follow the outline given in Robinson (1965, pp. 70-5). In order to prove this theorem, we will first prove two lemmas which jointly entail it.

Lemma A.1. *ϕ is downward persistent relative to T iff it is Π_1.*

Lemma A.2. *ϕ is persistent relative to T iff it is Σ_1.*

Proof of Lemma A.1. In one direction, let ϕ be a Π_1 sentence in the language of T, $\mathfrak{A}$ and $\mathfrak{B}$ two T-structures such that $\mathfrak{A} \subseteq \mathfrak{B}$, and suppose $\mathfrak{B} \models \phi$. Since ϕ is Π_1, there is a universal sentence, say θ, such that $T \vdash \theta \leftrightarrow \phi$. Thus, as $\mathfrak{B} \models \phi$, it follows that $\mathfrak{B} \models \theta$. By definition, this means that every sequence $\langle b_1, \ldots, b_n\rangle \in B$ satisfies θ. Let then $\langle a_1, \ldots, a_n\rangle \in A$ be an arbitrary sequence. Since $A \subseteq B$, this means that $\langle a_1, \ldots, a_n\rangle \in B$, whence, by the hypothesis, $\langle a_1, \ldots, a_n\rangle$ satisfies θ. Since this sequence was arbitrary, we may conclude that every such sequence from A also satisfies θ. Therefore, $\mathfrak{A} \models \theta$, whence $\mathfrak{A} \models \phi$.

For the other direction, let Γ be the set of all universal sentences γ in the language of T such that $T \vdash \phi \rightarrow \gamma$.[11] Consider now the set $\Lambda = T \cup \Gamma \cup \{\neg\phi\}$. Suppose, for reductio, that Λ is consistent. By the completeness theorem for first-order logic, it follows that Λ has a model, say, $\mathfrak{M}$. Since ϕ is downward persistent, it follows that there are no end-extensions $\mathfrak{M}'$ of $\mathfrak{M}$ such that $\mathfrak{M}'$ is a model for T and such that ϕ holds in $\mathfrak{M}'$, otherwise ϕ would hold in $\mathfrak{M}$ as well. Let $\Delta_\mathfrak{M}$ be the diagram of $\mathfrak{M}$. It's clear that $T \cup \Delta_\mathfrak{M} \vdash \neg\phi$. Thus, since only finitely many formulas were used in the proof, there is a finite subset $\Delta^*_\mathfrak{M}$ of $\Delta_\mathfrak{M}$ such that $T \cup \Delta^*_\mathfrak{M} \vdash \neg\phi$. Let $\delta(a_1, \ldots, a_k)$ be the conjunction of every formula in $\Delta^*_\mathfrak{M}$, with $a_1, \ldots, a_k$ individual constants not contained in T (if any). Since $T \cup \{\delta(a_1, \ldots, a_k)\} \vdash \neg\phi$, it follows by the Deduction Theorem that $T \vdash \delta(a_1, \ldots, a_k) \rightarrow \neg\phi$, so, by contraposition,

[11] Γ is never empty, as any tautological universal sentence will belong to it.

$T \vdash \phi \to \neg\delta(a_1, \ldots, a_k)$. As $a_1, \ldots, a_k$ are not contained in T, they are also not contained in ϕ, whence $T \vdash \phi \to \forall x_1, \ldots, x_n(\neg\delta(x_1, \ldots, x_n))$ (with $x_1, \ldots, x_n$ not free in δ). Now, $\forall x_1, \ldots, x_n(\neg\delta(x_1, \ldots, x_n))$ is a universal sentence implied by ϕ, so it belongs to Γ. Call this universal sentence γ. Since $\mathfrak{M}$ is a model for Γ, it follows that $\mathfrak{M} \models \gamma$. Notice, however, that $\neg\gamma$ is $\neg\forall x_1, \ldots, x_n \neg\delta(x_1, \ldots, x_n)$, which is equivalent to $\exists x_1, \ldots, x_n \delta(x_1, \ldots, x_n)$, which is clearly true in $\mathfrak{M}$, as $\delta(a_1, \ldots, a_k)$ is a conjunction of elements of $\Delta_{\mathfrak{M}}$ and hence holds in $\mathfrak{M}$. Thus, both γ and $\neg\gamma$ hold in $\mathfrak{M}$, which is absurd; therefore, Λ must be inconsistent.

Again, only finitely many formulas are used in the derivation of the inconsistency, so it follows that there is a finite $\Gamma^* \subseteq \Gamma$ such that $T \cup \Gamma^* \cup \{\neg\phi\}$ is inconsistent. Since Γ^* is finite, let θ be a universal sentence equivalent to the conjunction of every sentence in Γ^* (we know that such a θ exists by simple prenex equivalences). It's clear that $\theta \in \Gamma$ and that $T \vdash \neg(\theta \wedge \neg\phi)$. But $\neg(\theta \wedge \neg\phi)$ is equivalent to $\theta \to \phi$, whence $T \vdash \theta \to \phi$. Since $\theta \in \Gamma$, it follows that $T \vdash \phi \leftrightarrow \theta$, and, since θ is universal, it is Π_1, concluding the proof. ■

Proof of Lemma A.2. Suppose first that ϕ is Σ_1 relative to T, $\mathfrak{A} \subseteq \mathfrak{B}$ are two T-structures, and $\mathfrak{A} \models \phi$. Since ϕ is Σ_1, it follows that there is an existential sentence, say θ, such that $T \vdash \phi \leftrightarrow \theta$. As $\mathfrak{A} \models \phi$, $\mathfrak{A} \models \theta$. By definition, this means that there is a sequence $\langle a_1, \ldots, a_n\rangle \in A$ which satisfies θ. But $A \subseteq B$, so this sequence is also in B. Therefore, there is a sequence in B (the same sequence) which satisfies θ, whence $\mathfrak{B} \models \theta$. Therefore, $\mathfrak{B} \models \phi$.

Suppose now that ϕ is persistent relative to T. I claim that $\neg\phi$ is downward persistent relative to T. For suppose otherwise. Then there are models $\mathfrak{A}_{\mathfrak{M}}$ and $\mathfrak{B}_{\mathfrak{N}}$ such that $\mathfrak{A}_{\mathfrak{M}} \subseteq_{\text{end}} \mathfrak{B}_{\mathfrak{N}}$ and $\mathfrak{B}_{\mathfrak{N}} \models \neg\phi$ but $\mathfrak{A}_{\mathfrak{M}} \not\models \neg\phi$. Hence, $\mathfrak{A}_{\mathfrak{M}} \models \phi$. But then, since ϕ is persistent, $\mathfrak{B}_{\mathfrak{N}} \models \phi$, a contradiction. Thus, $\neg\phi$ is downward persistent. By Lemma A.1, this means that $\neg\phi$ is equivalent to a Π_1 sentence, say θ. From this, it follows that $\phi \leftrightarrow \neg\neg\phi \leftrightarrow \neg\theta$. Since θ is Π_1, $\neg\theta$ is Σ_1, concluding the proof. ■

Proof of Theorem A.2. A formula ϕ is absolute iff it is both persistent and downward persistent. So suppose it is absolute. Since it is persistent, by Lemma A.2, it is equivalent to a Σ_1 formula. Since it is downward persistent, by Lemma A.1, it is equivalent to a Π_1 formula. Thus, ϕ is Δ_1. ■

The Truncation Lemma

Another result which will be used here is the Truncation Lemma. Before stating the lemma, a few definitions are in order:

Definition A.3. Let $\mathfrak{A}_{\mathfrak{M}} = (\mathfrak{M}; A, E, \ldots)$ be any structure and consider $\mathscr{W} = \{\mathfrak{B}_{\mathfrak{M}} \subseteq_{\text{end}} \mathfrak{A}_{\mathfrak{M}} \mid \mathfrak{B}_{\mathfrak{M}} \text{ is well-founded}\}$. The largest $\mathfrak{B}_{\mathfrak{M}} \in \mathscr{W}$, i.e. the $\mathfrak{B}_{\mathfrak{M}}$ such that it's an end extension of all the other members of $\mathscr{W}$, is called the *well-founded part* of $\mathfrak{A}_{\mathfrak{M}}$.[12]

[12]That there is such a largest $\mathfrak{B}_{\mathfrak{M}}$ is a lemma proved in Barwise (1975, p. 72). The basic idea of the proof

Notice that, if $\mathfrak{B}_\mathfrak{M}$ is the well-founded part of a structure, then there is a unique isomorphism between it and a transitive structure in which the E relation is the real membership relation.[13] From now on, we will identify the well-founded part of a structure with its image under such an isomorphism.

Lemma A.3 (Truncation Lemma). *Let* $\mathfrak{A}_\mathfrak{M} = (\mathfrak{M}; A, E, \dots)$ *and* $\mathfrak{B}_\mathfrak{M} = (\mathfrak{M}; B, \in, \dots)$ *be* L^**-structures with* $\mathfrak{A}_\mathfrak{M} \models \mathsf{KPU}$ *and* $\mathfrak{B}_\mathfrak{M} \subseteq_{\mathrm{end}} \mathfrak{A}_\mathfrak{M}$*, where* $(\mathfrak{M}; B, \in)$ *is the well-founded part of* $\mathfrak{A}_\mathfrak{M}$*. Then* $\mathfrak{B}_\mathfrak{M} \models \mathsf{KPU}$*.*

I'll reproduce here the proof found in Barwise (1975, pp. 72-3). In order to prove this lemma, another lemma is needed:

Lemma A.4. *Let* $\mathfrak{A}_\mathfrak{M} \subseteq_{\mathrm{end}} \mathfrak{B}_\mathfrak{M}$*, with* $\mathfrak{B}_\mathfrak{M} \models \mathsf{KPU}$*. Suppose that, whenever* $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(a) = \alpha$*,* $a \in A$ *iff* $\alpha \in A$*. Suppose also that there is no ordinal* $\beta \in B$ *such that* β *is a least upper bound for the ordinals in* A*. Then, with the possible exception of foundation, all the axioms of* KPU *hold in* $\mathfrak{A}_\mathfrak{M}$*.*

The proof of Lemma A.3 then becomes merely a matter of showing that $\mathbb{B}_\mathfrak{M}$ satisfies the hypothesis of Lemma A.4. Before proving Lemma A.4, the following observation will prove useful:

Remark A.1. If $a, b \in B$, $\mathrm{rk}(a) \leq \mathrm{rk}(b)$ and $\mathrm{rk}(b) \in A$, then $\mathrm{rk}(a) \in A$.

Proof. If $\beta = \mathrm{rk}(b)$ and $\alpha = \mathrm{rk}(a)$, then $\alpha \leq \beta$, whence, by definition, $\alpha \in \beta$. But this means that $\alpha \in \beta_{E'}$, so, by Definition A.2, $\alpha \in \beta_E$. Again, by definition, this means that $\alpha \in M \cup A$; but M is the set of urelements and, since α is not an urelement (it's an ordinal, and being an ordinal is absolute), $\alpha \in A$.[14] So $\mathrm{rk}(a) \in A$. ■

Proof of Lemma A.4. Let's check each of the axioms in turn:

Extensionality: Suppose there are sets $a, b \in A$ such that $\mathfrak{A}_\mathfrak{M} \models \forall x(x \in a \leftrightarrow x \in b)$. Since this is a Δ_0 formula,[15] it's absolute, so $\mathfrak{B}_\mathfrak{M} \models \forall x(x \in a \leftrightarrow x \in b)$. As extensionality is true in $\mathfrak{B}_\mathfrak{M}$, it follows that $\mathfrak{B}_\mathfrak{M} \models a = b$; again, since this is absolute, it follows that $\mathfrak{A}_\mathfrak{M} \models a = b$.

is to consider the union of all structures in $\mathscr{W}$ and show that it's such that it's well founded and that $\mathfrak{A}_\mathfrak{M}$ is an end-extension of it.

[13] A proof of this fact is again to be found in Barwise (1975, p. 72).

[14] In particular, this shows that the ordinals in A are an initial segment of the ordinals in B.

[15] It's equivalent to $\forall x(x \in a \to x \in b) \wedge \forall x(x \in b \to x \in a)$.

Pair: Suppose that $x, y \in \mathfrak{A}_\mathfrak{M}$ and let $\alpha, \beta \in A$ be such that $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(x) = \alpha \wedge \mathrm{rk}(y) = \beta$. Now, if $\mathfrak{B}_\mathfrak{M} \models \gamma = (\alpha + 1) \cup (\beta + 1)$, then $\gamma \in A$, otherwise γ would be the least upper bound for the ordinals in A. Consider now the set b such that $\mathfrak{B}_\mathfrak{M} \models b = \{x, y\}$. Clearly, $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(b) = \gamma$. But then, since $\gamma \in A$, by the hypothesis, $b \in A$. Now, the formula $b = \{x, y\}$ is Δ_0,[16] thus absolute, so $\mathfrak{A}_\mathfrak{M} \models b = \{x, y\}$.

Union: Suppose $a \in A$; the objective is to find $b \in A$ such that $b = \bigcup a$. By hypothesis, $a \in B$, so $\mathfrak{B}_\mathfrak{M} \models b = \bigcup a$ for some b. Let $\alpha = \mathrm{rk}(a)$. It's clear that, for every $x \in b$, $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(x) \leq \mathrm{rk}(a)$, so $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(b) \leq \mathrm{rk}(a)$. Thus, $\mathrm{rk}(b) \in A$, whence, by hypothesis, $b \in A$. Since the formula $b = \bigcup a$ is absolute,[17] $\mathfrak{A}_\mathfrak{M} \models b = \bigcup a$.

Δ_0 Separation: Suppose $a, y \in \mathfrak{A}_\mathfrak{M}$. Let $\phi(x, y)$ be Δ_0. The objective is to find $a, b \in \mathfrak{A}_\mathfrak{M}$ such that $b = \{x \in a \mid \phi(x, y)\}$. In order to do so, we will use Δ_0 separation on $\mathfrak{B}_\mathfrak{M}$ and then "transfer" the result to $\mathfrak{A}_\mathfrak{M}$. So let $b \in B$ be such that it satisfies the above formula. It's clear that $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(b) \leq \mathrm{rk}(a)$, so, as $a \in A$, $\mathrm{rk}(a) \in A$, whence $\mathrm{rk}(b) \in A$. By hypothesis, this means that $b \in A$. Since $\phi(x, y)$ is absolute, it follows that $\mathfrak{A}_\mathfrak{M} \models b = \{x \in a \mid \phi(x, y)\}$.

Δ_0 Collection: In order to prove that this axiom holds in $\mathfrak{A}_\mathfrak{M}$, I'll need a theorem which I'll state but not prove, as it won't be used further:

Theorem A.3 (The Σ Reflection Principle). *For all Σ fomulas ϕ we have the following:*

$$\mathsf{KPU} \vdash \phi \leftrightarrow \exists a \phi^{(a)}$$

where a is a variable for sets not occurring in ϕ. In particular, every Σ formula is equivalent to a Σ_1 formula in KPU.

A proof sketch can be found in Barwise (1975, p. 16-7); the proof is by induction on the complexity of ϕ.

Now suppose the antecedent of the axiom, i.e. that $a \in \mathfrak{A}_\mathfrak{M}$, the formula $\phi(x, y)$ is Δ_0 with parameters from $\mathfrak{A}_\mathfrak{M}$ and that $\forall x \in a \exists y \phi(x, y)$ holds in $\mathfrak{A}_\mathfrak{M}$. Thus, since $y \in A$, by hypothesis $\mathfrak{B}_\mathfrak{M} \models \mathrm{rk}(y) = \alpha$ and $\alpha \in A$. Consider an arbitrary $x \in a$ and the corresponding y; it's clear that $\mathfrak{A}_\mathfrak{M} \models \phi(x, y)$. Since $\phi(x, y)$ is Δ_0, by absoluteness, $\mathfrak{B}_\mathfrak{M} \models \phi(x, y)$. Since x was arbitrary, it follows that $\mathfrak{B}_\mathfrak{M} \models \forall x \in a \exists \alpha \exists y(\mathrm{rk}(y) = \alpha \wedge \phi(x, y))$. As the initial quantifier is bounded, this is a Σ formula. Hence, by Theorem A.3, it's equivalent to a Σ_1 formula, that is, we can bound the second quantifier. Therefore, there is a $\beta \in B$ such that $\mathfrak{B}_\mathfrak{M} \models \forall x \in a \exists \alpha < \beta \exists y(\mathrm{rk}(y) = \alpha \wedge \phi(x, y))$.

[16] Its defining formula is given by $x \in b \wedge y \in b \wedge \forall z \in b(z = x \vee z = y)$.

[17] This formula is equivalent to the Δ_0 formula $\forall x \in b \forall y \in x(y \in a) \wedge \forall y \in a \exists x \in b(y \in x)$.

From the above, it follows that $\mathfrak{B}_\mathfrak{M} \models \forall x \in a \exists y(\mathrm{rk}(y) < \beta \wedge \phi(x,y)$. As the ordinals are well-founded in $\mathfrak{B}_\mathfrak{M}$, we can choose the least ordinal β which satisfies this condition. Clearly, $\beta \in A$, otherwise it would be a least upper bound for the ordinals in A, contradicting the hypothesis. Now, applying Δ_0 collection in $\mathfrak{B}_\mathfrak{M}$, there is a set $b \in B$ such that $\mathfrak{B}_\mathfrak{M} \models \forall x \in a \exists y \in b(\mathrm{rk}(y) < \beta \wedge \phi(x,y)$. Now, for every $y \in b$, $\mathrm{rk}(y) < \beta$, whence $\mathrm{rk}(b) \leq \beta$. Therefore, $b \in A$. But $\forall x \in a \exists y \in b(\mathrm{rk}(y) < \beta \wedge \phi(x,y)$ is a Δ_0 formula, hence absolute, and all its parameters are in $\mathfrak{A}_\mathfrak{M}$, so it holds in $\mathfrak{A}_\mathfrak{M}$. It follows that there is a set b satisfying the consequent of the axiom, concluding the proof. ■

It's now possible to prove Lemma A.3. Observe first that, since $\mathfrak{B}_\mathfrak{M}$ is the well-founded part of $\mathfrak{A}_\mathfrak{M}$, it clearly satisfies Foundation. It remains to be seen then that it satisfies the remaining axioms. As noted above, in order to show this, instead of proving each axiom directly, it'll suffice to show that $\mathfrak{B}_\mathfrak{M}$ satisfies the hypothesis of Lemma A.4. The following proposition will be useful in this proof:

Proposition A.1. *If $a \in A$ and $a_E \subseteq B$, then $a \in B$.*

Proof. Suppose, for reductio, that $a \notin B$ and consider the structure $\mathfrak{B}'_\mathfrak{M} = (\mathfrak{M}; B', \in, \dots)$ such that $B' = B \cup \{a\}$. But $\mathfrak{B}'_\mathfrak{M}$ is also well-founded, for suppose it's not; then there is an infinite descending $\in$-chain in $\mathfrak{B}'_\mathfrak{M}$. But the only set which is in $\mathfrak{B}'_\mathfrak{M}$ that is not in $\mathfrak{B}_\mathfrak{M}$ is a, so this chain must have already been in $\mathfrak{B}_\mathfrak{M}$, contradicting the hypothesis. Thus, $\mathfrak{B}_\mathfrak{M} \subseteq_{\mathrm{end}} \mathfrak{B}'_\mathfrak{M}$ and $\mathfrak{B}'_\mathfrak{M}$ is well-founded, contradicting the maximality of $\mathfrak{B}_\mathfrak{M}$. Therefore, $a \in B$. ■

Proof of Lemma A.3. Let $a \in B$ and suppose that $\mathfrak{A}_\mathfrak{M} \models \mathrm{rk}(a) = \alpha$. We will show that $\alpha \in B$ by $\in$-induction on a. First, it's clear that $\mathrm{rk}(\emptyset) = \emptyset \in B$. Now suppose that, for every $x \in a$, $\mathrm{rk}(x) \in B$. Consider an arbitrary $y \in \alpha$. By definition, we know that $\mathrm{rk}(a) = \sup\{\mathrm{rk}(x) + 1 \mid x \in a\}$, so $y \in \mathrm{rk}(x) + 1$ for some $x \in a$. Thus, $y \in \mathrm{rk}(x) \cup \{\mathrm{rk}(x)\}$, that is, $y \leq \mathrm{rk}(x)$. By Remark A.1, it follows that $y \in B$. Therefore, $\alpha \subseteq B$. By Proposition A.1, $\alpha \in B$.

Suppose now that $\alpha \in B$ and that $\mathfrak{A}_\mathfrak{M} \models \mathrm{rk}(a) = \alpha$. We'll show that $a \in B$ by $\in$-induction on α. The base case is obvious: if $\mathrm{rk}(a) = \emptyset$, then $a = \emptyset$, whence $a \in B$. Assuming the induction hypothesis, consider an arbitrary $b \in a$. Since $b \in a$, $\mathrm{rk}(b) < \mathrm{rk}(a)$, whence, by the induction hypothesis, $b \in B$. Since b was arbitrary, it follows that $a \subseteq B$. Thus, by Proposition A.1, $a \in B$.

Therefore, if $\mathfrak{A}_\mathfrak{M} \models \mathrm{rk}(a) = \alpha$, then $a \in B$ iff $\alpha \in B$. It remains to be seen that there is no least upper bound in A for the ordinals in B. But this follows clearly from Proposition A.1.[18] Therefore, $\mathfrak{B}_\mathfrak{M}$ satisfies the hypothesis of Lemma A.4, whence $\mathfrak{B}_\mathfrak{M} \models \mathsf{KPU}$. ■

[18] Suppose there is a least upper bound, say α. Then $\alpha = \sup\{\beta \mid \beta \in B\}$. Thus, for every $\beta \in \alpha$, $\beta \in B$, whence, by Proposition A.1, $\alpha \in B$, contradicting the hypothesis.

Sets of Hereditary Cardinality less than a Cardinal κ

In this section, we will develop a bit of the theory of $H(\kappa)$ for infinite κ. The importance of these sets is that each $H(\kappa)$ is a model for KPU, a fact that will be used in the proof of our main theorem. First, we define by recursion the sets V_α; remember that our metatheory is ZFC, that M is the set of urelements and $\mathcal{P}(X)$ is the powerset of X

$$\begin{aligned} V_0 &= \emptyset \\ V_{\alpha+1} &= \mathcal{P}(M \cup V_\alpha) \\ V_\alpha &= \bigcup_{\beta<\alpha} V_\beta \text{ for } \alpha \text{ limit.} \end{aligned}$$

We also define $V_M = \bigcup V_\alpha$. Obviously, V_M is not a set, yet it will be a useful abbreviation. Given this definition, it's possible to define $H(\kappa)$:

Definition A.4. For any infinite cardinal κ, we define the *set of hereditary cardinality less than* κ, $H(\kappa)$, as follows: $H(\kappa) = \{a \in V_M \mid \mathrm{TC}(a) \text{ has cardinality less than } \kappa\}$, where $\mathrm{TC}(a)$ is the transitive closure of a.

The main theorem of this section is the following:

Theorem A.4. *For all infinite cardinals $\kappa > \omega$, the set $H(\kappa)_{\mathfrak{M}} = (\mathfrak{M}; H(\kappa), \in)$ is a model for* KPU. *If $\kappa = \omega$, then $H(\omega)$ is a model for* KPU-Inf.

Proof of Theorem A.4. Let $H(\kappa)$ be as in the hypothesis. We will show that each of the axioms holds in $H(\kappa)$. Most of the axioms are rather straightforward; the only one which is a bit more involved is Δ_0 Collection, whose proof is different depending on whether κ is regular or singular.

Extensionality: We need to show that, $\forall z \in H(\kappa) \forall x \in H(\kappa) \forall y \in H(\kappa)[(z \in x \leftrightarrow z \in y) \rightarrow x = y]$. So consider arbitrary $x, y \in H(\kappa)$ and let $z \in x$. Since $H(\kappa)$ is transitive, $z \in H(\kappa)$, which implies that $z \in y$, that is, $x \subseteq y$. Similarly, for y, that is, $y \subseteq x$. Therefore, $x = y$.

Pair: Let $x, y \in H(\kappa)$. Consider the set $b = \{x, y\}$. Notice that $\mathrm{TC}(b) = \mathrm{TC}(x) \cup \mathrm{TC}(y) \cup \{x\} \cup \{y\}$, thus $|\mathrm{TC}(b)| < \kappa$, that is, $b \in H(\kappa)$.

Union: Let $x \in H(\kappa)$ be arbitrary and consider $\bigcup x$. It's clear that $\bigcup x \subseteq \mathrm{TC}(x)$, whence $|\mathrm{TC}(\bigcup x)| \leq |\mathrm{TC}(x)|$, so $\bigcup x \in H(\kappa)$.

Foundation: Notice that $H(\kappa) \subseteq V_\alpha$ for some α. Therefore, since each V_α is well-founded, it follows that $H(\kappa)$ is well-founded as well.

Δ_0 **Separation:** Consider $x \in H(\kappa)$ and $y \subseteq x$. Since $y \subseteq x$, $\mathrm{TC}(y) \subseteq \mathrm{TC}(x)$, whence $|\mathrm{TC}(y)| \leq |\mathrm{TC}(x)| < \kappa$, so that $y \in H(\kappa)$. Observe that this gives us a stronger result, namely, that $H(\kappa)$ actually satisfies full separation.

Δ_0 **Collection:** Consider $x \in H(\kappa)$ and suppose that, for every $y \in x$, there is a $z \in H(\kappa)$ such that $\phi(y, z)$. We need to show that there is a set $b \in H(\kappa)$ such that $b = \{z \in H(\kappa) \mid \phi(y, z)\}$. As mentioned above, the proof is divided by cases. Suppose first that κ is regular. Since $x \in H(\kappa)$, its cardinality is less than κ. Thus, there are less than κ zs satisfying the hypothesis. Let b be as desired. For every $z \in H(\kappa)$, $\mathrm{TC}(z) < \kappa$. Thus $|\mathrm{TC}(b)| = \sup\{|\mathrm{TC}(z)| \, | z \in b\}$. As κ is regular, it's clear that $|\mathrm{TC}(b)| < \kappa$. Thus, $b \in H(\kappa)$. Notice that we didn't use in the preceding the hypothesis that ϕ is Δ_0, so, for κ regular, we actually obtain full comprehension. This is not surprising, as it is known that $H(\kappa)$ for κ regular is a model for the Replacement axioms, and full collection is equivalent to Replacement.

If κ is singular, consider $\kappa+1$. Suppose $H(\kappa)_{\mathfrak{M}} \models \forall y \in x \exists z \phi(y, z)$. Since every successor cardinal is regular, it follows by the above that $H(\kappa+1) \models \mathsf{KPU}$. Since $H(\kappa) \subseteq_{\mathrm{end}} H(\kappa+1)$, and $\forall y \in x \exists z \phi(y, z)$ is a Σ_1 formula, it is persistent, thus $H(\kappa + 1) \models \forall y \in x \exists z \phi(y, z)$. Now ϕ has only a finite number of symbols, so we can consider the reduction of L^* to the language of ϕ. Since this language is countable, we may apply the Löwenheim-Skolem theorem[19] to $H(\kappa + 1)_{\mathfrak{M}}$ to obtain a structure $\mathfrak{A}_{\mathfrak{M}}$ with $|\mathfrak{A}_{\mathfrak{M}}| < \kappa$ such that $\mathfrak{A}_{\mathfrak{M}} \models \mathsf{KPU}$ and $\mathfrak{A}_{\mathfrak{M}} \models \forall y \in x \exists z \phi(y, z)$. Consider now the transitive collapse[20] of $\mathfrak{A}_{\mathfrak{M}}$, say, $\mathfrak{A}'_{\mathfrak{M}}$. As $\mathfrak{A}'_{\mathfrak{M}}$ is transitive and with cardinality less than κ, every set $a \in A$ is such that $|\mathrm{TC}(a)| < \kappa$, so $A \cup M \subseteq H(\kappa)$, whence $\mathfrak{A}'_{\mathfrak{M}} \subseteq_{\mathrm{end}} H(\kappa)_{\mathfrak{M}}$. Since $\mathfrak{A}'_{\mathfrak{M}} \models \mathsf{KPU}$ and $\mathfrak{A}'_{\mathfrak{M}} \models \forall y \in x \exists z \phi(y, z)$, we can apply Δ_0 Collection to obtain $\mathfrak{A}'_{\mathfrak{M}} \models \exists b \forall y \in x \exists z \in b \phi(y, z)$. But this formula is Σ_1, thus persistent. Therefore, $H(\kappa)_{\mathfrak{M}} \models \exists b \forall y \in x \exists z \in b \phi(y, z)$. ■

A.2 Model-theoretic Background

In order to prove our main theorem, we will need some tools from model theory. Particularly important will be the notions involved in the comparison between given logics and the notion of *adequacy to truth*, first developed by Feferman (1974), but used here as formulated by Väänänen (1985). We divide this section into three sub-sections: in the first, we will develop general model-theoretic tools needed for the proof. In the second, we will develop

[19] We rely here on the fact that KPU is a first-order theory and that the cardinality of the language in question is countable to apply the theorem and get a structure with the desired cardinality. For a proof of this theorem specific to the context of KPU, cf. Barwise (1975, p. 52-3).

[20] This technique is applicable in KPU. I won't prove this result here, as it would lengthen an already lengthy appendix. For a proof, cf. Barwise (1975, pp. 30-33). Note that $\mathfrak{A}_{\mathfrak{M}}$ is a model for Extensionality, so the transitive collapse is applicable here.

the notion of adequacy to truth, which is also crucial for the proof. Finally, in the third, we develop some further notions that will also be employed in the proof of the main theorem.

For completeness sake, we include here some standard definitions that will be used throughout the rest of the appendix.[21]

Definition A.5. We define a *multi-sorted* vocabulary as a non-empty set τ consisting of sort symbols $r, s, t, \ldots$, finitary relation symbols $P, R, S, \ldots$, finitary function symbols $f, g, \ldots$, and constants $c, d, \ldots$. Each constant and function symbol of a vocabulary τ is associated with a sort symbol from τ, as are the argument places of the relation and function symbols; the *one-sorted* case is a special case of the many-sorted, in which we simply drop the sort symbols. Structures $\mathfrak{A}$ are denoted in the obvious way, if necessary subscripting the domains, function, and relations with their respective sorts (e.g. the sort s domain is denoted by A_s, etc.); the class of τ-structures is denoted by $\mathrm{Str}[\tau]$.

Definition A.6. By the *reduct of* $\mathfrak{A}$ *to vocabulary* σ, in symbols $\mathfrak{A} \restriction \sigma$, we denote the result of restricting a τ structure $\mathfrak{A}$ to the σ structure ($\sigma \subseteq \tau$) which arises by "forgetting" the sorts, relations, etc., not in σ. If τ is one-sorted, the *relativized reduct* of $\mathfrak{A}$ toσ and $P^{\mathfrak{A}}$, in symbols $(\mathfrak{A} \restriction \sigma) \mid P^{\mathfrak{A}}$, we denote the reduct of $\mathfrak{A}$ to σ relativized to $P^{\mathfrak{A}}$, where $P^{\mathfrak{A}}$ is a unary relation symbol in τ and is such that $P^{\mathfrak{A}}$ is σ-closed in $\mathfrak{A} \restriction \sigma$; i.e. if $c^{\mathfrak{A}} \in P^{\mathfrak{A}}$ for $c \in \sigma$, and $P^{\mathfrak{A}}$ is closed under $f^{\mathfrak{A}}$ for $f \in \sigma$.

Definition A.7. A *logic* is a pair $(\mathcal{L}, \models_{\mathcal{L}})$, where $\mathcal{L}$ is a mapping defined on vocabularies τ such that $\mathcal{L}[\tau]$ is a class (the class of $\mathcal{L}$-sentences of vocabulary τ) and $\models_{\mathcal{L}}$ is a relation between structures and $\mathcal{L}$-sentences. Generally, a logic is also required to obey further properties, such as the reduct, isomorphism, and renaming properties, but those will not be all too important here. For more details, cf. Ebbinghaus (1985, p. 28).

Comparing Logics

In this section, we will deal mainly with the notions of elementary classes, projective classes,[22] and relativized projective classes.[23] Most of the definitions in this section can be found, almost verbatim, in Ebbinghaus (1985). The proofs, however, unless otherwise noted, are our own.

Definition A.8. Let $\mathcal{L}$ be a logic and $\mathscr{K}$ a class of τ-structures. We say that:

1. $\mathscr{K}$ is an *elementary class* in $\mathcal{L}$, in symbols $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}}$, iff there is $\phi \in \mathcal{L}[\tau]$ such that $\mathscr{K} = \mathrm{Mod}^{\tau}_{\mathcal{L}}(\phi)$.

[21] Most such definitions can be found in Ebbinghaus (1985, §1.1).

[22] In the current literature, one often sees the nomenclature *pseudo-elementary class*. Cf., e.g., Hodges (2004, p. 206).

[23] Sometimes also referred to as PC'_{Δ}. Cf. Hodges (2004, p. 208).

2. $\mathscr{K}$ is a *projective class* in $\mathcal{L}$, in symbols, $\mathscr{K} \in \mathrm{PC}_{\mathcal{L}}$, iff there is $\tau' \supseteq \tau$ having the same sort symbols as τ and a class $\mathscr{K}'$ of τ'-structures, $\mathscr{K}' \in \mathrm{EC}_{\mathcal{L}}$, such that $\mathscr{K} = \{\mathfrak{A} \restriction \tau \mid \mathfrak{A} \in \mathscr{K}'\}$, the class of τ-reducts of $\mathscr{K}'$.

3. $\mathscr{K}$ is a *relativized projective class* in $\mathcal{L}$, in symbols $\mathscr{K} \in \mathrm{RPC}_{\mathcal{L}}$ iff, in the one-sorted case, there is $\tau' \supseteq \tau$, a unary relation symbol $U \in \tau' \setminus \tau$, and a class $\mathscr{K}'$ of vocabulary τ', $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}}$, such that $\mathscr{K} = \{(\mathfrak{A} \restriction \tau) \mid U^{\mathfrak{A}} \mid \mathfrak{A} \in \mathscr{K}$ and $U^{\mathfrak{A}}$ is τ-closed in $\mathfrak{A}\}$, or, in the many sorted case, iff there is $\tau' \supseteq \tau$ and a class $\mathscr{K}'$ of τ'-structures, $\mathscr{K}' \in \mathrm{EC}_{\mathcal{L}}$, such that $\mathscr{K} = \{\mathfrak{A} \restriction \tau \mid \mathfrak{A} \in \mathscr{K}'\}$. Here, $(\mathfrak{A} \restriction \tau) \mid U^{\mathfrak{A}}$ is the relativized reduct of $\mathfrak{A}$.

Definition A.9. Let $\mathcal{L}$ and $\mathcal{L}^*$ be logics. We say that $\mathcal{L}^*$ is *as strong as* $\mathcal{L}$, in symbols $\mathcal{L} \leq \mathcal{L}^*$, iff every class EC in $\mathcal{L}$ is EC in $\mathcal{L}^*$. Similarly, $\mathcal{L}$ and $\mathcal{L}^*$ are *equally strong* or *equivalent*, in symbols $\mathcal{L} \equiv \mathcal{L}^*$, iff both $\mathcal{L} \leq \mathcal{L}^*$ and $\mathcal{L}^* \leq \mathcal{L}$. Finally, we say that $\mathcal{L}^*$ is *stronger than* $\mathcal{L}$, in symbols $\mathcal{L} < \mathcal{L}^*$, iff $\mathcal{L} \leq \mathcal{L}^*$ and not $\mathcal{L}^* \equiv \mathcal{L}$. The notions of $\mathcal{L} \leq_{(R)PC} \mathcal{L}^*$, $\mathcal{L} \equiv_{(R)PC} \mathcal{L}^*$, and $\mathcal{L} <_{(R)PC} \mathcal{L}^*$ are defined in an analogous way.

Definition A.10. A class $\mathscr{K}$ of τ-structures is said to be Δ in $\mathcal{L}$ iff $\mathscr{K}$ and $\bar{\mathscr{K}} = \mathrm{Str}[\tau] \setminus \mathscr{K}$ are (R)PC in $\mathcal{L}$. A logic $\mathcal{L}$ has the Δ-interpolation property iff every Δ class of $\mathcal{L}$ is EC in $\mathcal{L}$. The Δ-*closure* of $\mathcal{L}$, in symbols $\Delta(\mathcal{L})$, is the logic that has as elementary classes just the classes that are Δ in $\mathcal{L}$.

With these definitions, we can now prove the following lemmas:

Lemma A.5. *$\mathcal{L}' \leq_{RPC} \mathcal{L}$ iff $\mathcal{L}' \leq \Delta(\mathcal{L})$.*

Lemma A.6. $\mathcal{L}_{\omega\omega} \equiv \Delta(\mathcal{L}_{\omega\omega})$

Proof of Lemma A.5. Observe first that, if $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}}$ for some $\mathcal{L}$, then $\bar{\mathscr{K}} \in \mathrm{EC}_{\mathcal{L}}$ as well.[24] It's also clear that every EC class is also a (R)PC class.[25] So suppose $\mathcal{L}' \leq_{RPC} \mathcal{L}$ and let $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}'}$. From this it follows that $\mathscr{K}, \bar{\mathscr{K}} \in (\mathrm{R})\mathrm{PC}_{\mathcal{L}'}$, whence, by the hypothesis, $\mathscr{K}, \bar{\mathscr{K}} \in (\mathrm{R})\mathrm{PC}_{\mathcal{L}}$. But then, by definition, $\mathscr{K}$ is Δ in $\mathcal{L}$, whence $\mathscr{K} \in \mathrm{EC}_{\Delta(\mathcal{L})}$.[26]

Conversely, suppose $\mathcal{L}' \leq \Delta(\mathcal{L})$ and let $\mathscr{K} \in \mathrm{RPC}_{\mathcal{L}'}$. This means that there is a $\mathscr{K}' \in \mathrm{EC}_{\mathcal{L}'}$ such that $\mathscr{K} = \{(\mathfrak{A} \restriction \tau) \mid U^{\mathfrak{A}} \mid \mathfrak{A} \in \mathscr{K}$ and $U^{\mathfrak{A}}$ is τ-closed in $\mathfrak{A}\}$. But then, since $\mathscr{K}' \in \mathrm{EC}_{\mathcal{L}'}$, by the hypothesis, $\mathscr{K}' \in \mathrm{EC}_{\Delta(\mathcal{L})}$, whence, by definition, $\mathscr{K}' \in (\mathrm{R})\mathrm{PC}_{\mathcal{L}}$. It follows that there is a $\mathscr{K}'' \in \mathrm{EC}_{\mathcal{L}}$ such that $\mathscr{K}' = \{(\mathfrak{A}' \restriction \tau') \mid P^{\mathfrak{A}'} \mid \mathfrak{A}' \in \mathscr{K}''$ and $P^{\mathfrak{A}'}$ is τ'-closed in $\mathfrak{A}'\}$. Let's abbreviate $(\mathfrak{A}' \restriction \tau') \mid P^{\mathfrak{A}'}$ to $\mathfrak{A}'_P$. Thus, $\mathscr{K} = \{(\mathfrak{A}'_P \restriction \tau) \mid U^{\mathfrak{A}'_P} \mid \mathfrak{A}'_P \in \mathscr{K}$ and $U^{\mathfrak{A}'_P}$ is τ-closed in $\mathfrak{A}'_P\}$. Therefore, $\mathscr{K} \in (\mathrm{R})\mathrm{PC}_{\mathcal{L}}$, as desired. ■

[24] Proof: If $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}}$, then there is a sentence $\phi \in \mathcal{L}$ such that $\mathscr{K}$ is the class of all models of ϕ. Thus, $\bar{\mathscr{K}}$ is the class of all structures $\mathfrak{A} \in \mathrm{Str}[\tau]$ such that $\mathfrak{A} \not\models \phi$. But then, if $\mathfrak{A} \in \bar{\mathscr{K}}$, it follows that $\mathfrak{A} \models \neg\phi$, so $\bar{\mathscr{K}}$ is the class of all models of $\neg\phi$. Therefore, $\mathscr{K} \in \mathrm{EC}_{\mathcal{L}}$.

[25] One can always consider the reduct of each structure $\mathfrak{A} \in \mathscr{K}$ ($\mathscr{K} \in$ EC) to the basic vocabulary of ϕ; if the logic in question also contains sentences such as $\forall x U x$, then every PC class can be made into a RPC class.

[26] This part of the proof is based on the one found in Makowski et al. (1976, p. 168).

Proof of Lemma A.6. That $\mathcal{L}_{\omega\omega} \leq \Delta(\mathcal{L}_{\omega\omega})$ is clear from the definition of Δ-closure. For the other direction, observe first that $\mathrm{EC}_{\Delta(\mathcal{L}_{\omega\omega})} = \mathrm{EC}_{\mathcal{L}_{\omega\omega}} \cup \{\mathscr{K} \in \mathrm{PC}_{\mathcal{L}_{\omega\omega}} \mid \bar{\mathscr{K}} \in \mathrm{PC}_{\mathcal{L}_{\omega\omega}}\}$. Thus, in order to show that $\Delta(\mathcal{L}_{\omega\omega}) \leq \mathcal{L}_{\omega\omega}$, one needs to show that every Δ class in $\mathcal{L}_{\omega\omega}$ is also an elementary class. But that is just the Souslin-Kleene property, which follows from Craig's interpolation theorem.[27] Therefore, $\Delta(\mathcal{L}_{\omega\omega}) \leq \mathcal{L}_{\omega\omega}$, which ends the proof of the lemma. ■

Adequacy to Truth

We can now deal with adequacy to truth and related notions. First, a few definitions are in order:

Definition A.11. Consider any set a. Define $\mu_a(x)$ by recursion as the possibly infinitary formula in the vocabulary $\tau_{\mathrm{set}} = \{\in\}$:

$$\mu_a(x) = \forall y(y \in x \leftrightarrow \bigvee_{b \in a} \mu_b(y)).$$

The idea is that $\mu_a(x) \leftrightarrow x = a$, or, at least, that x has the same set-theoretical structure as a ($x = a$ will indeed take place in any transitive set containing $\mathrm{TC}(\{a\})$). Define now by recursion $\pi_a(x)$ as follows:

$$\pi_a(x) = \mu_a(x) \wedge \bigwedge_{b \in \mathrm{TC}(a)} \exists y \mu_b(y).$$

Let $\mathfrak{B} = (B, E)$ be a model for extensionality and consider $\mathfrak{B}_0$, the well-founded part of $\mathfrak{B}$ (cf. Definition A.3). Since $\mathfrak{B}$ is extensional and $\mathfrak{B}_0$ is well-founded, it's possible to apply the transitive collapse theorem to obtain an isomorphism $i : \mathfrak{B}_0 \to \mathfrak{A}$ (A being a transitive set) and such that $\mathfrak{B} \models \pi_a(x)$ iff $x \in B_0, a \in A$, and $i(x) = a$.

Definition A.12. Throughout this appendix, we will simply assume that associated with each logic $\mathcal{L}$ there is a transitive set A such that $\mathcal{L}[\tau] \subseteq A$ for all τ considered. This set will be called the *syntax* set of $\mathcal{L}$. If A is closed under primitive recursive set functions, we say that the syntax of $\mathcal{L}$ is *represented on* A. The idea is, roughly, if A is primitive recursively closed, it has "enough" functions to code the syntax of $\mathcal{L}$.[28] Finally, we assume that the logic $\mathcal{L}$ is strong enough to fix each element of A, i.e. $\mathrm{Mod}(\pi_a(x)) \in \mathrm{EC}_{\mathcal{L}[\tau_{\mathrm{set}}]}$ for $a \in A$. It's possible to impose further constraints on the syntax set, yet here those won't be necessary. As a matter of convention, we generally use A, A', A'', ... for the syntax sets of, respectively, $\mathcal{L}, \mathcal{L}', \mathcal{L}'', \ldots$.

[27] For a proof of Craig's theorem for $\mathcal{L}_{\omega\omega}$, cf. Bell and Slomson (1974, p. 153-7).

[28] For a definition of primitive recursive set functions, cf. e.g.Simpson (1978) and Jensen and Karp (1971); here, however, since we're dealing with $\mathcal{L}_{\omega\omega}$, we will be mostly dealing with ordinary primitive recursive functions, which are enough to code its syntax.

Definition A.13. A logic $\mathcal{L}$ is said to be *adequate to truth in* $\mathcal{L}'$ if for every τ there is $\tau^+ = [\tau, \tau_{\text{set}}, \text{Th}, \tau']$ and $\theta \in \mathcal{L}[\tau^+]$ such that for every $\mathfrak{M} \in \text{Str}[\tau]$, the following conditions hold:

(AT1) $(\mathfrak{M}, \mathfrak{A}', \text{Th}_{\mathcal{L}'}(\mathfrak{M}), \mathfrak{N}) \models_{\mathcal{L}} \theta$ for some $\mathfrak{N}$;

(AT2) If $(\mathfrak{M}, \mathfrak{B}, T, \mathfrak{N}) \models_{\mathcal{L}} \theta \wedge \pi_\phi(b)$, then $b \in T$ iff $\mathfrak{M} \models_{\mathcal{L}'} \phi$, whatever $\phi \in A'$ and $b \in B$.

Some observations: first, the role of τ' is to provide us with auxiliary symbols necessary, for instance, to write the definition of satisfaction for $\mathcal{L}'$. Second, $\mathfrak{A}'$ in (AT1) is the syntax set of $\mathcal{L}'$, whereas $\mathfrak{B}$ is a set theoretical structure $\mathfrak{B} = (B, E)$ which is used to "pin down" the sentence ϕ and its subformulas regarded as set-theoretical objects, i.e. $\mathfrak{B} \models \pi_\phi(b)$ iff $b = i(\phi)$, following Definition A.11.[29] $\text{Th}_{\mathcal{L}'}(\mathfrak{M})$ is defined in the usual way as $\text{Th}_{\mathcal{L}'}(\mathfrak{M}) = \{\phi \in \mathcal{L}'[\tau_{\mathfrak{M}}] \mid \mathfrak{M} \models_{\mathcal{L}'} \phi\}$. Finally, $\mathfrak{N}$ will serve as a sort of "interpretation function", mapping sequences of elements from M into their α coordinate.

As for the symbol Th in τ^+, it is used in the definition of θ. In fact, as emphasized by Feferman (1974, p. 218), θ above will need to contain: (i) formalizations of the recursive equations needed to code much of the information above; the auxiliary symbols for those are mainly contained in τ'; (ii) elementary statements about sequences used in the satisfaction clauses; (iii) the satisfaction clauses for $\mathcal{L}'$ themselves; (iv) the definition of Th, the truth predicate, in terms of the satisfaction predicate. Here, the satisfaction predicate, $S(x, y)$, is such that $S(\phi, s)$ iff there is a sequence s from the model $\mathfrak{M}$ such that s satisfies ϕ. Thus, if η is the sentence containing all the information from (i)-(iii) above, we can make (iv) explicit by writing θ as:

$$\eta \wedge \forall x(\text{Th}(x) \leftrightarrow \exists s S(x, s)).$$

Thus, what (AT1) is saying is that θ is basically a sentence providing a satisfaction definition for $\mathcal{L}'$ in $\mathcal{L}$, whereas (AT2) is providing $\mathcal{L}'$ with a truth definition representable in $\mathcal{L}$.

Our next lemma will relate the above notions to the notion of Δ closure from the preceding section:

Lemma A.7. *If $\mathcal{L}$ is adequate to truth in $\mathcal{L}'$ and $A' \subseteq A$, then $\mathcal{L}' \leq \Delta(\mathcal{L})$.*

Proof of Lemma A.7. Suppose the hypothesis. Observe first that, as $A' \subseteq A$, π_ϕ is $\text{RPC}_{\mathcal{L}}$ definable. Let $\mathscr{K} \in \text{EC}_{\mathcal{L}'}$. By definition, this means that there is a $\phi \in \mathcal{L}'$ such that $\mathscr{K} = \{\mathfrak{M} \mid \mathfrak{M} \models \phi\}$. Consider now the sentence θ as in Definition A.13. By hypothesis, it follows from (AT2) that the following conditions are equivalent:

[29] Notice that the class $\text{Mod}(\pi_\phi)$ will generally not be an $\text{EC}_{\mathcal{L}}$ class, but merely a $\text{RPC}_{\mathcal{L}}$ class.

(a) $\mathfrak{M} \models_{\mathcal{L}'} \phi$;

(b) $(\mathfrak{M}, \mathfrak{B}, \mathfrak{N}) \models \theta \wedge \pi_\phi(b) \wedge \mathrm{Th}(b)$ for some $b \in B$, $\mathfrak{B}$, and $\mathfrak{N}$;

(c) $(\mathfrak{M}, \mathfrak{B}, \mathfrak{N} \models \theta \wedge \pi_\phi(b) \rightarrow \mathrm{Th}(b)$ for all $b \in B$, $\mathfrak{M}$, and $\mathfrak{N}$.

Thus, by substituting the $\mathrm{RPC}_\mathcal{L}$ definition of π_ϕ in (b) and (c), we obtain a RPC-definition of $\mathrm{Mod}(\phi)$, that is, $\mathscr{K}$ is RPC in $\mathcal{L}$. Since $\mathscr{K}$ was arbitrary, it follows that all $\mathrm{EC}_{\mathcal{L}'} \subseteq \mathrm{RPC}_\mathcal{L}$; since the complement of an elementary class is also an elementary class, it follows that every pair $\mathscr{K}, \bar{\mathscr{K}}$ is $\mathrm{RPC}_\mathcal{L}$, whence they are all Δ in $\mathcal{L}$. Therefore, they are elementary classes in $\Delta(\mathcal{L})$.[30] ■

Corollary A.1. *If $\mathcal{L}$ is adequate to truth in $\mathcal{L}'$ and $A' \subseteq A$, then $\mathcal{L}' \leq_{\mathrm{RPC}} \mathcal{L}$.*

Proof. Suppose the hypothesis. By Lemma A.7, $\mathcal{L}' \leq \Delta(L)$. Thus, by Lemma A.5, $\mathcal{L}' \leq_{\mathrm{RPC}}$.■

Some further notions

In this section, we will prove two more lemmas that are necessary in order to demonstrate our main theorem. In order to state the lemmas, one more definition is necessary.

Definition A.14. A logic $\mathcal{L}$ is said to *capture* a set-theoretical predicate R if there is an $\mathrm{RPC}_\mathcal{L}$-class $\mathscr{K}$ of set-theoretical structures such that:

(C1) For any set a there is a transitive set M such that $a \in M$, and $(M, \in |_M) \in \mathscr{K}$;

(C2) If $\mathfrak{M} \in \mathscr{K}$ and $\mathfrak{M} \models \pi_{a_i}(m_i)(i = 1, \ldots, n)$, then $R(a_1, \ldots, a_n)$ if and only if $\mathfrak{M} \models R(m_1, \ldots, m_n)$.

In the above, $\pi_{a_i}(m_i)$ is defined as in Definition A.11.

We can now state and prove the lemma:

Lemma A.8. *If R is a predicate which is Δ_1 in* KPU-Inf*, then $\mathcal{L}_{\omega\omega}$ captures R.*

Proof of Lemma A.8. Let $\mathscr{L}$ be the $\mathrm{EC}_{\mathcal{L}_{\omega,\omega}}$-class of a large enough finite fragment of KPU-Inf. Now, by Theorem A.4, for each κ, $H_\kappa \in \mathscr{K}$, thus (C1) above is satisfied.

In order to see that (C2) holds, suppose $\mathfrak{M} \in \mathscr{K}$ and $\mathfrak{M} \models \pi_{a_i}(m_i)(i = 1, \ldots, n)$ and let $R(x_1, \ldots, x_n)$ be a Δ_1 predicate in KPU-Inf. Let $\mathfrak{N}$ be the well-founded part of $\mathfrak{M}$; by Lemma A.3, $\mathfrak{N} \in \mathscr{K}$ as well, so consider its transitive collapse, $\mathfrak{N}'$. Suppose that $R(a_1, \ldots, a_n)$. Since R is a Δ_1 predicate, by Theorem A.2 it's absolute, so $\mathfrak{N}' \models R(a_1, \ldots, a_n)$. But $\mathfrak{N}' \simeq \mathfrak{N}$, whence $\mathfrak{N} \models R(m_1, \ldots, m_n)$. Since R is absolute, it follows that $\mathfrak{M} \models$

[30] The preceding proof is based on the one in Väänänen (1985, p. 604).

$R(m_1, \ldots, m_n)$. Conversely, if $\mathfrak{M} \models R(m_1, \ldots, m_n)$, then, by the absoluteness of R, $\mathfrak{N} \models R(m_1, \ldots, m_n)$. Using again the fact that $\mathfrak{N}' \simeq \mathfrak{N}$, it follows that $\mathfrak{N}' \models R(a_1, \ldots, a_n)$ so, by the absoluteness of R, $R(a_1, \ldots, a_n)$.[31] ■

Now comes our main lemma, whose proof is a bit more involved:

Lemma A.9. *Let $\mathcal{L}$ and $\mathcal{L}'$ be arbitrary logics and suppose $\mathcal{L}$ captures the predicate $S(x, y)$ such that*

$$S(\mathfrak{M}, \phi) \text{ if and only if } \phi \in \mathcal{L}' \text{ and } \mathfrak{M} \models_{\mathcal{L}'} \phi,$$

for all ϕ and all $\mathfrak{M}$. Then $\mathcal{L}$ is adequate to truth in $\mathcal{L}'$.

Proof of Lemma A.9. By the hypothesis, there's a $\mathrm{RPC}_{\mathcal{L}}$ class $\mathscr{K}$ such that:

(C1) For any set a there's a transitive set M such that $a \in M$ and $\langle M, \in | M\rangle \in \mathscr{K}$;

(C2) If $\mathfrak{C} \in \mathscr{K}$ and $\mathfrak{C} \models \pi_{a_i}(m_i)$ $(i = 1, 2)$, then $S(a_1, a_2)$ iff $\mathfrak{C} \models S(m_1, m_2)$.

Our goal is to show that, given these conditions, for every τ there is a $\tau^+ = [\tau, \tau_{\mathrm{set}}, \mathrm{Th}, \tau']$ and $\theta \in \mathcal{L}[\tau^+]$ such that for every $\mathfrak{M} \in \mathrm{Str}[\tau]$, both (AT1) and (AT2) from Definition A.13 hold.[32]

So let τ'_{set} be the vocabulary of $\mathscr{K}$, disjoint from τ_{set}, and consider an arbitrary type τ. In order to simplify the proof, we will assume throughout that τ is one-sorted and contains only one binary predicate R, yet nothing hinges on this (the proof is obviously generalizable to the more complex case). Define $\tau^+ = [\tau, \tau_{\mathrm{set}}, T, \tau']$, with τ' containing τ'_{set}, plus three constant symbols, $m, n,$ and r of the sort of τ'_{set}, and whatever additional vocabulary we may need. Let $S'(x, y)$ be the predicate $S(x, y)$ in the language of τ'_{set} (that there is such a predicate is ensured by (C2)).

Consider now the structures $\mathfrak{M} \in \mathrm{Str}[\tau]$ and define $\mathscr{K}'$ as the class of τ^+-structures $\mathfrak{M}' = [\mathfrak{M}, \mathfrak{B}, T, \mathfrak{N}, m, n, r, f]$ obeying the following strictures:

(a) $\mathfrak{N} \in \mathscr{K}$;

(b) $\mathfrak{B} \subseteq_{\mathrm{end}} \mathfrak{N}$;

(c) $\mathfrak{N} \models$ "m is a structure (n, r) of type $\langle 2\rangle$ and n is a set of urelements";

(d) $\forall x(x \in M \leftrightarrow \mathfrak{N} \models f(x) \in n)$ and f is a bijection between M and n;

(e) $\forall x, y \in M(R(x, y)) \leftrightarrow \mathfrak{N} \models \langle f(x), f(y)\rangle \in r$;

[31] The preceding proof is based on the one in Väänänen (1985, p. 620).

[32] What follows is essentially the proof found in Väänänen (1985, p. 620-1).

(f) $\forall x \in B(T(x) \leftrightarrow S'(m, x))$;

As Väänänen puts it, the idea behind $\mathscr{K}'$ is this: the first condition ensures that $\mathfrak{N}$ is the set-theoretical universe in which $S(x, y)$ is captured, whereas the second marks $\mathfrak{B}$ off as the syntax set for $\mathcal{L}$. Inside $\mathfrak{N}$, there is a structure $m = (n, r)$ which, by conditions, (d) and (e), is exactly like $\mathfrak{M}$, whose true sentences we're trying to define. That n (and thus M) is a set of urelements helps to "filter out" any extra-elements that could come into the the picture when taking transitive closures, something that, although not necessary in general, is actually crucial in the case of $\mathcal{L}_{\omega\omega}$.[33] As for condition (f), it defines the truth-predicate T in a natural way.

It's clear that the class $\mathscr{K}'$ consists exactly of the projections to τ^+ of the structures satisfying θ (as defined in Definition A.13), being thus $\mathrm{RPC}_{\mathcal{L}}$. Suppose now $\mathfrak{M} \in \mathrm{Str}[\tau]$. By (C1), there is a transitive set N such that $A', \mathfrak{M} \in N$ and $\mathfrak{N} = (N, \in | N) \in \mathscr{K}$. Define $\mathfrak{M}' = [\mathfrak{M}, \mathfrak{A}, \mathrm{Th}_{\mathcal{L}'}(\mathfrak{M}), \mathfrak{N}, n, m, r, f]$ in such a way to make conditions (a)-(e) true. By (C2), it follows that $\mathfrak{N} \models S'(\mathfrak{M}, \phi) \iff S(\mathfrak{M}, \phi) \iff \phi \in \mathrm{Th}_{\mathcal{L}'}(\mathfrak{M}) \iff T(\phi)$, thus satisfying condition (f). Therefore, $\mathfrak{M}' \in \mathscr{K}'$, whence, by the preceding, it expands to a model of θ. Thus, (AT1) holds.

Let now $\mathfrak{M}'$ be such that $\mathfrak{M}' \models \theta \wedge \pi_\phi(b)$ for $\phi \in A'$ and $b \in B$. Consider the well-founded part $\mathfrak{N}'$ of $\mathfrak{N}$, with $i : \mathfrak{N}' \to (N, \in)$ its transitive collapse. By hypothesis, $\mathfrak{B} \models \pi_\phi(b)$, so, by the definition of π_ϕ, $b \in N'$ and $i(b) = \phi$. As n is a set of urelements, $m = (n, r) \in N'$ (because the collapsing function fixes the urelements by definition; cf. Barwise (1975, p. 30)), whence $i(m)$ is a structure $\mathfrak{m}$ isomorphic to $\mathfrak{M}$. We are thus able to prove the following equivalences:

$$b \in T \iff \mathfrak{N} \models S'(m, b); \tag{4.1}$$

$$\iff S(\mathfrak{m}, \phi); \tag{4.2}$$

$$\iff \phi \in \mathcal{L}' \text{ and } \mathfrak{m} \models_{\mathcal{L}'} \phi; \tag{4.3}$$

$$\iff \phi \in \mathcal{L}' \text{ and } \mathfrak{M} \models_{\mathcal{L}'} \phi. \tag{4.4}$$

(1) is justified by condition (f) above; (2) and (4) by the isomorphism condition outlined in the last paragraph; (3) by the definition of S.

Thus, (AT2) also holds. ■

A.3 Proof of the Main Theorem

In this section, I'll prove Theorem A.1, as stated in the Introduction. We begin with a more rigorous definition of a logic being absolute relative to a theory T:

[33] I thank Väänänen for clarifying this point to me through written communication.

Definition A.15. Let $\mathcal{L}$ be a logic and T a set theory, generally an extension of KPU. We say that $\mathcal{L}$ is *absolute relative to* T iff there is a Δ_1 predicate $S(x, y)$ such that for any $\phi \in A$ and for any $\mathfrak{M}$,

$$S(\mathfrak{M}, \phi) \iff \phi \in \mathcal{L} \text{ and } \mathfrak{M} \models \phi.$$

We also require that the syntactic operations be Δ_1 with respect to T.

The proof of the theorem is basically the one found in Väänänen (1985) and can be divided into two parts. First, one shows that $\mathcal{L}_{\omega\omega}$ does have the required properties, i.e. it is represented in HF and is absolute relative to KPU-Inf. Secondly, one shows that if $\mathcal{L}'$ also has these properties, then $\mathcal{L}' \leq \mathcal{L}_{\omega\omega}$. Here, however, I'll focus only on the more difficult second part; for a summary of how to proceed with the first part, the reader is directed to Barwise (1975, p. 78-83).[34] The result is actually incredibly simple given the work done in the previous sections.

Proof. Suppose the hypothesis. By hypothesis, the predicate $S(x, y)$, which defines the satisfaction relation in $\mathcal{L}$, is Δ_1 in KPU-Inf, whence, by Lemma A.8, $\mathcal{L}_{\omega\omega}$ captures this predicate. Thus, by Lemma A.9, $\mathcal{L}_{\omega\omega}$ is adequate to truth in $\mathcal{L}$. By Lemma A.7 and Corollary A.1, it follows that $\mathcal{L} \leq_{RPC} \mathcal{L}_{\omega\omega}$, so, by Lemma A.5, $\mathcal{L} \leq \Delta(\mathcal{L}_{\omega\omega})$. Therefore, by Lemma A.6, $\mathcal{L} \leq \mathcal{L}_{\omega\omega}$. ■

B Appendix: Feferman's Proposal

In his 1999 paper, "Logic, Logics, and Logicism" (Feferman 1999), Feferman proposed an interesting modification of Tarski's criterion. Instead of considering only those notions which are invariant under all *permutations* of the domain of individuals, Feferman proposes to consider the notions which are invariant under all *similarity relations* between domains of individuals. The main goal of this chapter is to explain and evaluate this proposal. Thus, in the first section, I present the formalism underlying the proposal; since, as shown by Casanovas (2007), there are some subtle questions here in the choice of the formalism, I will be very detailed in its presentation. In the next appendix, I will also present Casanovas's analysis of this proposal.

B.1 Preliminary remarks and definitions

Definition B.1. I will use TS for the set of types. This set is defined recursively as follows:

[34] That the syntax of $\mathcal{L}_{\omega\omega}$ is represented on HF is a simple result, once we note that ω is primitive recursively closed. One only needs then to prove that the satisfaction relation for $\mathcal{L}_{\omega\omega}$ is absolute, which is a tedious, but not too complicated result.

1. $0, b \in \mathrm{TS}$ (where 0 is the type of the individuals and b is the type of the truth values);
2. For any natural number n, if $\tau_1, \ldots, \tau_n, \sigma \in \mathrm{TS}$, then $\langle \tau_1, \ldots, \tau_n \to \sigma \rangle \in \mathrm{TS}$ (intuitively, the new type is for an n-ary function from types $\tau_1, \ldots, \tau_n$ to σ).

Definition B.2. Given now a non-empty set D, for each $\tau \in \mathrm{TS}$, we associate a domain D_τ as follows:

1. $D_0 = D$ and $D_b = \{T, F\}$;
2. $D_{\langle \tau_1, \ldots, \tau_n \to \sigma \rangle} = {}^{D_{\tau_1} \times \cdots \times D_{\tau_n}} D_\sigma$ (where ${}^A B$ is the set of all functions from A to B).[35]

Definition B.3. A functional finite type structure over D is defined as $\mathcal{D} = \langle D, D_t \langle D_\tau \rangle_{\tau \in \mathrm{TS}} \rangle$.

Observe that, aside from the basic types, all other types consist of functions. Therefore, if we want to consider relations in such a functional finite type structure, we need to identify them with their characteristic functions, e.g., the identity relation over D_0, $I = \{\langle x, x \rangle \mid x \in D_0\}$ is identified with the function f_I of type $\langle 0, 0 \to t \rangle$ such that $f(x, y) = T$ iff $x = y$.

Since Feferman makes an essential use of the lambda calculus in the proof of his main theorem, it will be convenient to supply a definition of this language as well.

Definition B.4. The *terms* of the language are defined by the following clauses:

1. For each $\tau \in \mathrm{TS}$, $x^n_\tau (n \in \omega)$ is a term (in practice, in order not to overload notation, I will drop the superscript and use variables x, y, z instead);
2. If s, t are terms of types $\langle \tau \to \sigma \rangle$ and τ, respectively, then $s(t)$ is a term of type σ;
3. If s is a term of type σ and x_τ a variable of type τ, then $\lambda x_\tau(s)$ is a term of type $\tau \to \sigma$.

The denotation of terms is defined by variable assignments in the usual way:

Definition B.5. Let α be a function from the set of variables to $\mathcal{D}$. The denotation of each term under α, $[[t]]^{\mathcal{D}}[\alpha]$, is defined as follows:

1. If x_τ is a variable of type τ, then $[[x_\tau]]^{\mathcal{D}}[\alpha] = \alpha(x_\tau)$;
2. If s, t are terms of types $\langle \tau \to \sigma \rangle, \tau$, then $[[s(t)]]^{\mathcal{D}}[\alpha] = [[s]]^{\mathcal{D}}[\alpha]([[t]]^{\mathcal{D}}[\alpha])$;
3. If s is a term of type σ and x_τ is a variable of τ, then $[[\lambda x_\tau(s)]]^{\mathcal{D}}[\alpha]$ is a function $f \in^{D_\tau} D_\sigma$ such that, for any $d \in D_\tau$, $f(d) = [[s]]^{\mathcal{D}}[\alpha^{x_\tau}_d]$, where $\alpha^{x_\tau}_d$ is a variable assignment such that $\alpha^{x_\tau}_d(x_\tau) = d$ and is otherwise the same as α.

[35] As Feferman indicates, it's not necessary, at each level, to take *all* such functions: it's possible to take only, e.g., the recursive functions—cf. Mitchell (1990, p. 371-2) for discussion. When we do take all functions, the hierarchy thus generated is called *maximal*. I'll follow Feferman in considering only maximal hierarchies.

Sometimes I will employ constants of a given type, whose denotation will be the obvious one. I'll also write image of the variables under α instead of α in $[\alpha]$ if that's more convenient.

We saw in the main body of the chapter how one of Feferman's main criticisms of Tarski's proposal was that it didn't allow for cross-domain operations. Thus, it's not very surprising that his framework is designed to allow for a precise definition of such operations. I'll quote his definition *verbatim*:

Definition B.6. An operation O is of *type τ across domains* if, for each functional type structure $\mathcal{D}$, we have an associated $O^{\mathcal{D}} \in D_\tau$.

In other words, an operation of type τ across domains is a function that takes as arguments functional type structures and gives as values objects of type τ. Next, let's define definability:

Definition B.7. An operation O is said to be *definable from operations $O_1, \ldots, O_k$* if it is given by a definition from them uniformly over each $\mathcal{D}$, i.e. if there's a term $t(x_1, \ldots, x_n)$ such that each x_i is of the same type as $O_i (i \leq k)$, t is of the same type as O, and, in each $\mathcal{D}$, $O^{\mathcal{D}} = [[t(x_1, \ldots, x_n)]]^{\mathcal{D}}[O_1^{\mathcal{D}}, \ldots, O_k^{\mathcal{D}}]$.

Finally, I'll also consider operations determined by formulas. As I will be concerned mostly with the first-order predicate calculus without identity, I'll restrict my definition to formulas from that language.

Definition B.8. A formula ϕ of the first-order predicate calculus without identity is said to *determine* an operation O if, whenever ϕ contains exactly n predicate variables $P_1, \ldots, P_n$ and exactly m free variables $x_1, \ldots, x_m$, then $(\mathcal{D}, P_1, \ldots, P_n) \models \phi[a_1, \ldots, a_n]$ iff $O^{\mathcal{D}}(p_1, \ldots, p_n, a_1, \ldots, a_m) = 1$, where $p_1, \ldots, p_n$ are the characteristic functions of $P_1, \ldots, P_n$.

B.2 The main theorem

In the previous chapter, I analyzed how Tarski's proposal depended heavily on his strong metaphysical assumptions. I also raised the question of what would happen if we chose a larger class of transformations in the formulation of our proposal. Here, I'll examine Feferman's proposal and main result, namely, that we take as the class of transformations the class of all surjections between domains of the functional type structures defined above. Let's state this more precisely.

Definition B.9. Let $D = \langle D_0, D_t\langle D_\tau\rangle_{\tau\in \mathrm{TS}}\rangle$ and $D' = \langle D'_0, D_t\langle D'_\tau\rangle_{\tau\in \mathrm{TS}}\rangle$. By a *similarity relation* $\sim$ between D and D' we mean a collection of relations $\sim_\tau$ for each $\tau \in \mathrm{TS}$ such that, if $\tau = \langle \tau_1, \ldots, \tau_n \to \sigma\rangle$, then:

1. $\forall x \in D_0 \exists x' \in D'_0 (x \sim_0 x') \wedge \forall x' \in D'_0 \exists x \in D_0 (x \sim_0 x')$;

2. $\forall x \in D_b \forall x' \in D'_b (x \sim_b x' \iff x = x')$;

3. For each $\tau = \langle \tau_1, \ldots, \tau_n \to \sigma \rangle$ and $p \in D_\tau$ and $p' \in D'_\tau$, we have:

$$p \sim_\tau p' \iff \forall \bar{x} \in D_\tau \bar{x'} D'_\tau ((\bar{x} \sim_\tau \bar{x'} \to p(\bar{x}) \sim_\sigma p'(\bar{x'}))).$$

where $\bar{x}$ abbreviates $x_1, \ldots, x_n$ and $\bar{x} \sim_\tau \bar{x'}$ abbreviates $x_1 \sim_{\tau_1} x'_1 \wedge \cdots \wedge x_n \sim_{\tau_n} x'_n$.

Clause 1 above basically states that $\sim_0$ is total and surjective, clause 2 that the truth values are always invariant, and clause 3 gives a recursive definition of $\sim_\tau$ for the higher types.

Given a surjective function $h : D_0 \to D'_0$, it's possible to define a partial extension h_τ to the other types that satisfy the clauses for a similarity relation. Here's the definition:

Definition B.10. Given a surjective function $h : D_0 \to D'_0$, it's partial extension h_τ is defined recursively as:

1. $h_0 = h$;

2. $h_b(x) = x$ for $x \in D_b$;

3. If $\tau = \langle \tau_1, \ldots, \tau_n \to \sigma \rangle$ and every $h_{\tau_i} (i \leq n)$ is surjective, then we define h_τ as follows.

   (a) Its domain is the subset of D_τ consisting of all mappings p which satisfy the following two conditions: (i) $p(x_1, \ldots, x_n) \in \mathrm{dom}(h_\sigma)$ for all $x_i \in \mathrm{dom}(h_{\tau_i})(i \leq n)$, and (ii) for all $x_i, x'_i \in \mathrm{dom}(h_{\tau_i})(i \leq n)$, if $h_{\tau_i}(x_i) = h_{\tau_i}(x'_i)$, then $h_\sigma(p(x_1, \ldots, x_n)) = h_\sigma(p(x'_1, \ldots, x'_n))$.

   (b) For any $p \in D_\tau$ which satisfies the above conditions, we define $h_\tau(p)$ as the mapping from D'_τ such that, for all $x_i \in \mathrm{dom}(h_{\tau_i})(i \leq n)$,

$$h_\tau(p)(h_{\tau_1}(x_1), \ldots, h_{\tau_n}(x_n)) = h_\sigma(p(x_1, \ldots, x_n)).$$

On the other hand, given that every surjective function h will be total in D_0 and surjective over D'_0, it's also possible to use Clauses 2, 3 above to define directly the similarity relation induced by h, $\sim_h$, by setting $x_0 \sim x'_0$ iff $h(x_0) = x'_0$ and then extending the relation to the higher types using clauses 2 and 3. In fact, as the next theorem shows, these are really different ways of doing the same thing, so that every surjective function between base domains is actually a similarity relation:

Theorem B.1. *If $\tau = \langle \tau_1, \ldots, \tau_n \to \sigma \rangle$ is a functional type and $h : D_0 \to D'_0$ is a surjective function, then $\sim_{h_\tau} = h_\tau$.*

Proof. The proof is by induction on the complexity of τ. The base case, for 0 and b, is immediate. So let τ be given and suppose the hypothesis is true for all types of lower complexity than τ. I must show that, given $p \in D_\tau$ and $p' \in D'_\tau$, $p \sim_{h_\tau} p'$ iff $h_\tau(p) = p'$.

Suppose first that $p \sim_{h_\tau} p'$. I'll show (a) $p \in \mathrm{dom}(h_\tau)$ and (b) $h_\tau(p) = p'$. To see (a), consider first $x_i \in \mathrm{dom}(h_{\tau_i})(i \leq n)$. Let $h_{\tau_i}(x_i) = y_i(i \leq n)$ for some $y_i \in D_{\tau_i}$. By the induction hypothesis, $x_i \sim_{\tau_i} y_i$ for $i \leq n$, so, by Clause 3 of the definition of a similarity relation, $p(x_1, \ldots, x_n) \sim_\sigma p'(y_1, \ldots, y_n)$. Therefore, applying the induction hypothesis again, $h_\sigma(p(x_1, \ldots, x_n)) = p'(y_1, \ldots, y_n)$, so $p(x_1, \ldots, x_n) \in \mathrm{dom}(h_\sigma)$ for any $x_i \in \mathrm{dom}(h_{\tau_i})$. Suppose now $h_{\tau_i}(x_i) = h_{\tau_i}(x'_i) = y_i(i \leq n)$ for $x_i, x'_i \in \mathrm{dom}(h_{\tau_i})$ and $y_i \in D_{\tau_i}(i \leq n)$. By the induction hypothesis, $x_i \sim_{h_{\tau_i}} y_i$ and $x'_i \sim_{h_{\tau_i}} y_i(i \leq n)$. Therefore, by Clause 3, $p(x_1, \ldots, x_n) \sim_{h_\sigma} p'(y_1, \ldots, y_n)$ and $p(x'_1, \ldots, x'_n) = p'(y_1, \ldots, y_n)$. Applying the induction hypothesis again, this means that $h_\sigma(p(x_1, \ldots, x_n)) = p'(y_1, \ldots, y_n) = h_\sigma(p(x'_1, \ldots, x'_n))$, as required. So $p \in \mathrm{dom}(h_\tau)$. It remains to be seen that $h_\tau(p) = p'$.

This is easy. Let $x_1 \in D_{\tau_1}, \ldots, x_n \in D_{\tau_n}$ be given such that $h_{\tau_i}(x_i) = y_i(i \leq n)$. By the induction hypothesis, $x_i \sim_{h_{\tau_i}} y_i(i \leq n)$, so, by Clause 3 and the hypothesis that $p \sim_{h_\tau} p'$, $p(x_1, \ldots, x_n) \sim_{h_\sigma} p'(y_1, \ldots, y_n)$. But then, by the induction hypothesis again, $h_\sigma(p(x_1, \ldots, x_n)) = p'(h_{\tau_1}(x_1), \ldots, h_{\tau_n}(x_n))$. Therefore, by clause 3.b of Definition B.10, $h_\tau(p) = p'$.

Conversely, suppose $h_\tau(p) = p'$ and that $x_i \sim_{\tau_i} y_i(i \leq n)$. By the induction hypothesis, for each $i \leq n$, $h_{\tau_i}(x_i) = y_i$. Thus, by Definition B.10, $h_\sigma(p(x_1, \ldots, x_n)) = p'(y_1, \ldots, y_n)$, whence, by the induction hypothesis, $p(x_1, \ldots, x_n) \sim_{h_\sigma} p'(y_1, \ldots, y_n)$. Therefore, $p \sim_{h_\tau} p'$, as required. ■

I'll adopt for this section Feferman's terminology and call surjective functions between base domains and their extensions *homomorphisms*. Now comes the main definition of this chapter.

Definition B.11. An operation O of type τ across domains is said to be *similarity invariant* if, for each $\mathcal{D}, \mathcal{D}'$ and similarity relation $\sim$ between $\mathcal{D}$ and $\mathcal{D}'$, we have $O^{\mathcal{D}} \sim O^{\mathcal{D}'}$. It is said to be *homomorphism invariant* if we only require O to be invariant under similarity relations determined by homomorphisms.

Considering that homomorphisms are a subset of similarity relations, it could seem that homomorphism invariance is actually a weaker notion than similarity invariance. In reality, however, as shown by Casanovas (2007), every similarity relation can be decomposed into surjective mappings, so that homomorphism invariance and similarity invariance coincide. However, as the result is more naturally stated using Casanovas's terminology, I'll postpone its analysis for the section on his article.

Moving on, it's possible to show that some operations are *not* homomorphism invariant. In particular, almost none of the cardinality quantifiers and the equality function turn out to be homomorphism invariant.

Theorem B.2. *Let I be the operation of type $\langle 0, 0 \to b\rangle$ defined at each $\mathcal{D}$ by*

$$I^{\mathcal{D}}(x, y) = T \text{ if x=y, and } F \text{ otherwise.}$$

Then I is not homomorphism invariant.

Proof. Let $D_0 = \{a, b\}$ and $D_0' = \{c\}$ and consider the homomorphism $h : D_0 \to D_0'$ defined pointwise as $h(a) = h(b) = c$. Thus, $a \sim_0 c$ and $b \sim_0 c$. However, $I^{\mathcal{D}}(a, b) = F$, whereas $I^{\mathcal{D}'}(c, c) = T$, that is, $I^{\mathcal{D}}(a, b) \not\sim I^{\mathcal{D}}(c, c)$. Therefore, by clause 3, $I^{\mathcal{D}} \not\sim I^{\mathcal{D}'}$, whence I is not homomorphism invariant. ■

Theorem B.3. *For each non-zero cardinal κ, the quantifier E_κ of type $\langle\langle 0 \to b\rangle \to b\rangle$ is defined by*

$$E_\kappa^{\mathcal{D}}(p) = T \text{ if there are at least } \kappa \text{ distinct } x \text{ such that } p(x) = T, \text{ and } F \text{ otherwise.}$$

Then, for $\kappa \geq 2$, E_κ is not homomorphism invariant.

Proof. Let D_0 be any domain of cardinality κ, $D_0' = \{a\}$, and p be such that $E_\kappa^{\mathcal{D}}(p^{\mathcal{D}}) = T$ with $\kappa \geq 2$. Set $p^{\mathcal{D}'}(a) = T$ and consider the homomorphism h defined as, for every $x \in D_0$, $h(x) = a$. By definition, for every $x \in D_0$, $x \sim_0 a$ and, by hypothesis, $p^{\mathcal{D}}(x) \sim p^{\mathcal{D}'}(a)$. But $E_\kappa^{\mathcal{D}'}(p^{\mathcal{D}'}) = F$, so $E_\kappa^{\mathcal{D}} \not\sim E_\kappa^{\mathcal{D}'}$, that is, E_κ is not similarity invariant. ■

On the other hand, given invariant operations $O_1, \ldots, O_n$, it's possible to build further invariant operations.

Lemma B.1. *Let t be a term of the typed-lambda calculus, $\mathcal{D}, \mathcal{D}'$ two type hierarchies and $\sim \subseteq D_0 \times D_0'$ a similarity relation. Then, if $\alpha \sim \alpha'$, $[[t]]^{\mathcal{D}}[\alpha] \sim [[t]]^{\mathcal{D}'}[\alpha']$.*

Proof. By induction on the complexity of t. ■

Lemma B.2. *If an operation O is defined by operations $O_1, \ldots, O_n$, all of which are similarity invariant, then O itself is similarity invariant.*

Proof. Suppose the hypothesis, i.e. that O is defined by invariant operations $O_1, \ldots, O_n$. Let $\mathcal{D}, \mathcal{D}'$ be arbitrary and consider a similarity relation $\sim$. By the hypothesis, $O_i^{\mathcal{D}} \sim O_i^{\mathcal{D}'}$, so, by Lemma B.1, $O^{\mathcal{D}} = [[t]]^{\mathcal{D}}[O_1^{\mathcal{D}}, \ldots, O_n^{\mathcal{D}}] \sim [[t]]^{\mathcal{D}'}[O_1^{\mathcal{D}'}, \ldots, O_n^{\mathcal{D}'}] = O^{\mathcal{D}'}$, as required. ■

Definition B.12. Take π for the type $\langle 0 \to b\rangle$, i.e. for the type of monadic predicates. A type $\tau = \langle \tau_1, \ldots, \tau_n \to \sigma\rangle$ is said to be *monadic* if $\sigma = b$ and, for each $\tau_i (i \leq n)$, τ_i is either π, or b, or 0. τ is said to be *pure monadic* if it has the form $\langle \pi^n \to b\rangle$.

As noted by Feferman, the following monadic operations *are* homomorphism invariant:

Proposition B.1. *The operation N of negation of type $\langle b \to b\rangle$, defined by $N^{\mathcal{D}}(p) = F$ if $p = T$, T otherwise, is similarity invariant.*

Proposition B.2. *The operation C of conjunction of type $\langle b, b \to b\rangle$, defined by $C^{\mathcal{D}}(p, q) = T$ if $p = q = T$ and F otherwise, is similarity invariant.*

Proposition B.3. *The operation E of existential quantification over the domain of individuals of type $\langle\langle 0 \to b\rangle \to b\rangle$, defined by $E^{\mathcal{D}}(p) = T$ if there is an $x \in D_0$ such that $p(x) = T$, F otherwise, is similarity invariant.*

The first two propositions follow immediately from clause 2. As for the third, here's the proof:

Proof. Our goal is to show that, if p and p' are arbitrary operations of type $\langle 0 \to b\rangle$ over $\mathcal{D}$ and $\mathcal{D}'$, respectively, such that $p \sim p'$, then $E^{\mathcal{D}}(p) \sim E^{\mathcal{D}'}$. If we show this, then, by clause 3, $E^{\mathcal{D}} \sim E^{\mathcal{D}'}$.

So let p, p' be as desired. As $p \sim p'$, this means that, whenever $x \in D_0$ and $x' \in D'_0$ are such that $x \sim x'$, then $p(x) = p'(x')$. Suppose now $E^{\mathcal{D}}(p) = T$. Then, by definition, there is $x \in D_0$ such that $p(x) = T$. Applying the first similarity condition, there must be $x' \in D'_0$ such that $x \sim x'$, so, by the hypothesis, $p'(x') = T$, whence, by definition, $E^{\mathcal{D}'}(p') = T$. An analogous reason shows that, if $E^{\mathcal{D}'} = T$, then $E^{\mathcal{D}}(p) = T$. Therefore, $E^{\mathcal{D}}(p) = E^{\mathcal{D}'}(p')$ for any such p, which is what we wanted to prove. ■

The above three propositions are useful because of the following theorem:

Theorem B.4. *Every formula ϕ of the first-order predicate calculus without identity determines an operation O which is definable in the typed λ-calculus from N, C, E, and λ-abstraction, plus the characteristic functions of whatever predicates appear in ϕ.*

Proof. The proof is by induction on the complexity of ϕ. For the base case, suppose ϕ is atomic. Since the language doesn't contain equality, ϕ must be of the form $P(x_1, \ldots, x_m)$. Thus, ϕ determines $p(x_1, \ldots, x_m)$, where p is the characteristic function of P.

Assuming now the theorem is true for formulas of lesser complexity, if ϕ is of the form $\neg\psi$, then ψ must determine an operation O which satisfies the theorem. Thus, ϕ determines

an operation definable by $N(O)$. Similarly, if ϕ is $\psi \wedge \theta$, then ϕ determines an operation definable by $C(O, O')$, where O, O' are the operations determined by ψ, θ, respectively.

Finally, if ϕ is of the form $\exists x\psi$, then, if O is the operation defined by ψ, then ϕ determines an operation definable by $E(\lambda x O)$. ■

Here's an example from Feferman. Consider the formula $\forall x(P(x,z) \wedge \exists y Q(x,y,z))$. This determines an operation O of type $\langle\langle 0,0 \to b\rangle, \langle 0,0,0 \to b\rangle, 0 \to b\rangle$, i.e. for each domain $\mathcal{D}$ an operation of the form $O^{\mathcal{D}}(p,q,z)$, where p and q are the characteristic functions of P and Q, respectively. This is definable by:

$$O^{\mathcal{D}}(p,q,z) = N(E(\lambda x(C(p(x,z), N(E(\lambda y(q(x,y,z))))))))).$$

I come now to the main theorem:

Theorem B.5. *Let O be of monadic type and homomorphism invariant. Then O is definable by a formula of the first-order predicate calculus without equality.*

Proof. I'll first prove the pure monadic case, which is simpler. For convenience, I'll make a slight change of notation and take $D_b = \{0,1\}$ from now on, with 0 taking the place of F and 1 the place of T. The basic idea of the proof is to show that the operation O can be completely determined by considering its behavior only over finite structures and then using this information to construct a first-order formula corresponding to such behavior.

So let O be a pure monadic homomorphism invariant operation of type $\langle \pi^n \to b\rangle$. Consider a structure of the form $(\mathcal{D}, p_1, \ldots, p_n)$, where each $p_i (i \leq n)$ is of type π. Let $\bar{p}(y)$ abbreviate $p_1(y) \ldots p_n(y)$ and let $\bar{k}$ be an n-termed sequence of 0s and 1s, in such a way that $\bar{p}(y) = \bar{k}$ for some $\bar{k}$. Define on D_0 an equivalence relation as follows:

$$y \equiv z \iff \bar{p}(y) = \bar{p}(z)$$

.

Denote by $[y]$ the equivalence class of y under $\equiv$. Note that, since there are exactly 2^n n-termed sequences of 0s and 1s, there will be at most 2^n equivalence classes under $\equiv$. Thus, we can use these equivalence classes to construct the desired finite domain. Define then a new structure $(\mathcal{D}', \bar{p}')$ such that $D_0' = \{[y] \mid y \in D_0\}$ and such that $\bar{p}'([y]) = \bar{p}(y)$. Set $h : M_0 \to M_0'$ by $h(y) = [y]$. Further, set $\Delta(\mathcal{D}, \bar{p}) = \{\bar{k} \mid \exists y \bar{p}(y) = k\}$ (this works like a diagram of $\mathcal{D}$ under $\bar{p}$) and set $[\![\bar{k}]\!] = \{y \mid \bar{p}(y) = \bar{k}\}$. Note that:

Claim 1: $[\![\bar{k}]\!] \neq \varnothing \iff \bar{k} \in \Delta(\mathcal{D}, \bar{p})$.

Proof. Suppose $[\![\bar{k}]\!] \neq \varnothing$. Then there's $y \in [\![\bar{k}]\!]$, that is, by definition, $\bar{p}(y) = \bar{k}$, whence $\bar{k} \in \Delta(\mathcal{D}, \bar{p})$. On the other hand, if $\bar{k} \in \Delta(\mathcal{D}, \bar{p})$, by definition there's $y \in D_0$ such that $\bar{p}(y) = \bar{k}$, so $y \in [\![\bar{k}]\!]$, whence $[\![\bar{k}]\!] \neq \varnothing$.

Claim 2: if $[\![\bar{k}]\!] \neq \varnothing$, then $y \in [\![\bar{k}]\!] \iff [\![\bar{k}]\!] = [y]$.

Proof. Suppose the hypothesis and let $y \in [\![\bar{k}]\!]$, i.e. $\bar{p}(y) = \bar{k}$. I'll show that $[\![\bar{k}]\!] = [y]$. If $x \in [\![\bar{k}]\!]$, then $\bar{p}(x) = \bar{k} = \bar{p}(y)$, so $x \in [y]$, and if $x \in [y]$, then $\bar{p}(x) = \bar{p}(y) = \bar{k}$, so $x \in [\![\bar{k}]\!]$. On the other hand, if $[\![\bar{k}]\!] = [y]$, then, since $y \in [y]$, $y \in [\![\bar{k}]\!]$.

The above result shows that the equivalence classes are completely determined by the $\bar{k}$s, so that we can consider them instead of working directly with the elements from D_0.

Claim 3: If $\bar{k} \neq \bar{l}$, then $[\![\bar{k}]\!] \cap [\![\bar{l}]\!] = \varnothing$.

Proof. Suppose $y \in [\![\bar{k}]\!] \cap [\![\bar{l}]\!]$. Then $\bar{k} = \bar{p}(y) = \bar{l}$. The claim then follows by contraposition.

Therefore, there's a bijective function $f : \Delta(\mathcal{D}, \bar{p}) \to D_0'$ defined as $f[\bar{k}] = [\![\bar{k}]\!]$.

Claim 4: $O^{\mathcal{D}}(\bar{p}) = O^{\mathcal{D}'}(\bar{p}')$.

Proof. Suppose towards a contradiction that $O^{\mathcal{D}}(\bar{p}) \neq O^{\mathcal{D}'}(\bar{p}')$. Let h be as defined above, i.e. $h(y) = [y]$. Clearly h is surjective, so it determines a similarity relation induced by $y \sim_0 [y]$. Since, by construction, $\bar{p}(y) = \bar{p}'([y])$, it follows that $\bar{p} \sim \bar{p}'$. But then, by the hypothesis, $O^{\mathcal{D}} \not\sim O^{\mathcal{D}'}$, contradicting the fact that O is similarity invariant.

Let now $(\mathcal{D}^*, \bar{p}^*)$ be such that $\Delta(\mathcal{D}^*, \bar{p}^*) = \Delta(\mathcal{D}, \bar{p})$. Define $(\mathcal{D}^{*'}, \bar{p}^{*'})$ in the same way as above.

Claim 5: $(\mathcal{D}', \bar{p}') \simeq (\mathcal{D}^{*'}, \bar{p}^{*'})$.

Proof. By the above, there are functions $f : \Delta(\mathcal{D}, \bar{p}) \to D_0'$ and $g : \Delta(\mathcal{D}^*, \bar{p}^*) \to D_0^{*'}$ such that both f, g are bijections. As $\Delta(\mathcal{D}^*, \bar{p}^*) = \Delta(\mathcal{D}, \bar{p})$, f and g have the same domain, whence we can take the composition $\theta = g \circ f^{-1}$, which will be a function $\theta : D_0' \to D_0^{*'}$. I claim θ is the desired isomorphism. It's clearly a bijection, since the composition of bijections is still a bijection. Further, if $\bar{p}'([y]) = \bar{k}$, then $f[\bar{k}] = [y]$, so, since $f^{-1}([y]) = \bar{k}$ and $g(\bar{k}) = [y^*]$, it follows that $\bar{p}'([y]) = \bar{k}$ iff $\bar{p}^{*'}(\theta[y]) = \bar{k}$, which concludes the proof.

Hence, $O^{\mathcal{D}}(\bar{p}) = O^{\mathcal{D}'}(\bar{p}') = O^{\mathcal{D}^{*'}}(\bar{p}^{*'}) = O^{\mathcal{D}^*}(\bar{p}^*)$, so that the behavior of O at a structure is completely determined by its diagram. Define now $O^+ = \{\Delta(\mathcal{D}, \bar{p}) \mid O^{\mathcal{D}}(\bar{p}) = 1\}$ and $O^- = \{\Delta(\mathcal{D}, \bar{p}) \mid O^{\mathcal{D}}(\bar{p}) = 0\}$. Since each $\Delta(\mathcal{D}, \bar{P})$ is a subset of ${}^n2$, and $|{}^n2| = 2^n$, and $|\mathcal{P}({}^n2)| = 2^{2^n}$, it follows that there are at max 2^{2^n} such Δs, whence each O^+ and O^- are finite. Enumerate them each as $O^+ = \{\Delta_1, \ldots, \Delta_r\}$ and $O^- = \{\Delta_{r+1}, \ldots, \Delta_s\}$, setting $O^* = O^+ \cup O^-$. It follows that:

$$O^{\mathcal{D}}(\bar{p}) = \begin{cases} 1 \text{ if } \Delta(\mathcal{D}, \bar{p}) = \Delta_i \text{ for } 1 \leq i \leq r \\ 0 \text{ if } \Delta(\mathcal{D}, \bar{p}) = \Delta_i \text{ for } r+1 \leq i \leq s. \end{cases}$$

But, for a given Δ_i, $\Delta(\mathcal{D}, \bar{p}) = \Delta_i$ is definable as follows. Enumerate each $\bar{k} \in^n 2$ as $\bar{k}_1, \ldots, \bar{k}_m$. For each such $\bar{k}_j$, set

$$\phi_{jl} := \begin{cases} P_l(y) \text{ if } 1 = l \in \bar{k}_j \\ \neg P_l(y) \text{ if } 0 = l \in \bar{k}_j. \end{cases}$$

Then, for each $\bar{k}_j$, let ψ_j be defined as:

$$\psi_j := \begin{cases} \exists \bar{y} \bigwedge\limits_{l \leq n} \phi_{jl} \text{ if } \bar{k}_j \in \Delta_i \\ \neg \exists \bar{y} \bigwedge\limits_{l \leq n} \phi_{jl} \text{ if } \bar{k}_j \notin \Delta_i \end{cases}$$

Finally, set $\psi^{\Delta_i} := \bigwedge\limits_{j \leq m} \psi_j$, so that $\psi^{\Delta_i} = 1$ iff $\Delta_i = \Delta(\mathcal{D}, \bar{p})$ for some $\mathcal{D}$ and $O^{\mathcal{D}}(\bar{p}) = 1$. Define $\theta := \bigvee\limits_{\Delta_i \in O^*} \psi^{\Delta_i}$. Then θ completely describes the operation in question. ■

C Appendix: Casanovas's Analysis of Feferman's Proposal

As mentioned in the previous appendix, Casanovas (2007) provides a detailed analysis of Feferman's theorem. Specifically, building on the work of Casanovas et al. (1996), he provides suitable definitions for similarity invariance in a relational type setting, instead of Feferman's functional type setting. Strikingly, this change of setting results in a remarkably different analysis: whereas Feferman is able to prove that the operations determined by first-order formulas in a language without equality are similarity invariant, Casanovas shows that, in this new setting, the operations determined by negation, conjunction, and universal quantification are not similarity invariant. Since these results are very surprising, I'll focus in this section on Casanovas's analysis of the main differences between his and Feferman's approaches.

C.1 Types of similarity

In this section, I want to analyze closely the different types of invariance proposed by both Feferman and Casanovas, as well as state a few results concerning their relations. For completeness sake, I'll restate here the definitions from the Appendix to the first chapter regarding finite relational type structures.

Definition C.1. Define a hierarchy of types in the following way, using TS as the set of types:

1. $0 \in \text{TS}$ (we take 0 as the type of individuals);

2. For any natural number n, if $\tau_1, \ldots, \tau_n \in \mathrm{TS}$, then $\langle \tau_1, \ldots, \tau_n \rangle \in \mathrm{TS}$ (the new type is for n-ary relations among types $\tau_1, \ldots, \tau_n$).

Definition C.2. Given now a non-empty set D, we associate, for each type τ, a domain D_τ in the following way:

1. $D_0 = D$;
2. $D_{\langle \tau_1, \ldots, \tau_n \rangle} = \mathcal{P}(D_{\tau_1} \times \cdots \times D_{\tau_n})$.

In the same way that permutations of the base domain can be extended to all types, so do mappings between domains:

Definition C.3. Let $f : D \to E$ be a mapping between base domains D, E and let τ be a relational type. The *induced mapping* f_τ can be defined recursively as follows:

1. $f_0 = f$;
2. $f_\tau : D_\tau \to E_\tau$, with $\tau = \langle \tau_1, \ldots, \tau_n \rangle$ is defined as:

$$f_\tau(a) = \{\langle f_{\tau_1}(a_1), \ldots, f_{\tau_n}(a_n) \rangle \mid \langle a_1, \ldots, a_n \rangle \in a\}.$$

Given the differences between a functional type structure and a relational type structure, we need to adapt the definition of similarity relation to this new context. One way of doing it is by straightforwardly adapting Clauses 1 and 3 of Feferman's definition (Cluase 2 is omitted, as there is no boolean type in the relational context).

Definition C.4. A relation π between D_0 and D'_0 is a *similarity relation* iff for every $a \in D_0$ there is $b \in D'_0$ such that $\pi(a, b)$ and for every $b \in D'_0$ there is $a \in D_0$ such that $\pi(a, b)$. In other words, $\mathrm{dom}(\pi) = D_0$ and $\mathrm{rng}(\pi) = D'_0$.

The above is the natural adaptation of Clause 1. Clause 3 is adapted in the following way:

Definition C.5. Let π be a similarity relation between base domains D, E and let τ be a relational type. The *induced similarity relation* π_τ is defined recursively as follows:

1. $\pi_0 = \pi$;
2. If $\tau = \langle \tau_1, \ldots, \tau_n \rangle$, then the similarity π_τ between D_τ and E_τ is given, for $a \in D_\tau$ and $b \in E_\tau$, by: $\pi_\tau(a, b)$ iff:
   (a) for each $\langle a_1, \ldots, a_n \rangle \in a$, there's $\langle b_1, \ldots, b_n \rangle \in b$ such that $\pi_{\tau_i}(a_i, b_i)(i \leq n)$ and

(b) for each $\langle b_1, \ldots, b_n \rangle \in b$, there's $\langle a_1, \ldots, a_n \rangle \in a$ such that $\pi_{\tau_i}(a_i, b_i)(i \leq n)$.

That's not, however, the definitions provided by Casanovas. Instead of working directly with the similarity relations, he provides definitions by way of compositions of surjective mappings. Although seemingly more complicated, Casanovas's definition actually helps to simplify the proof of a few theorems, as we will see. Here's his analogue to Clause 1:

Definition C.6. A relation $\pi \subseteq D \times E$ is a *similarity relation* between D and E iff for some $n \geq 2$, there are sets $D_1, \ldots, D_n$ and mappings $f_1, \ldots, f_{n-1}$ such that $D_1 = D$, $D_n = E$, and for every $i = 1, \ldots, n$, f_i is a mapping from D_i onto D_{i+1} or it is a mapping from D_{i+1} onto D_i, and π is the relational composition $R_1 \circ \cdots \circ R_{n-1}$ where $R_i = f_i$ if f_i is from D_i onto D_{i+1}, and $R_i = f_i^{-1}$ if it is from D_{i+1} onto D_i.

And here's his analogue to Clause 3:

Definition C.7. Given a similarity relation $\pi : D_0 \to D_0'$, there are, by Definition C.6, sets $D_1, \ldots, D_n$ and mappings $f_1, \ldots, f_{n-1}$, such that $D_1 = D_0$, $D_n = D_0'$, and, for each f_i, f_i is either a mapping from D_i onto D_{i+1}, or from D_{i+1} onto D_i, and π is the relational composition $R_1 \circ \cdots \circ R_{n-1}$ where each $R_i = f_i$ if the former holds, or else $R_i = f_i^{-1}$ if the latter holds. We already know, by Definition C.3, how to extend each f_i to f_{i_τ}, so we can use this to define π_τ: let π_τ be the relational product $R_{1_\tau} \circ \cdots \circ R_{n-1_\tau}$, where each R_{i_τ} has the obvious definition.

It's not too difficult to show that these definitions are actually equivalent. The proof is by induction on the complexity of types. The next theorem will take care of the base case:

Theorem C.1. *A binary relation π between D_0 and D_0' is a similarity relation according to Definition C.4 iff it is a similarity relation according to Definition C.6.*

Proof. Suppose first that π is a similarity relation between D_0 and D_0' according to Definition C.4. I'll show that it is a similarity relation between D_0 and D_0' according to Definition C.6 by providing a decomposition of π into surjective mappings or inverses thereof.

First, define an equivalence relation $\equiv$ on D_0' as follows:

$$b \equiv b' \text{ iff } \pi(a, b) \text{ and } \pi(a, b') \text{ for some } a \in D_0.$$

Let $D_0/_\equiv$ be the set of all equivalence classes of D_0 by $\equiv$. Define $f_1 : D_0 \to D_0/_\equiv$ by setting $f_1(a) = [b]$ iff $\pi(a, b)$. By construction, this will be a function, and by the hypothesis, for every $b \in D_0'$ there's $a \in D_0'$ such that $\pi(a, b)$, so f_1 is surjective. Next, let f_2 be the canonical mapping from D_0' to $D_0'/_\equiv$. This mapping is obviously surjective. I claim that $\pi = f_2^{-1} \circ f_1$. To see this, suppose $\langle a, b \rangle \in \pi$. Then $f_1(a) = [b]$ and, since $\langle [b], b \rangle \in f_2^{-1}$, it

follows that $\langle a, b\rangle \in f_2^{-1} \circ f_1$. On the other hand, suppose $\langle a, b\rangle \in f_2^{-1} \circ f_1$. Then there's a $[c]$ such that $\langle [c], b\rangle \in f_2^{-1}$ and $f_1(a) = [c]$. But then, by the definition of f_2^{-1}, $[c] = [b]$, so $f_1(a) = [b]$, whence $\pi(a, b)$.

The proof in the other direction is by induction on the length n of sequence of sets $D_1, \ldots, D_n$ in Definition C.6.

The base case is $n = 2$. In that case, either $\pi = f_1$ or $\pi = f_1^{-1}$. If the former, then f_1 is a function from D_0 onto D_0', so $\mathrm{dom}(\pi) = D_0$ and $\mathrm{rng}(\pi) = D_0'$, as required. If the latter, then f_1 is a function from D_0' onto D_0, whence f_1^{-1} is a relation whose domain is D_0 (because f_1 is surjective) and whose range is D_0' (because f_1 is a function). Thus, either way π will be a total surjective relation, which is what we wanted to prove.

Suppose now the hypothesis is true for n and consider $n + 1$. By definition, $\pi = R_1 \circ \cdots \circ R_n$. Let $\pi' = R_1 \circ \cdots \circ R_{n-1}$. Notice that π' is also a similarity relation between $D_1 = D$ and D_n, so the induction hypothesis applies and $\mathrm{dom}(\pi') = D$ and $\mathrm{rng}(\pi') = D_n$. But then, as $\pi = \pi' \circ R_n$, it follows that $\mathrm{dom}(\pi) = \mathrm{dom}(\pi') = D$ and $\mathrm{rng}(\pi) = \mathrm{rng}(R_n) = D_0'$ (by the definition of R_n), as desired. ■

Next comes the induction step.

Theorem C.2. *Let π be a similarity relation between base domains D_0, D_0', π_τ be the extension for τ according to Definition C.5 and π_τ' be the extension for τ according to Definition C.7. Then $\pi_\tau = \pi_\tau'$.*

Proof. The induction hypothesis is that the theorem is true for all types of lower complexity than τ. Suppose first that $\pi_\tau'(a, b)$. I'll show that $\pi_\tau(a, b)$, whence $\pi_\tau' \subseteq \pi_\tau$. In order to do this, I'll show that a, b satisfy conditions (a) and (b) laid out in Definition C.5.

By definition, there are $f_1, \ldots, f_n$ such that, for each f_i and $d \in D_i$ or $d \in D_{i+1}$ (accordingly as $f_i : D_i \to D_{i+1}$ or $f_i : D_{i+1} \to D_i$), we have:

$$f_i(d) = \{\langle f_{i_{\tau_1}}(d_1), \ldots, f_{i_{\tau_n}}(d_n)\rangle \mid \langle d_1, \ldots, d_n\rangle \in d\}$$

.

Therefore, for each $\langle a_1, \ldots, a_n\rangle \in a$, there will be, for each a_j, a sequence of surjective mappings or inverse of surjective mappings taking a_j to b_j, whence $\pi_{\tau_j}'(a_j, b_j)$. By the induction hypothesis, $\pi_{\tau_j}' = \pi_{\tau_j}$, so $\pi_{\tau_j}(a_j, b_j)$. Hence, for each $\langle a_1, \ldots, a_n\rangle \in a$, there's a $\langle b_1, \ldots, b_n\rangle$ such that $\pi_{\tau_j}(a_j, b_j)$. A similar reasoning shows that, for each $\langle b_1, \ldots, b_n\rangle \in b$, there will be $\langle a_1, \ldots, a_n\rangle \in a$ such that $\pi_{\tau_j}(a_j, b_j)$. Thus, $\pi_\tau' \subseteq \pi_\tau$.

In the other direction, suppose $\pi_\tau(a, b)$. Again, this means that, for each $\langle a_1, \ldots, a_n\rangle \in a$, there's $\langle b_1, \ldots, b_n\rangle \in b$ such that $\pi_{\tau_j}(a_j, b_j)(j \leq n)$ and similarly with the roles of a and b reversed. Thus, by the induction hypothesis, for each a_j, b_j, there will be a sequence of

mappings $f_{i_{\tau_j}}$ and their inverses taking a_j to b_j. Taking the extension of each $f_{i_{\tau_j}}$ to f_{i_τ} will thus give a sequence of mappings and their inverses taking a to b, that is, $\pi'_\tau(a, b)$. ■

The fact that every similarity relation can be decomposed into a sequence of surjective mappings and their inverses is rather surprising, as it implies that invariance under surjective mappings and invariance under similarity relations actually coincide. In other words, as remarked in the previous section, Feferman's distinction between homomorphism invariance and similarity invariance actually collapses. The proof here was presented in a relational type setting, but its adaptation to the functional type setting used by Feferman is straightforward.[36] This fact will be useful later on.

C.2 Types of invariance

This section will be concerned with types of invariance. I'll first present Casanovas's definition of mapping-invariance and, afterwards, present the natural translation of Feferman's definition to a relational setting. I'll then present Casanovas's proof that it coincides with another type of invariance, what he calls *preimage-invariance*. The proof that these last type of invariance *does not* coincide with mapping-invariance will be postponed to the next section, when I'll present Casanovas's characterization of the mapping-invariant objects.

As in Feferman's case, we define an object a of type τ to be a function which associates, with every $\mathcal{D}$, a corresponding $a_\mathcal{D} \in D_\tau$. As remarked above, by Theorem C.2, mapping-invariance and similarity-invariance actually coincide. However, since both versions will be useful, I'll present the two definitions below.

Definition C.8. An object a of type τ is said to be *similarity invariant* if, for every $\mathcal{D}, \mathcal{D}'$ and every similarity relation π between D_0, D'_0, $\pi_\tau(a_\mathcal{D}, a_{\mathcal{D}'})$.

Definition C.9. An object a of type τ is said to be *mapping-invariant* iff for every D_τ, D'_τ and surjective mapping $f_\tau : D_\tau \to D'_\tau$, $f(a_\mathcal{D}) = a_{\mathcal{D}'}$.

Let $\tau_1, \ldots, \tau_n, \tau$ be relational types. A $\langle \tau_1, \ldots, \tau_n \to \tau \rangle$*-ary operator* is a function F such that, for any $\mathcal{D}$, gives a mapping

$$F_\mathcal{D} : D_{\tau_1} \times \cdots \times D_{\tau_n} \to D_\tau.$$

Definition C.10. A $\langle \tau_1, \ldots, \tau_n \to \tau \rangle$-ary operator F is *similarity invariant* if for every $\mathcal{D}, \mathcal{D}'$ and every similarity relation π between D_0, D'_0, for every $a_i \in D_{\tau_i} (i \leq n)$ and every $b_i \in D_{\tau_i} (i \leq n)$, if $\pi_{\tau_i}(a_i, b_i)(i \leq n)$, then:

$$\pi_\tau(F_\mathcal{D}(a_1, \ldots, a_n), F_{\mathcal{D}'}(b_1, \ldots, b_n)).$$

[36] For a somewhat different proof, cf. Casanovas (2007, section 7).

Definition C.11. A $\langle \tau_1, \ldots, \tau_n \to \tau \rangle$-ary operator F is *mapping-invariant* if, for all D_0, D_0' and surjective mapping $f : D_0 \to D_0'$, for all $a_1 \in D_{\tau_1}, \ldots, a_n \in D_{\tau_n}$, we have:

$$f_\tau(F_{\mathcal{D}}(a_1, \ldots, a_n)) = F_{\mathcal{D}'}(f_{\tau_1}(a_1), \ldots, f_{\tau_n}(a_n)).$$

Let now TS be a set of relational type symbols and $\mathcal{D}$ a relational type hierarchy. I'll present a way of translating every type $\tau \in \mathrm{TS}$ to a type τ^* in a functional type hierarchy and, similarly, every object $a \in D_\tau$ to an object $a^* \in D_{\tau^*}$. For the type symbols, define * recursively as follows:

1. $0^* = 0$;
2. $\langle \tau_1, \ldots, \tau_n \rangle^* = \langle \tau_1^*, \ldots, \tau_n^* \to b \rangle$.

For objects, * can also be defined recursively in the following way:

1. If $a \in D_0$, then $a^* = a$;
2. If $\tau = \langle \tau_1, \ldots, \tau_n \rangle$ and $a \in D_\tau$, let χ_a be the characteristic function of a. Then a^* is a function from $D_{\tau_1} \times \cdots \times D_{\tau_n}$ to b such that:

$$a^*(a_1^*, \ldots, a_n^*) = \chi_a(a_1, \ldots, a_n).$$

Finally, if F is an operator, then we have:

$$F^*_{\mathcal{D}^*}(a_1^*, \ldots, a_n^*) = (F_{\mathcal{D}}(a_1, \ldots, a_n))^*$$

Notice that the above function is a bijection between D_τ and D_{τ^*}. Using this translation, it's possible to translate Feferman's homomorphism invariance into the relational setting. Let $f : D_0 \to D_0'$ be a surjective mapping.For any type τ, the *Feferman extension* of f, f_τ^{Fe}, is defined recursively as follows:

Definition C.12. 1. $f_0^{Fe} = f$;

2. For the other types, the definition will proceed in two steps: first, we will specify the domain of f_τ^{Fe}, and then we will specify its behavior. Let $\tau = \langle \tau_1, \ldots, \tau_n \rangle$.

   (a) The domain of f_τ^{Fe} will consist of all $a \subseteq D_{\tau_1} \times \cdots \times D_{\tau_n}$ such that, for all $a_i \in \mathrm{dom}(f_{\tau_i}^{Fe}), a_i' \in \mathrm{dom}(f_{\tau_i}^{Fe})(i \leq n)$, if $f_{\tau_i}^{Fe}(a_i) = f_{\tau_i}^{Fe}(a_i')(i \leq n)$, then $(a_1, \ldots, a_n) \in a$ iff $(a_1', \ldots, a_n') \in a$.

   (b) For any a that satisfies the above condition, $f_\tau^{Fe}(a)$ is defined as:

$$\{(f_{\tau_1}^{Fe}(a_1), \ldots, f_{\tau_n}^{Fe}(a_n)) \mid (a_1, \ldots, a_n) \in a \cap \mathrm{dom}(f_{\tau_1}^{Fe}) \times \cdots \times \mathrm{dom}(f_{\tau_n}^{Fe})\}$$

Notice that, in general, the Feferman extension of a given surjective mapping will only be a *partial* surjective mapping. This corresponds to the fact that, over functional type settings, in general the extension of a surjective mapping is also only partially defined (recall Definition B.10). The notion of *Feferman-invariance* is defined in the following way:

Definition C.13. An object a of type τ is *Feferman-invariant* iff for any surjective mapping $f : D_0 \to D'_0$, $a_{\mathcal{D}} \in \mathrm{dom}(f^{Fe}_{\tau})$ and $f^{Fe}_{\tau}(a_{\mathcal{D}}) = a_{\mathcal{D}'}$.

Definition C.14. A $\langle \tau_1, \ldots, \tau_n \to \sigma \rangle$-operator F is *Feferman-invariant* iff for any surjective mapping $f : D_0 \to D'_0$ and $a_1, \ldots, a_n$ of type $\tau_1, \ldots, \tau_n$, respectively, $F_{\mathcal{D}}(a_1, \ldots, a_n) \in \mathrm{dom}(f^{Fe}_{\sigma})$ and $f^{Fe}_{\sigma}(F_{\mathcal{D}}(a_1, \ldots, a_n)) = F_{\mathcal{D}'}(f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n))$.

In order to show that Feferman-invariance is the exact relational counterpart of (what Feferman calls) homomorphism-invariance in the functional setting, I'll need the following lemma:

Lemma C.1. *Let $f : D_0 \to D_0$ be a surjective mapping, τ be a relational type and suppose $a \in D_{\tau}$. Then:*

1. *$a \in \mathrm{dom}(f^{Fe}_{\tau})$ iff $a^* \in \mathrm{dom}(f_{\tau^*})$;*
2. *If $a \in \mathrm{dom}(f^{Fe}_{\tau})$, then $f^{Fe}_{\tau}(a) = f_{\tau^*}(a^*)$.*

Proof. The proof is by induction on the complexity of τ. The base case is immediate. So suppose that both 1 and 2 from the theorem are true for all types of lower complexity than τ.

Let $a \in \mathrm{dom}(f^{Fe}_{\tau})$ and $a^*_i \in \mathrm{dom}(f_{\tau^*_i})(i \leq n)$. By the definition of *, $a^*(a^*_1, \ldots, a^*_n)$ will be either 0 or 1. But, by Definition B.10, $0, 1 \in \mathrm{dom}(f_b)$, so $a^*(a^*_1, \ldots, a^*_n) \in \mathrm{dom}(f_b)$. On the other hand, suppose $a^*_1, b^*_1 \in \mathrm{dom}(f_{\tau^*_1}), \ldots, a^*_n, b^*_n \in \mathrm{dom}(f_{\tau^*_n})$ are such that $f_{\tau^*_i}(a^*_i) = f_{\tau^*_i}(b^*_i)$. By the induction hypothesis, $f^{Fe}_{\tau_i}(a_i) = f^{Fe}_{\tau_i}(b_1)(i \leq n)$, whence by Definition C.12, $(a_1, \ldots, a_n) \in a$ iff $(b_1, \ldots, b_n) \in a$. By the definition of $*$, this means that $a^*(a^*_1, \ldots, a^*_n) = a^*(b^*_1, \ldots, b^*_n)$, so, by Definition B.10, $f_b(a^*(a^*_1, \ldots, a^*_n)) = f_b(a^*(b^*_1, \ldots, b^*_n))$. Therefore, a^* meets the two conditions from clause 3 of Definition B.10, that is, $a^* \in \mathrm{dom}(f_{\tau^*})$.

Conversely, suppose $a^* \in \mathrm{dom}(f_{\tau^*})$. Consider $a_i, b_i (i \leq n)$ such that $f^{Fe}_{\tau_i}(a_i) = f^{Fe}_{\tau_i}(b_i)$. By the induction hypothesis, $a^*_i, b^*_i \in \mathrm{dom}(f_{\tau^*_i})$ and, moreover, $f_{\tau^*_i}(a^*_i) = f_{\tau^*_i}(b^*_i)$. But then, by Definition B.10, $f_b(a^*(a^*_1, \ldots, a^*_n)) = f_b(a^*(b^*_1, \ldots, b^*_n))$. By the definition of $*$, this means that $(a_1, \ldots, a_n) \in a$ iff $(b_1, \ldots, b_n) \in a$. Therefore, by Definition C.12, $a \in \mathrm{dom}(f^{Fe}_{\tau})$. This takes care of 1.

To see that 2 also holds, suppose $a \in \mathrm{dom}(f^{Fe}_{\tau})$ and let $f^{Fe}_{\tau}(a) = b$. Thus, by Definition C.12, we have:

$$b = \{(f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n)) \mid (a_1, \ldots, a_n) \in a \cap \mathrm{dom}(f^{Fe}_{\tau_1}) \times \cdots \times \mathrm{dom}(f^{Fe}_{\tau_n})\}.$$

Consider thus b^*. I'll show that, for any $(a^*_1, \ldots, a^*_n)$, we have $b^*(f_{\tau^*_i}(a^*_1), \ldots, f_{\tau^*_n}(a^*_n)) = f_b(a^*(a^*_1, \ldots, a^*_n))$. Notice that:

$$\begin{aligned} f_b(a^*(a^*_1, \ldots, a^*_n)) = 1 &\iff a^*(a^*_1, \ldots, a^*_n)) = 1 \\ &\iff (a_1, \ldots, a_n) \in a \\ &\iff (f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n)) \in b \\ &\iff b^*(f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n)) = 1 \end{aligned}$$

This concludes the proof. ■

Corollary C.1. *An object a is Feferman-invariant iff a^* is homomorphism-invariant (in the sense of Feferman).*

The proof of the corollary is immediate from the lemma.

Theorem C.3. *An operator F is Feferman-invariant iff its corresponding operator F^* in the functional type-setting is homomorphism invariant.*

Proof. Let F be a $\langle \tau_1, \ldots, \tau_n \to \sigma \rangle$-operator and $f : D_0 \to D_0$ be a surjective mapping. Suppose first that F is Feferman-invariant and let $a_1, \ldots, a_n$ be such that $a^*_1, \in \mathrm{dom}(f_{\tau^*_1}), \ldots, a^*_n \in \mathrm{dom}(f_{\tau^*_n})$. By the above lemma, for each a_i, $a_i \in \mathrm{dom}(f_{\tau_i})$ and, moreover, $f_{\tau_i}(a_i)^* = f_{\tau^*_i}(a^*_i)$. By the definition of Feferman-invariance, $F_{\mathcal{D}}(a_1, \ldots, a_n) \in f^{Fe}_{\sigma}$ and:

$$f^{Fe}_{\sigma}(F_{\mathcal{D}}(a_1, \ldots, a_n)) = F_{\mathcal{D}'}(f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n))$$

Hence, since, by the above lemma, $F^*_{\mathcal{D}}(a^*_1, \ldots, a^*_n) = F_{\mathcal{D}}(a_1, \ldots, a_n)^* \in f_{\sigma^*}$ and

$$f_{\sigma^*}(F^*_{\mathcal{D}}(a^*_1, \ldots, a^*_n)) = f^{Fe}_{\sigma}(F_{\mathcal{D}}(a_1, \ldots, a_n))^*,$$

it follows by Feferman-invariance and the above lemma again that

$$\begin{aligned} f_{\sigma^*}(F^*_{\mathcal{D}}(a^*_1, \ldots, a^*_n)) &= F_{\mathcal{D}'}(f^{Fe}_{\tau_1}(a_1), \ldots, f^{Fe}_{\tau_n}(a_n)) \\ &= F^*_{\mathcal{D}'}(f_{\tau^*_1}(a^*_1), \ldots, f_{\tau^*_n}(a^*_n)), \end{aligned}$$

as required.

The proof of the converse is very similar. Suppose F^* is homomorphism invariant and let $a^*_1, \ldots, a_n*$ be such that $a_i \in \mathrm{dom}(f^{Fe}_{\tau_i})$ for each $i \leq n$. By the lemma, $a^*_i \in \mathrm{dom}(f_{\tau^*_i})$ and,

moreover, $f_{\tau_i^*}(a_i) = f^{Fe}_{\tau_i}(a_i)$. By homomorphism-invariance, $F^*_{\mathcal{D}}(a_1^*, \ldots, a_n^*) \in \mathrm{dom}(f_{\sigma^*})$. Therefore, we have:

$$\begin{aligned} F_{\mathcal{D}'}(f^{Fe}_{\tau_1}(a_1), \ldots, f_{\tau_n}(a_n))^* &= F^*_{\mathcal{D}'}(f_{\tau_1^*}(a_1), \ldots, f_{\tau_n^*}(a_n)) \\ &= f_{\sigma^*}(F^*_{\mathcal{D}}(a_1^*, \ldots, a_n^*)) \\ &= f_{\sigma^*}(F_{\mathcal{D}}(a_1, \ldots, a_n)^*) \\ &= f^{Fe}_{\sigma}(F_{\mathcal{D}}(a_1, \ldots, a_n))^* \end{aligned}$$

Since * is a bijection, this gives the desired result. ■

The last type of invariance introduced by Casanovas is what he calls *preimage invariance*. Again, we start with a basic surjective function $f : D_0 \to D_0'$ and then we define recursively the *preimage extension* of f as follows:

Definition C.15. Let $f : D_0 \to D_0'$ be a surjective function. Define f^p_τ as follows:

1. $f_0 = f$;
2. If $\tau = \langle \tau_1, \ldots, \tau_n \rangle$, then:
   (a) The domain of f^p_τ consists of all relations $a \subseteq M_{\tau_i} \times M_{\tau_n}$ for which there's some $b \subseteq N_{\tau_1} \times \cdots \times N_{\tau_n}$ such that
   $$a = \{(a_1, \ldots, a_n) \in \mathrm{dom}(f^p_{\tau_1}) \times \mathrm{dom}(f^p_{\tau_n}) \mid (f^p_{\tau_1}(a_1), \ldots, f^p_{\tau_n}(a_n)) \in b\}.$$
   (b) For each such $a \in \mathrm{dom}(f^p_\tau)$, set
   $$f^p_\tau(a) = \{(f^p_{\tau_1}(a_1), \ldots, f^p_{\tau_n}(a_n)) \mid (a_1, \ldots, a_n) \in a\}.$$

The name "preimage extension" should be clear: this essentially consists in taking the preimages of the extensions of f defined at previous levels of the hierarchy. The definitions of preimage invariance for objects and operators is as expected:

Definition C.16. An object a of type τ is *preimage-invariant* iff for any surjective function $f : D_0 \to D_0'$, $a_{\mathcal{D}} \in \mathrm{dom}(f^p_\tau)$ and $f^p_\tau(a_{\mathcal{D}}) = a_{\mathcal{D}'}$.

Definition C.17. A $\langle \tau_1, \ldots, \tau_n \to \sigma \rangle$-operator F is *preimage-invariant* iff for any surjective function $f : D_0 \to D_0'$, if $a_i \in \mathrm{dom}(f^p_{\tau_i})(i \leq n)$, then $F_{\mathcal{D}}(a_1, \ldots, a_n) \in \mathrm{dom}(f^p_\sigma)$ and $f^p_\sigma(F_{\mathcal{D}}(a_1, \ldots, a_n)) = F_{\mathcal{D}'}(f^p_{\tau_1}(a_1), \ldots, f^p_{\tau_n}(a_n))$.

The next result shows that preimage invariance in a relational type and homomorphism invariance over a functional type coincide over the basic levels, which were the focus of Feferman's analysis. In the following, I'll abbreviate $\langle 0_1, \ldots, 0_n \rangle$ as 0^n and $\langle 0^{m_1}, \ldots, 0^{m_n} \to n \rangle$ as $\langle m_1, \ldots, m_r \to n \rangle$. I'll make use of the following lemma:

Lemma C.2. *If $\tau = 0^n$ and $f : D_0 \to D_0'$ is surjective, then $f_\tau^{Fe} = f_\tau^p$.*

Proof. I'll prove the lemma by showing, firstly, that $\mathrm{dom}(f_\tau^p) = \mathrm{dom}(f_\tau^{Fe})$ and, secondly, that $f_\tau^{Fe}(a) = f_\tau^p(a)$ for every $a \in \mathrm{dom}(f_\tau^{Fe})$.

Suppose first that $a \in \mathrm{dom}(f_\tau^{Fe})$. By Definition C.12, we have the following equality:

$$f_\tau^{Fe}(a) = \{(f_0^{Fe}(a_1), \ldots, f_0^{Fe}(a_n)) \mid (a_1, \ldots, a_n) \in a \cap (\mathrm{dom}(f_0^{Fe} \times \cdots \times \mathrm{dom}(f_0^{Fe}))\}$$

Let $f_\tau^{Fe}(a) = b$. Since $f_0 = f$ and f is total, it follows that the above can be simplified to:

$$f_\tau^{Fe}(a) = b = \{(f(a_1), \ldots, f(a_n)) \mid (a_1, \ldots, a_n) \in a\}.$$

But then, it follows immediately that:

$$a = \{(a_1, \ldots, a_n) \in \mathrm{dom}(f_0^p) \times \mathrm{dom}(f_0^p) \mid (f_0^p(a_1), \ldots, f_0^p(a_n)) \in b\}$$

whence $a \in \mathrm{dom}(f_\tau^p)$.

On the other hand, suppose $a \in \mathrm{dom}(f_\tau^p)$. By Definition C.15, there is $b \subseteq D_0' \times \cdots \times D_0'$ such that:

$$a = \{(a_1, \ldots, a_n) \in \mathrm{dom}(f_0^p) \times \mathrm{dom}(f_0^p) \mid (f_0^p(a_1), \ldots, f_0^p(a_n)) \in b\}$$

Suppose then that there are $a_i, a_i' (i \leq n)$ such that $f_0^{Fe}(a_i) = f_0^{Fe}(a_i')$. Since, by definition, $f_0^{Fe} = f_0 = f_0^p$, it follows that $f_0^p(a_i) = f_0^p(a_i')$. Therefore, by the above, $(a_1, \ldots, a_n) \in a$ iff $(a_1', \ldots, a_n') \in a$, that is, $a \in \mathrm{dom}(f_\tau^{Fe})$.

The next part follows easily from the definitions. Suppose $a \in \mathrm{dom}(f_\tau^{Fe})$. We have:

$$\begin{aligned} f_\tau^{Fe}(a) &= \{(f_0^{Fe}(a_1), \ldots, f_0^{Fe}(a_n)) \mid (a_1, \ldots, a_n) \in a \cap \mathrm{dom}(f_0^{Fe}) \times \cdots \times \mathrm{dom}(f_0^{Fe})\} \\ &= \{(f(a_1), \ldots, f(a_n)) \mid (a_1, \ldots, a_n) \in a\} \\ &= \{(f_0^p(a_1), \ldots, f_0^p(a_n)) \mid (a_1, \ldots, a_n) \in a\} \\ &= f_\tau^p(a) \end{aligned}$$

This concludes the proof. ∎

Observe that we used the fact that f is total and surjective in the above, so the proof can't be readily extended to higher types.

Corollary C.2. *If a is an object of type 0^n, then a is Feferman-invariant iff it is preimage invariant.*

Corollary C.3. *If a is an object f type 0^n, then a is preimage-invariant iff a^* is homomorphism-invariant in the sense of Feferman.*

Theorem C.4. *A $\langle m_1, \ldots, m_r \to n\rangle$-operator F is preimage invariant iff it is Feferman-invariant.*

Proof. Suppose $f : D_0 \to D'_0$ is a surjective mapping. By the above lemma, it follows that $F_{\mathcal{D}}(a_1, \ldots, a_n) \in \mathrm{dom}(F^{Fe}_{0^n})$ iff $F_{\mathcal{D}}(a_1, \ldots, a_n) \in \mathrm{dom}(F^{p}_{0^n})$, so this part is taken care of. Using the above lemma, we obtain the following equations:

$$F_{\mathcal{D}'}(f^{Fe}_{0^{m_1}}(a_1), \ldots, f^{Fe}_{0^{m_r}}(a_n)) = F_{\mathcal{D}'}(f^{p}_{0^{m_1}}(a_1), \ldots, f^{p}_{0^{m_n}}(a_n))$$

and

$$f^{Fe}_{0^n}(F_{\mathcal{D}}(a_1, \ldots, a_n)) = f^{p}_{0^n}(F_{\mathcal{D}}(a_1, \ldots, a_n)).$$

Therefore, we obtain that

$$F_{\mathcal{D}'}(f^{Fe}_{0^{m_1}}(a_1), \ldots, f^{Fe}_{0^{m_r}}(a_n)) = f^{Fe}_{0^n}(F_{\mathcal{D}}(a_1, \ldots, a_n))$$

iff

$$f^{Fe}_{0^n}(F_{\mathcal{D}}(a_1, \ldots, a_n)) = f^{p}_{0^n}(F_{\mathcal{D}}(a_1, \ldots, a_n)).$$

That is, F is Feferman-invariant iff it is preimage-invariant. ■

Corollary C.4. *A $\langle m_1, \ldots, m_r \to n\rangle$-operator F is preimage invariant iff F^* is homomorphism invariant.*

As Casanovas mentions, the above corollary helps to explain why conjunction acting on formulas with the same free variables comes out as invariant in Feferman's analysis, but not on his. Conjunction acting on formulas with the same free variables is basically the operation of intersection, which is preserved by preimages but not in general by surjective functions. I'll give a more precise characterization of the mapping-invariant objects and operators in the next section.

C.3 Invariant Objects and Operators

In this section, I'll present Casanovas's characterization of mapping-invariant objects and operators for the first levels of the hierarchy. For objects, I'll provide a characterization of 0^n and $\langle\langle 0\rangle\rangle$ mapping-invariant objects, whereas for operators I'll provide a characterization

of $\langle m \to n\rangle$-ary mapping-invariant operators; the more general case of $\langle m_1, \ldots, m_r \to n\rangle$-ary mapping-invariant operators is obtained by an easy generalization from the previous case. Since, as we saw in the last section, similarity invariance is equivalent to mapping invariance, I will switch between these two notions according to what is more convenient in each situation.

Mapping-invariant objects

The following lemma will prove to be very useful in the sequence:

Lemma C.3. *The union of mapping-invariant objects is also mapping-invariant.*

Proof. Let $a = \bigcup_{i\in I} b^i$ for mapping-invariant b_is, i.e. for each $\mathcal{D}$, $a_\mathcal{D} = \bigcup_{i\in I} b^i_\mathcal{D}$. Consider a surjective function $f : D_0 \to D'_0$. We have:

$$\begin{aligned} f_{0^n}(a_\mathcal{D}) &= \{(f_0(a_1), \ldots, f_0(a_n)) \mid (a_1, \ldots, a_n) \in b^i_\mathcal{D} \text{ for some } i \in I\} \\ &= \bigcup_{i\in I}\{(f_0(a_1), \ldots, f_0(a_n)) \mid (a_1, \ldots, a_n) \in b^i_\mathcal{D}\} \\ &= \bigcup_{i\in I} b^i_{\mathcal{D}'} \\ &= a_{\mathcal{D}'} \end{aligned}$$

This concludes the proof. ∎

In order to prove the first characterization theorem, I'll first define the *diagonal* objects and then state a couple of lemmas that will be useful in the proof.

Definition C.18. The *diagonal object* d^I of type 0^n is defined, for each $\mathcal{D}$ and $I \subseteq \{1, \ldots, n\}$, as follows: $d^I_\mathcal{D} = \{(a_1, \ldots, a_n) \in D_0^n \mid a_i = a_j \forall i, j \in I\}$.

Notice that, by this definition, $D_0 = d^I_\mathcal{D}$ for $I = \{1\}$.

Lemma C.4. *Suppose a is an object of type 0^n and that the following holds: if $d^I_\mathcal{D} \cap a_\mathcal{D} \neq \varnothing$ for $|D_0| > n$, then $d^I_{\mathcal{D}'} \subseteq a_{\mathcal{D}'}$ for any $\mathcal{D}'$. Then a is uniformly a union of objects d^I.*

Proof. Suppose a satisfies the hypothesis of the theorem and let $(a_1, \ldots, a_n) \in a$. Now, either there are $i, j \in I$ for $I \subseteq \{1, \ldots, n\}$ such that $a_i = a_j$ or not. If the former, then $(a_1, \ldots, a_n) \in d^I_\mathcal{D}$ for some I. If the latter, then $(a_1, \ldots, a_n) \in d^{\{1\}}_\mathcal{D}$. Either way, $(a_1, \ldots, a_n) \in d^I_\mathcal{D}$ for some I. Thus, for each element of a, there is some I that contains it. Let K be the set of all Is which contain elements of a. Clearly $a \subseteq \bigcup_{I\in K} I$. Conversely, for each such $I \in K$, $d^I_\mathcal{D} \cap a_\mathcal{D} \neq \varnothing$, so by the hypothesis of the theorem, $d^I_\mathcal{D} \subseteq a_\mathcal{D}$. Therefore, $\bigcup_{I\in K} I \subseteq a$, whence $\bigcup_{I\in K} = a$. ∎

Lemma C.5. *Each d^I is a mapping-invariant object and, moreover, any object which is uniformly a union of such d^Is is also mapping-invariant.*

Proof. Let $f : D_0 \to D_0$ be a surjective mapping. By definition, we have:

$$f_{0^n}(d^I_{\mathcal{D}}) = \{(f_0(a_1), \ldots, f_0(a_n)) \mid (a_1, \ldots, a_n) \in d^I_{\mathcal{D}}\}.$$

Since $a_i = a_j$ for every $i, j \in I$ and f is a function, it follows easily that $f_{0^n}(d^I_{\mathcal{D}}) \subseteq d^I_{\mathcal{D}'}$. Moreover, since f is surjective, for every $(b_1, \ldots, b_n) \in d^I_{\mathcal{D}'}$, there will be $(a_1, \ldots, a_n) \in d^I_{\mathcal{D}}$ such that $f(a_k) = b_k (k \leq n)$. Thus $f_{0^n}(d^I_{\mathcal{D}}) = d^I_{\mathcal{D}'}$, that is, d^I is mapping invariant.

By Lemma C.3, the union of these objects is also mapping-invariant. ■

Theorem C.5. *An object a of type 0^n is mapping-invariant iff it is the empty object ($a_{\mathcal{D}} = \varnothing$), or the universe ($a_{\mathcal{D}} = D_0^n$), or it is uniformly a union of objects d^I for $I \subseteq \{1, \ldots, n\}$.*

Proof. As both the empty set and universe are obviously mapping-invariant, the above lemma takes care of the right-to-left direction. Thus we only need to show that, if a is mapping-invariant, it has one of three forms specified above. Because of Lemma C.4, I only need to show that, for any mapping-invariant a of the appropriate type, if $d^I_{\mathcal{D}} \cap a_{\mathcal{D}} \neq \varnothing$, then, for any $\mathcal{D}'$, $d^I_{\mathcal{D}'} \subseteq a_{\mathcal{D}'}$.

So assume $|D_0| > n$ and let $(a_1, \ldots, a_n) \in d^I_{\mathcal{D}} \cap a_{\mathcal{D}}$. Consider $(b_1, \ldots, b_n) \in d^I_{\mathcal{D}'}$ for some $\mathcal{D}'$. Define a similarity relation π as follows:

$$\pi = \{((a_1, \ldots, a_n), (b_1, \ldots, b))\} \cup (D_0 \setminus \{a_1, \ldots, a_n\} \times D'_0).$$

By the mapping invariance of a, it follows that, for any $\mathcal{D}'$, $\pi(a_{\mathcal{D}}, a_{\mathcal{D}'})$. But then, by construction, $(b_1, \ldots, b_n) \in a_{\mathcal{D}'}$, as desired. ■

Since, in general, the intersection of d^I, d^J and the complement of d^I for some $I, J \subseteq \{1, \ldots, n\}$ is not the union of d^Is, it follows at once that the intersection and complement of mapping-invariant objects are not necessarily themselves mapping-invariant. This is in sharp contrast to the situation with permutation-invariant objects, which are preserved by intersection and complement. It also indicates that, in a relational type hierarchy, the objects determined by certain conjunctions and negations will not in general be mapping-invariant, again in contrast to the situation we saw above when discussing Feferman's theorem. Since, however, these questions are considered more naturally in the context of operators determined by first-order formulas, I'll postpone discussion of this latter point to the next section.

The next set-theoretical lemma will be useful when characterizing the mapping-invariant objects of type $\langle\langle 0\rangle\rangle$.

Lemma C.6. *Let A be a set such that $|A| < \kappa$. Then, for any function f whose domain is A, $|\mathrm{rng}(f)| < \kappa$.*

Proof. Suppose the hypothesis. By the axiom of choice, it's possible to well-order A. Relative to this well-ordering, define a function $g : \mathrm{rng}(f) \to A$ as follows: for any $b \in \mathrm{rng}(f)$, $g(b) = a$ such that a is the least element of A such that $f(a) = b$. Clearly g is injective, for, if $g(b) = g(b') = a$, then, by definition, $b = f(a) = b'$. Therefore, $|\mathrm{rng}(f)| \leq |A| < \kappa$. ■

Theorem C.6. *Suppose a is uniformly the union of the following objects:*

1. *the object e such that, for every $\mathcal{D}$, $e_{\mathcal{D}} = \{\varnothing\}$;*
2. *the object u such that, for every $\mathcal{D}$, $u_{\mathcal{D}} = \{D_0\}$;*
3. *the object p such that, for every $\mathcal{D}$, $p_{\mathcal{D}} = \mathcal{P}(D_0) \setminus \{\varnothing\}$;*
4. *for some cardinal number κ, the object b^κ such that, for every $\mathcal{D}$, $b^\kappa_{\mathcal{D}} = \{A \subseteq D_0 \mid A \neq \varnothing$ and $|A| < \kappa\}$.*

Then a is mapping-invariant.

Proof. Again, because of Lemma C.3, I only need to show that each of these objects is mapping-invariant. The first three objects are obviously mapping-invariant. To see that b^κ is also mapping-invariant, let $f : D_0 \to D_0'$ be a surjective mapping. In that case, we have the following:

$$f_{\langle\langle 0\rangle\rangle}(b^\kappa_{\mathcal{D}}) = \{f_{\langle 0\rangle}(A) \mid A \in b^\kappa\}.$$

By the above lemma, for each $A \in b^\kappa$, $|f_{\langle 0\rangle}(A)| < \kappa$. Therefore, $f_{\langle\langle 0\rangle\rangle}(b^\kappa_{\mathcal{D}}) \subseteq b^\kappa_{\mathcal{D}'}$.

Conversely, suppose $B \in b^\kappa_{\mathcal{D}'}$. By definition, $B \subseteq D_0'$ and $|B| < \kappa$. As f is surjective, for each $b \in B$, there's $a \in D_0$ such that $f(a) = b$. Define $A \subseteq D_0$ to be the set of elements in D_0 such that $f(a) = b$ for some b; if $f(a) = f(a')$, then pick only the least of them. Then $f \restriction A : A \to B$ is a bijection between A and B, whence $|A| = |B| < \kappa$, so $A \in b^\kappa_{\mathcal{D}}$, and, by construction, $f_{\langle 0\rangle}(A) = B$, i.e. $B \in f_{\langle\langle 0\rangle\rangle}(b^\kappa_{\mathcal{D}})$. Therefore, $b^\kappa_{\mathcal{D}'} \subseteq f_{\langle\langle 0\rangle\rangle}(b^\kappa_{\mathcal{D}})$, whence $b^\kappa_{\mathcal{D}'} = f_{\langle\langle 0\rangle\rangle}(b^\kappa_{\mathcal{D}})$. So b^κ is mapping-invariant, as required. ■

To prove the converse, I'll need the following lemma:

Lemma C.7. *Suppose a is a mapping-invariant object of type $\langle\langle 0\rangle\rangle$ and suppose that, for a given $\mathcal{D}$, $a_{\mathcal{D}}$ has non-empty elements A such that $A \neq D_0$. Let κ be a cardinal such that, for each $\mu < \kappa$, $a_{\mathcal{D}}$ has non-empty elements $A \neq D_0$ such that $|A| = \mu$. Then, for any $\mathcal{D}$, $b^\kappa_{\mathcal{D}} \subseteq a_{\mathcal{D}}$.*

Proof. Suppose the hypothesis of the theorem and let $B \subseteq D_0'$ be non-empty and such that $|B| = \mu < \kappa$. I must show that $B \in a_{\mathcal{D}'}$. By the hypothesis, there's a set $A \in a_{\mathcal{D}}$ such that $A \neq D_0'$ and such that $|A| = \mu$. Since $|A| = |B|$, there must be a bijection f between them; set then $\pi = f \cup ((D_0 \setminus A) \times D_0')$. Clearly π is a similarity relation between D_0 and D_0', and, by construction, $\pi(A, B)$. But a is mapping invariant, whence $B \in a_{\mathcal{D}'}$. ■

Remark C.1. If, in the above lemma, $b^\kappa_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$ for some $\mathcal{D}$, it follows that κ must be regular. Otherwise, κ could be reached by a union of less than κ sets of size less than κ, whence b^κ would also be reachable by such a union. Since $a_{\mathcal{D}}$ would contain all sets of sizes less than κ (by the lemma), it would follow that $b^\kappa_{\mathcal{D}} \subseteq a_{\mathcal{D}}$, contradicting the hypothesis.

Remark C.2. Another corollary of the above lemma is that, if $b^\kappa_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$, and $\kappa = \lambda^+$, then there can be no $A \neq D_0$ such that $|A| \geq \lambda$ and $A \in a_{\mathcal{D}}$. The above lemma shows that, if $a_{\mathcal{D}}$ contains one set of cardinality μ, it must contain all of them. Thus, if there was $A \in a_{\mathcal{D}}$ satisfying the above conditions, $b^\kappa_{\mathcal{D}} \subseteq a_{\mathcal{D}}$, contrary to the hypothesis.

Theorem C.7. *Suppose a is a mapping-invariant object of type $\langle\langle 0\rangle\rangle$. Then a is either the empty object or uniformly a union of the following objects:*

1. *the object e such that, for every $\mathcal{D}$, $e_{\mathcal{D}} = \{\varnothing\}$;*
2. *the object u such that, for every $\mathcal{D}$, $u_{\mathcal{D}} = \{D_0\}$;*
3. *the object p such that, for every $\mathcal{D}$, $p_{\mathcal{D}} = \mathcal{P}(D_0) \setminus \{\varnothing\}$;*
4. *for some cardinal number κ, the object b^κ such that, for every $\mathcal{D}$, $b^\kappa_{\mathcal{D}} = \{A \subseteq D_0 \mid A \neq \varnothing$ and $|A| < \kappa\}$.*

Proof. There are, at first, two cases to consider: either (i) there is no $\mathcal{D}$ such that $a_{\mathcal{D}}$ has a nonempty element $A \neq D_0$ or (ii) there is such a $\mathcal{D}$. If (ii), there are two more cases to consider: (a) for arbitrarily large κ, there is $\mathcal{D}$ such that $b^\kappa_{\mathcal{D}} \subseteq a_{\mathcal{D}}$ or (b) there's a least κ such that, for some $\mathcal{D}$, $b^\kappa_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$. Finally, if (b), then, by the above remark, $\kappa = \lambda^+$ for some λ. In that case, there are two further cases to consider: either (b′) there's no $\mathcal{D}$ such that $|D_0| > \lambda$ and $D_0 \in a_{\mathcal{D}}$ or (b″) there's $\mathcal{D}$ such that $|D_0| \geq \lambda$ and $D_0 \in a_{\mathcal{D}}$. Let's tackle each such case in turn.

Case (i): There's no $\mathcal{D}$ such that $a_{\mathcal{D}}$ has a non-empty element $A \neq D_0$. Thus, for each $\mathcal{D}$, $a_{\mathcal{D}} \subseteq \{\varnothing, D_0\}$. By the mapping invariance of a, it follows that either $a_{\mathcal{D}} = \varnothing$ for each $\mathcal{D}$, or $a_{\mathcal{D}} = \{D_0\}$ for each $\mathcal{D}$, or else $a_{\mathcal{D}} = \{\varnothing, D_0\}$ for each $\mathcal{D}$. Thus, in this case, a satisfies the theorem.

Case (iia): Suppose that, for arbitrarily large κ, there is $\mathcal{D}$ such that $b^{\kappa}_{\mathcal{D}} \subseteq a_{\mathcal{D}}$. Consider an arbitrary $\mathcal{D}$ such that $|D_0| = \kappa$ for some κ. By the hypothesis and the mapping invariance of a, $b^{\kappa}_{\mathcal{D}} \subseteq a_{\mathcal{D}}$. But then, either $a_{\mathcal{D}} = \mathcal{P}(D_0)$, or else $a_{\mathcal{D}} = \mathcal{P}(D_0) \setminus \{\varnothing\}$. Since $\mathcal{D}$ was arbitrary, this holds uniformly for each $\mathcal{D}$, so the theorem is also true in this case.

Case (iib′): There's a least $\kappa = \lambda^{+}$ such that $b^{\kappa}_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$ for some $\mathcal{D}$ and, moreover, there's no $\mathcal{D}$ such that $|D_0| \geq \lambda$ and $D_0 \in a_{\mathcal{D}}$. I claim that either $a = b^{\lambda}$, or $a = b^{\kappa} \cup e$. Note that, since κ is the least cardinal such that $b^{\kappa}_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$ for some $\mathcal{D}$, it follows immediately that $b^{\lambda} \subseteq a$. To see the converse, suppose $A \in a_{\mathcal{D}}$ for an arbitrary $\mathcal{D}$. Either $A = \varnothing$ or $A \neq \varnothing$. Suppose the latter. Then $|A| = \mu$ for some $\mu < \lambda$ (otherwise, as remarked above, $b^{\kappa} \subseteq a$, contradicting the hypothesis). But then $A \in b^{\lambda}_{\mathcal{D}}$, as required. Therefore, either $a_{\mathcal{D}} = b^{\lambda}_{\mathcal{D}}$, or else $a_{\mathcal{D}} = b^{\lambda}_{\mathcal{D}} \cup \{\varnothing\}$, which is what we wanted to prove.

Case (iib″): There's a least $\kappa = \lambda^{+}$ such that $b^{\kappa}_{\mathcal{D}} \not\subseteq a_{\mathcal{D}}$ for some $\mathcal{D}$ and, moreover, there's $\mathcal{D}$ such that $|D_0| \geq \lambda$ and $D_0 \in a_{\mathcal{D}}$. I claim that, for every $\mathcal{D}'$, $D'_0 \in a_{\mathcal{D}'}$. There are three possibilities: either $|D'_0| < \lambda$, or $|D_0| = |D'_0| \geq \lambda$ or $|D_0|, |D_0| \geq \lambda$ but $|D_0| \neq |D'_0|$. If the first, then $D'_0 \in a_{\mathcal{D}'}$. If the second, then there's a bijection f between D_0 and D'_0, which is also a similarity relation. By the mapping invariance of a, it follows that $D'_0 \in a_{\mathcal{D}'}$. Finally, if the third possibility obtains, suppose, without loss of generality, that $|D_0| > |D'_0|$. Let then $A \subseteq D_0$ and $B \subseteq D'_0$ be such that $|A| = |B| \geq \lambda$, with $A \neq D_0$. As $|A| = |B|$, there's a bijection f between them; set $\pi = f \cup ((D_0 \setminus A) \times D'_0)$. Clearly, π is a similarity relation, whence there must be $C \in a_{\mathcal{D}'}$ such that $\pi(D_0, C)$; by construction, $C = D'_0$. Therefore, $D'_0 \in a_{\mathcal{D}'}$. Hence, it follows that either $a = b^{\lambda} \cup u$, or $a = b^{\lambda} \cup e \cup u$, uniformly for each $\mathcal{D}$. So the theorem also applies in this case. ■

Mapping-invariant $\langle m \to n \rangle$-ary operators

I'll show in this section, first, that every mapping-invariant $\langle m \to n \rangle$-ary operator can be obtained from certain basic ones by means of three main operations. Using this result, I'll then show that every such operator can be determined by a first-order formula.

The basic operators are:

1. The *constant* $\langle m \to 1 \rangle$-ary operators $C^{m}_{\top}$ and $C^{m}_{\bot}$ whose actions on $R \subseteq D^{m}_{0}$ are:
   (a) $C^{m}_{\top}(R) = D_0$;
   (b) $C^{m}_{\bot}(R) = \varnothing$.
2. The $\langle m \to nm \rangle$*-ary diagonal operator* Δ^{m}_{n} such that, for any $R \subseteq D^{m}_{0}$,

$$\Delta^{m}_{n}(R) = \{n \times a \mid a \in R\}$$

where $n \times a$ is the n-fold concatenation of the tuple a, i.e., if $a = (a_1, \ldots, a_n)$, then $n \times a = (b_1, \ldots, b_{nm})$, where, for each $0 \leq k < n$ and $1 \leq i \leq m$, $b_{km+i} = a_i$. That is, Δ_n^m produces n copies of a and concatenates them.

3. The $\langle m \to m-1 \rangle$*-ary* i*-projection operator* Π_i^m (for $m \geq 2$ and $1 \leq i \leq m$), such that, for any $R \subseteq D_0$,

$$\Pi_i^m(R) = \{(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_m \mid (a_1, \ldots, a_m) \in R\}$$

4. For any $\sigma \in \mathrm{Sym}\{1, \ldots, m\}$ (the symmetric group of $\{1, \ldots, m\}$), the $\langle m \to m \rangle$*-ary permutation operator* P_σ, such that, for any $R \subseteq D_0^m$,

$$P_\sigma(R) = \{(a_{\sigma(1)}, \ldots, a_{\sigma(m)}) \mid (a_1, \ldots, a_n) \in R\}$$

The *generating operations for operators* are the following ones:

1. *Product.* If F is $\langle m \to n_1 \rangle$-ary and G is $\langle m \to n_2 \rangle$-ary, the product $F \times G$ is the $\langle m \to n_1 + n_2 \rangle$-ary operator such that, for $R \subseteq D_0^m$,

$$F \times G(R) = F(R) \times G(R)$$

2. *Sum.* If F and G are $\langle m \to n \rangle$-ary operators, the sum $F \cup G$ is the $\langle m \to n \rangle$-ary operator such that, for any $R \subseteq D_0$,

$$F \cup G(R) = F(R) \cup G(R)$$

3. *Composition.* If F is $\langle m \to n \rangle$-ary and G is $\langle n \to k \rangle$-ary, then the composition $G \circ F$ is the $\langle m \to k \rangle$-ary operator such that, for any $R \subseteq D_0$,

$$G \circ F(R) = G(F(R))$$

If F is an operator generated only from projections, diagonals, and permutations by composition, then F is called *intern.*

Lemma C.8. *An* $\langle m \to n \rangle$*-ary operator* F *is intern iff there's a map* $\sigma : \{1, \ldots, n\} \to \{1, \ldots, m\}$ *such that for any* $R \subseteq D_0^m$*,*

$$F(R) = \{(a_1, \ldots, a_n) \mid \textit{for some } (a'_1, \ldots, a'_m) \in R, a_i = a'_{\sigma(i)} \textit{ for all } 1 \leq i \leq n\}$$

Proof. From left to right, the proof is by induction on the generating sequence of F. If F is a projection Π_i^m, then, for $1 \leq j \leq n$, let σ be defined as:

$$\sigma(j) = \begin{cases} j & \text{if } j < i \\ j+1 & \text{if } j \geq i \end{cases}$$

If F is a permutation P_ζ, then $\sigma = \zeta$. If F is a diagonal Δ_n^m, then, for $1 \leq j \leq nm$, define σ as

$$\sigma(j) = \begin{cases} j & \text{if } j \leq m \\ j - km & \text{if } j > m \end{cases}$$

where k is such that $km + i = j$ for some $1 \leq i \leq m$. Finally, if F is a composition $G \circ H$ such that G and H satisfy the hypothesis of the proposition, let ζ be the permutation defined for G and ζ' be the permutation defined for H. Then $\sigma = \zeta \circ \zeta'$.

Conversely, suppose F is obtained in the manner described by the proposition. There are three cases to consider: either $n < m$, or $n = m$, or $m < n$. Let's tackle each in turn. If $n = m$, then $\sigma \in \mathrm{Sym}\{1, \ldots, m\}$ and $F = P_\sigma$. If $n < m$, let ζ be an extension of σ to m defined in the following way: $\zeta(i) = \sigma(i)$ for $i \leq n$, and $\zeta(i) = i$ if $n < i$. Then $\zeta \in \mathrm{Sym}\{1, \ldots, m\}$. Define the $\langle m \to n \rangle$-ary operator G as follows:

$$G(R) = \Pi_{n+1}^m \circ \cdots \circ \Pi_m^m \circ P_\zeta(R)$$

I claim $G(R) = F(R)$ for any $R \subseteq D_0^m$. We have:

$$\begin{aligned} (a_1, \ldots, a_n) \in G(R) &\iff \text{for some } (a'_1, \ldots, a'_m) \in R, a_i = a'_{\zeta(i)} \text{ for all } i \leq n \\ &\iff \text{for some } (a'_1, \ldots, a'_m) \in R, a_i = a'_{\sigma(i)} \text{ for all } i \leq n \\ &\iff (a_1, \ldots, a_n) \in F(R) \end{aligned}$$

Finally, suppose $m < n$. There are two cases to consider: either m divides n or not. If the former, then $n = km$ for some $k < n$. This means that we can partition n into k blocks, in such a way that σ can be decomposed into the union of smaller $\sigma_0, \ldots, \sigma_{k-1}$ such that each $\sigma_j (j < k)$ is a bijection between the jth block and m. Define then the $\langle m \to n \rangle$-ary operator G as follows:

$$G(R) = P_\zeta \circ \Delta_k^m(R)$$

where ζ is defined as:

$$\zeta(i) = \sigma_j(i) + jm$$

with j being the block in which i is in. Again, I claim that $G(R) = F(R)$. Notice that, by construction, $a_{\zeta(i)} = a_{\sigma(i)} (i \leq n)$. Therefore, we have:

$$\begin{aligned}(a_1,\ldots,a_n)\in G(R) &\iff \text{for some } (a'_1,\ldots,a'_n)\in\Delta^m_k(R), a_i=a'_{\sigma(i)}(i\leq n)\\ &\iff \text{for some } (a^*_1,\ldots,a^*_m)\in R, a_i=a^*_{\sigma(i)}\\ &\iff (a_1,\ldots,a_n)\in F(R)\end{aligned}$$

The other case is when m doesn't divide n. In that case, by the division algorithm, $n=km+r$ for some $0<r<m$. We proceed in the same way as in the first case, but instead of employing Δ^m_k, we use Δ^m_{k+1} and the projection functions $\Pi^{km+m}_{km+(r+1)},\ldots,\Pi^{km+m}_{km+m}$. It's clear that we can suitably modify the argument above to show that this operator is the same as F. ■

Definition C.19. A *component* is either a permutation of products of intern operators, or else an operator of the form $\Delta^1_{n_l}\circ C^m_\top$ for some n_l and m.

The idea is to show that every mapping-invariant operator can be decomposed into components of the above form; the indices will become clear when I state the theorem. In order to show this, I'll define a technical notion which will be very useful in what follows, namely the notion of what Casanovas calls a *free system*:

Definition C.20. Let $R\subseteq D^m_0$ for some $\mathcal{D}$. Then (D_0,R) is a *free system* if R is infinite, for all $a\in D_0$ there is at most one tuple $(a_1,\ldots,a_m)\in R$ such that $a=a_i$ for some $i\leq m$, and for all $(a_1,\ldots,a_m)\in R$, $a_i\neq a_j$ if $i\neq j$.

Before stating the theorem, a remark on notation: if R is an n-ary relation, I'll follow Casanovas and employ $\mathrm{field}_j(R)=\{a\mid (a_1,\ldots,a_n)\in R \text{ and } a=a_j(j\leq n)\}$, i.e. $\mathrm{field}_j(R)$ is the set of all jth coordinates of R; $\mathrm{field}(R)$ is the union of all $\mathrm{field}_j(R)(j\leq n)$.

Theorem C.8. *Let F be an $\langle m\to n\rangle$-ary mapping-invariant operator. Let $R\subseteq D^m_0$ and suppose (D_0,R) is a free system. Then for any $(a_1,\ldots,a_n)\in F_{\mathcal{D}}(R)$ there are both a decomposition*

$$\{1,\ldots,n\}=I_1\cup\cdots\cup I_k$$

with each $I_j, I_l(1\leq i,j,\leq k$ pairwise disjoint and also operators $F_1,\ldots,F_k$ such that:

1. *F_l is $\langle m\to n_l\rangle$-ary, where $n_l=|I_l|$;*
2. *If $i\in I_l$ and $a_i=a_j$, then $j\in I_l$;*
3. *F_l is intern if $\{a_i\mid i\in I_l\}$ has more than one element or if it has just one element a and $a\in\mathrm{field}_j(R)$ for some j;*

4. *If* $\{a_i \mid i \in I_l\}$ *has only one element* a *and* $a \notin \mathrm{field}(R)$*, then* $F_l = \Delta^1_m \circ C^m_\top$*, that is,* $F_l(S) = \{n_l \times a \mid a \in D_0\}$ *for all* $S \subseteq D_0^m$*;*

5. $(a_i : i \in I_l) \in F_l$*;*

6. *There is a* $\sigma \in \mathrm{Sym}\{1, \ldots, m\}$ *such that for all* $S \subseteq D_0'^m$*, if* (D_0, S) *is a free system, then:*

$$P_\sigma(F_1(S) \times \cdots \times F_k(S)) \subseteq F(S)$$

Proof. The construction is a bit involved, but nevertheless clear. I'll construct a sequence of operators and sets of indices in such a way that they are tailor-made for verifying the theorem. The idea is to construct each operator and index by recursion.

We're given a mapping-invariant $\langle m \to n\rangle$-ary operator F, a m-ary relation R such that (D_0, R) is a free system, and a tuple $(a_1, \ldots, a_n) \in F(R)$. We proceed as follows. If $a_1 \in \mathrm{field}(R)$, select I_1 as a maximal subset of $\{1, \ldots, m\}$ for which $1 \in I_1$ and there is an intern $\langle m \to n_l\rangle$-ary operator G such that $(a_k : j \in I_1) \in G(R)$ and put $F_1 = G$. If $a_1 \notin \mathrm{field}(R)$, then take $I_1 = \{i \mid a_1 = a_i\}$ and put $F_1 = \Delta^1_{n_1} \circ C^m_\top$. Next, suppose l is the least element of $\{1, \ldots, m\} \setminus I_1$. We repeat the procedure for I_1: if $a_l \in \mathrm{field}(R)$, choose a maximal subset I_2 of $\{1, \ldots, m\} \setminus I_1$ for which $l \in I_2$ and there is an intern $\langle m \to n_2\rangle$-ary operator G such that $(a_i : i \in I_2) \in G(R)$, and put $F_2 = G$. If, on the other hand, $a \notin \mathrm{field}(R)$, set $I_2 = \{i \mid a_l = a_i\}$ and $F_2 = \Delta^1_{n_2} \circ C^m_\top$. This procedure will eventually exhaust all of $\{1, \ldots, m\}$, generating a sequence of operators $F_1, \ldots, F_k$ and sets of indices $I_1, \ldots, I_k$. I'll show that this sequence satisfies the six conditions laid out above.

Condition 1: If $a_i \in \mathrm{field}(R)$, then $F_l = G$ for some $\langle m \to n_l\rangle$-ary intern operator G; otherwise, $F_l = \Delta^1_{n_l} \circ C^m_\top$, which is also of the required arity.

Condition 2: If $a_i \notin \mathrm{field}(R)$, the result is immediate by the construction of I_l. So suppose $a_i \in \mathrm{field}(R)$ and $i \in I_l$. Suppose also, towards a contradiction, that there is a j such that $a_i = a_j$, but $j \notin I_l$. Consider $I_{l'} = I_l \cup \{j\}$. As there's an intern operator G such that $(a_i : i \in I_l) \in G(R)$, and $a_i = a_j$ for some $i \in I_l$, it follows also that $(a_j : j \in I_{l'}) \in G(R)$, contradicting the maximality of I_l. Thus, if $i \in I_l$ and $a_i = a_j$, $j \in I_l$ as well.

Condition 3: Suppose either $\{a_i \mid i \in I_l\}$ has more than one element, or that it has only one element a such that $a \in \mathrm{field}(R)$. If the latter, this means that $F_l = G$ for some intern operator G. If the former, suppose toward contradiction that F_l is not intern. Then $I_l = \{i \mid a_i = a_j\}$ for some j, whence $\{a_i \mid ; i \in I_l\}$ has only one element, contradicting the hypothesis. Thus, if $\{a_i \mid i \in I_l\}$ has more than one element, F_l is intern.

Condition 4: Suppose $\{a_i \mid i \in I_l\}$ has only one element a and $a \notin \mathrm{field}(R)$. Then $I_l = \{i \mid a_i = a\}$ and $F_l = \Delta^1_{n_l} \circ C^m_\top$, as desired.

Condition 5: Consider $(a_i : i \in I_l)$. If $I_l = \{i \mid a_i = a\}$ for some $a \notin \text{field}(R)$, then, for each $a_i \in (a_i : i \in I_l)$, $a_i = a$, so $(a_i \mid i \in I_l)$ is an n_l-ary sequence of repeated as, that is, $(a_i \mid i \in I_l) = n_l \times a \in \Delta^1_{n_l} \circ C^m_\top$. Otherwise, the result follows by the condition we imposed on G.

Condition 6: This is the trickiest. Assume that $S \subseteq D_0'^m$ is such that (D_0, S) is a free system and $(b_1, \ldots, b_n)$ is such that $(b_i : i \in I_l) \in F_l(S)$ for $l \leq k$. I'll show that $(b_1, \ldots, b_n) \in F(S)$. But first, let's construct the permutation σ. For each I_l, let s_l be an enumeration of I_l and $s_1{}^\frown s_2{}^\frown \ldots {}^\frown s_k$ be their concatenation. We set $\sigma = s_1{}^\frown s_2{}^\frown \ldots {}^\frown s_k$. Set J as the set of all $l \in \{1, \ldots, k\}$ such that F_l is intern. By Lemma C.8, if $l \in J$, there's a mapping $\sigma_l : I_l \to \{1, \ldots, m\}$ such that for all $T \subseteq D_0^{*m}$,

$$F_l(T) = \{(c_i : i \in I_l) \mid \text{ for some } (c'_1, \ldots, c'_m) \in T, c_i = c'_{\sigma_l(i)} \text{ for all } i \in I_l\}.$$

Therefore, for R and S in particular, this generates $(a^l_1, \ldots, a^l_m) \in R$ and $(b^l_1, \ldots, b^l_m) \in S$ such that $a_i = b^l_{\sigma_l(i)}$ and $b_i = b^l_{\sigma_l(i)}$ for all $i \in I_l$.

If $l \notin J$, then set $a^l = a_i$ and b^l_i for all $i \in I_l$. Let L be the set of all $l \in \{1, \ldots, k\} \setminus J$ such that $b^l \in \text{field}(S)$. Given that $b_l \in \text{field}(S)$, it follows that $b^l = b^l_i$ for some i and some tuple $(b^l_1, \ldots, b^l_m) \in S$. Choose one such tuple and another arbitrary tuple $(a^l_1, \ldots, a^l_m) \in R$; the only requirement on the tuple so chosen is that, for each $l, l' \in \{1, \ldots, k\}$, the corresponding tuples must be distinct. This is possible because R is infinite.

Set now:

$$A = D_0 \setminus (\{a^l_i \mid 1 \leq i \leq m \text{ and } l \in J \cup L\} \cup \{a^l \mid l \notin J\})$$

and

$$B = D'_0 \setminus (\{b^l_i \mid 1 \leq i \leq m \text{ and } l \in J \cup L\} \cup \{b^l \mid l \notin J\})$$

Define now a similarity relation π as follows:

$$\pi = (A \times B) \cup \{(a^l_i, b^l_i) \mid 1 \leq i \leq m \text{ and } l \in J \cup L\} \cup \{(a^l, b^l) \mid l \notin J\}$$

Clearly, $\pi(R, S)$, so, by the mapping invariance of F, $\pi(F(R), F(S))$. Thus, there is $(c_1, \ldots, c_n) \in F(S)$ such that $\pi(a_i, c_i)$ for $i \leq n$. But, by our construction of π and the l tuples, if $(l, i) \neq (l', i')$, then $a^l_i \neq a^{l'}_{i'}$, $a^l_i \neq a^{l'}$ for any l, i, l', and, if $l \neq l'$, $a^l \neq a^{l'}$. It follows that $(b_1, \ldots, b_n) = (c_1, \ldots, c_n) \in F(S)$, as required. ■

Theorem C.9. *Let F be an $\langle m \to n \rangle$-ary mapping-invariant operator. Assume D_0 is infinite. For any $(a_1, \ldots, a_n) \in F_{\mathcal{D}}(\varnothing)$, there are both a decomposition*

$$\{1, \ldots, n\} = I_1 \cup \cdots \cup I_k$$

with I_k, I_l disjoint for $k \neq l$ and also operators $F_1, \ldots, F_k$ such that:

1. *$F_l = \Delta^1_{n_l} \circ C^m_\top$ for $n_l = |I_l|$;*
2. *If $i \in I_l$ and $a_i = a_j$, then $j \in I_l$;*
3. *$(a_i : i \in I_l) \in F_l(\varnothing)$;*
4. *There is a $\sigma \in \mathrm{Sym}\{1, \ldots, m\}$ such that for any infinite set D'_0,*

$$P_\sigma(F_{1_{\mathcal{D}'}}(\varnothing) \times \cdots \times F_{k_{\mathcal{D}'}}(\varnothing)) \subseteq F_{\mathcal{D}'}(\varnothing)$$

Proof. Essentially the same as in the previous theorem. ■

Theorem C.10. *For any D_0, there are D'_0, f such that D'_0 is infinite and $f : D'_0 \to D_0$ is surjective.*

Proof. Let D_0 be arbitrary, I an infinite index set and $\{a_i \mid i \in I\}$ an enumeration of D_0 by I. Set $I = D'_0$ and let $f(i) = a_i$. Clearly f is a surjection, thus concluding the proof. ■

Theorem C.11. *For all R, D_0 such that $R \subseteq D_0^m$ and $R \neq \varnothing$, there are S, D'_0, f such that $S \subseteq D_0'^m$, (D'_0, S) is a free system and $f : D'_0 \to D_0$ is a surjection such that $f(S) = R$.*

Proof. Let I be an infinite index set and enumerate R as $\{(a^i_1, \ldots, a^i_n) \mid i \in I\}$. Let J be another index set such that $(b_j : j \in J)$ is an enumeration of $D_0 \setminus \mathrm{field}(R)$ and such that $J \cap (I \times \{1, \ldots, m\}) = \varnothing$. Define $D'_0 = J \cup (I \times \{1, \ldots, m\})$ and set $f : D'_0 \to D_0$ as $f(j) = b_j$ for $i \in J$ and $f((i,k)) = a^i_k$ for $(i,j) \in I \times \{1, \ldots, m\}$. Finally, define $S = f^{-1}(R)$. Clearly D'_0, S, and f satisfy the hypothesis of the theorem. ■

Theorem C.12. *Let F be an $\langle m \to n \rangle$-ary mapping-invariant operator. Let R and S be two m-ary relations over D_0, D'_0, respectively, and assume $f : D'_0 \to D_0$ is a surjective mapping such that $f(S) = R$. Then:*

1. *$f(F_{\mathcal{D}'}(S)) = F_{\mathcal{D}}(R)$;*
2. *If G is also an $\langle m \to n \rangle$-ary mapping-invariant operator and $F_{\mathcal{D}'}(S) = G_{\mathcal{D}'}(S)$, then $F_{\mathcal{D}}(R) = G_{\mathcal{D}}(R)$.*

Proof. For 1, we have, by mapping invariance, $f(F_{\mathcal{D}'}(S)) = F_{\mathcal{D}}(f(S)) = F_{\mathcal{D}}(R)$. For 2, because of 1, we have:

$$\begin{aligned} F_{\mathcal{D}}(R) &= f(F_{\mathcal{D}'}(S)) \\ &= f(G_{\mathcal{D}'}(S)) \\ &= G_{\mathcal{D}}(f(S)) \\ &= G_{\mathcal{D}}(R) \end{aligned}$$

Theorem C.13. *An $\langle m \to n\rangle$-ary operator F is mapping-invariant iff there are two $\langle m \to n\rangle$-ary operators G, H generated from the basic operators by sum, product, and composition and such that:*

1. *For all $\mathcal{D}$, $F_{\mathcal{D}}(\varnothing) = G_{\mathcal{D}}(\varnothing)$;*
2. *For all $\mathcal{D}$ and nonempty $R \subseteq D_0^m$, $F_{\mathcal{D}}(R) = H_{\mathcal{D}}(R)$.*

Proof. The right-to-left direction of the theorem is established by an easy induction on the generating sequence of G and H. As for the other direction, let's tackle the empty case first. Because of Theorem C.10 and Theorem C.12, we can focus on only those $\mathcal{D}$ such that D_0 is infinite. But then, by Theorem C.9, $F_{\mathcal{D}}(\varnothing)$ will be a finite sum of permutations of components; let G be this sum. Then G satisfies the theorem.

Next, consider case 2. Again, using Theorem C.11 and Theorem C.12, we can focus on those cases when (D_0, R) is a free system. But then, by Theorem C.8, $F_{\mathcal{D}}(R)$ will be a finite union of components, thus satisfying the theorem. ■

Definition C.21. Let P be an m-ary relation symbol and let $\phi(x_1, \ldots, x_n)$ be a first-order formula having only P as its extralogical symbol. The $\langle m \to n\rangle$-ary operator G^{ϕ} attached to ϕ is defined as: for any nonempty D_0 and $R \subseteq D_0^m$,

$$G^{\phi}_{\mathcal{D}}(R) = \{(a_1, \ldots, a_n) \mid (D_0, R) \models \phi(a_1, \ldots, a_n)\}$$

Definition C.22. Let $\mathcal{L} = \{P\}$, where P is an m-ary relation symbol. Let $\top$ be any tautology and $\bot$ be any contradiction. The set of *mapping-invariant* formulas of $\mathcal{L}$ is the least set satisfying the following:

1. $\top$ and $\bot$ are mapping-invariant;
2. $P(y_1, \ldots, y_m)$ is mapping-invariant for any distinct variables $y_1, \ldots, y_m$;
3. $\neg\exists y_1, \ldots, y_m P(y_1, \ldots, y_m)$ is mapping-invariant for any distinct variables $y_1, \ldots, y_m$;

4. If ϕ, ψ are mapping-invariant, then $(\phi \vee \psi)$ is mapping-invariant;

5. If ϕ, ψ are mapping-invariant and have no common free variables, then $(\phi \wedge \psi)$ is mapping-invariant;

6. If $\phi(x_1, \ldots, x_n)$ is mapping-invariant and y is a variable distinct from x_i for $i \leq n$, then $(\phi(x_1, \ldots, x_n) \wedge x_i = y)$ is mapping-invariant for all $i \leq n$;

7. If ϕ is mapping invariant, then $\exists x \phi$ is mapping-invariant.

Theorem C.14. *An $\langle m \to n \rangle$-ary operator F is mapping-invariant iff $F = G^{\phi}$ for some mapping-invariant ϕ in a language $\mathcal{L} = \{P\}$.*

Proof. Right-to-left follows easily by induction on ϕ. For left-to-right, we need to show, first, that every operator obtained from the basic ones by sum, product, or composition is definable by a mapping-invariant formula. This can be done by an induction on the generating sequence of the operator. Finally, suppose this result. By Theorem C.13, it follows that the behavior of F is determined by two operators G, H which are obtained from the basic ones by the above operations. By the result, it follows that there are formulas $\psi(x_1, \ldots, x_n)$ and $\chi(x_1, \ldots, x_n)$ such that $G = G^{\psi}$ and $H = H^{\chi}$. Let then $\phi(x_1, \ldots, x_n)$ be the following formula:

$$(\psi(x_1, \ldots, x_n) \wedge \neg \exists \bar{y} P(\bar{y})) \vee (\chi(x_1, \ldots, x_n) \wedge \exists \bar{y} P(\bar{y}))$$

Then ϕ is a mapping-invariant formula and $F = G^{\phi}$, as required. ■

C.4 Conclusion

As we have seen, in his article, Feferman presented three main criticisms of Tarski's proposal:

1. It assimilated logic to set theory;

2. The notions involved in explaining the semantics of $\mathcal{L}_{\infty\infty}$ are not set-theoretically robust, i.e. they're not absolute;

3. It gives no explanation of what constitutes the *same* operation over basic domains.

He tried to meet these three main criticisms with by proposing to consider the logical operations as precisely the homomorphism-invariant operations. Indeed, it seemed clear that, by ruling out the numerical quantifiers, Feferman's proposal avoided assimilating logic to set theory, thus ruling out 1. The notions captured were shown to be exactly those defined by the first-order predicate calculus without identity, which, as shown by Väänänen (1985),

are absolute even under the theory Kripke-Platek with urelements and without the axiom of infinity (KPU-inf). Finally, given his functional-type framework, he was able to give a uniform account of, e.g., the operation determined by conjunction, which will be uniformly the same operation from truth-values to truth-values.

Unfortunately, setting aside the first criticism, we have seen that matters are not so simple. As shown by Casanovas, Feferman's theorem is highly sensitive to which type-theoretical framework one adopts. In particular, if one adopts a *relational*-type framework, it's not clear that Feferman is still able to retain 3, as, e.g., conjunction acting on formulas without common variables will be mapping-invariant, but conjunction acting on formulas with common variables will not in general be mapping-invariant, which shows that even on the same domain the operations determined by this connective will not be the same. Moreover, while the semantics of first-order logic without equality are set-theoretically absolute, it's not clear that they are robust, given this wild disparity in the results depending on whether one adopts a functional- or a relational-type framework. Therefore, it seems that, by his own lights, Feferman's proposal fails as a good logicality criterion.

Conclusion

In the first part, I presented a historical reconstruction of Carnap's, Quine's, and Tarski's differing, and sometimes evolving, views on activity that can be broadly called *explication*. This activity is central to the philosophical output of those three, which is why I decided to give it center stage in the beginning of my analysis. In fact, given how much ink has been spilled on whether or not Tarski's proposal captures our "intuitive" concepts of truth, consequence, and logicality, focusing on Carnap's and Quine's dismissal of such appeals to "intuition" seemed to me a good strategy for making a fresh start on this debate. Nevertheless, there remains much to be done in this connection, be it on the historical, technical, or philosophical front.

On the historical side, there are least two issues that I believe are worthy of further exploration. First, Tarski's environment during his formative years should be better explored, in particular his connection with other Polish logicians such as Ajdukiewicz. In this same vein, the Polish school's ties with more "mainstream" figures, especially Frege, could also be more developed. This research would give us a fuller picture of what Arianna Betti (2008) has called the Classical Ideal of Science, which guided logical research in the 19th century and in the early days of the 20th, before Hilbert's modern position took hold. The second historical point that I believe also merits attention is the connection between Tarski, on the one hand, and Carnap and Quine, on the other. In this study, I limited myself to a few comparisons, but I believe a case for direct influence on specific points, e.g., Quine and Goodman's nominalist strategy, could be easily built, particularly in light of the new archival material made available by Frost-Arnold (2013).

On the technical side, it would perhaps be interesting to see how far Tarski's substitutional strategy could be made to work, and to compare the strength of the resulting theory with the simple theory of types. Burgess and Rosen (1997, pp. 183–5) seem to imply that Tarski's proposal is merely a rewritten version of the simple theory of types, but it would be nice to have a more detailed examination of this result. It would be also of interest to see what happens when one tinkers with the framework to better fit Tarski's finitist strictures, e.g, by considering only a finite set of expressions at the start.

Finally, on the more philosophical side, there remains the issue of examining from a

closer perspective the merits of each type of explication project. I have already indicated in Chapter 2 some of my worries regarding Quine's "paraphrase" method, but I have not done so for Carnap as well. In a footnote, I indicated that a comparison with Mark Wilson's view in *Wandering Significance* was called for, which would require a more extensive treatment of particular case studies from the history of mathematics. Similarly, comparison with Shapiro's development of Waismann's concept of open texture and how informal concepts can give rise to different sharpenings would make for some nice contrastive analysis.

More importantly, however, is how certain considerations from the second part of the study interact with the first part. The Carnapian position outlined in the first part implies that the activity of concept creation is a largely *voluntaristic* process, in which fruitfulness is the only measure of success. However, the second part of the study outlined a different picture. By introducing the notion of a property or an object being *natural*, I thereby also opened the possibility for another, more platonic measure of success, namely whether the concept thus created is also in some derivative sense natural, that is, if it tracks a natural property or object. Of course, these two metrics need not come apart. Indeed, they often go together, so much so that Tappenden (2008a,b) even proposed to *identify* the two, by simply equating naturalness with fruitfulness. As should be clear by my remarks in the third chapter, I reject this identification. For me, naturalness should be explained in each case by an appeal to *intrinsic* properties of the theory or structure in question (Manders's *conceptual settings*), following the example set forth by Manders's analysis of domain extension. Therefore, in my view, naturalness is not identical with fruitfulness, but *explains* fruitfulness.

In any case, this idea that our activity of concept creation is constrained by naturalness would go some way towards limiting the voluntaristic aspects of Carnapian (and Quinean, for the matter) explication. This would also explain why the activity of defining new mathematical concepts is so important and at the same time so difficult. Two examples: Wussing (2007) shows how the emergence of the abstract group concept was a rather slow process, that depended on the development of at least three different areas, namely geometry, number theory, and algebra, in the form of the theory of polynomial equations. It is interesting to see how distinguished authors such as Lagrange, Euler, Gauss, and others were groping towards this concept, proving special cases of results that are much more *naturally* formulated using group-theoretic concepts. Indeed, that is why their names ended up attached to results from a subject matter that was created only after their death, such as the famous Lagrange Theorem on the relation between the order of a group and the order of its subgroups. Not surprisingly, when it finally emerged, the abstract group concept proved to be fundamental for all these areas. In a somewhat different direction, and this is the second example, it would be interesting to see how concepts that did *not* naturally belong together were eventually refined into distinct notions, such as the notions of continuity, differentia-

bility, convergence, uniform convergence, etc. Grabiner's (1981/2005) book is a rich source in this regard.

These historical examples are also interesting in limiting the appeal to "intuitions" in this kind of context. If there is something to be learned from these two examples, is that natural concepts are anything but "intuitive" (in fact, mathematicians' intuitions about continuity and convergence famously led them astray numerous times). Rather, they are obtained after a painstaking labor of unearthing natural objects and properties. My hope is that a more in depth study of these historical examples could help to reveal some of the features that make those concepts so natural.

Moving on to the second part of the study, since this part is more opinionated and less historical, it also contain more loose ends. Two of them I indicated in Chapter 3 itself: a more thorough investigation of neo-Fregean abstractionism and a formal account of "analytic" that could be useful in building a bridge between logical objects and logical constants. I believe that the latter will prove to be a mere exercise in formal semantics, since the outline of such an account has been more than adequately provided by Lewis. The former, on the other hand, is more difficult, since the nature of abstraction principles is still very controversial. First, given the so-called bad company objection, there is the pressing question of which such principles are natural in a metaphysical sense. Some global technical constraints have been outlined in this connection by Fine (2002), but are there any other more metaphysical ones? And how do these constraints interact with the broadly Kleinian framework outlined here?

Second, there is also the question of the nature of abstract objects. Typically, philosophers simply assume that the nature of this division between concrete and abstract objects is well understood, but this division is not so clear to me.[37] As Heck (2017) suggests, the neo-Fregean may have a promising line here, but this account needs to be fleshed out, and objections such as those raised by Lewis (1986, p. 85) should be met. Perhaps a generalization of my proposal for types could work here, but, again, there is much to be investigated in this connection.

In this regard, another question that arises in the interaction between the first and the second part is about the *essence* of mathematical objects. If we conceive the essence of an object to be simply the set of its necessary properties, then every property of a mathematical object will presumably be an essential property. However, at least since Fine (1994), philosophers have begun to search for a more robust conception of essence, one that is somehow connected to the way that a thing's essence *constitutes* that thing. Perhaps this could be tied to the explanation I gave of natural properties in the text. Suppose, as is somewhat plau-

[37] For early doubts about this, cf. Hale (1987, Chap. 3); for a more recent discussion, including a very useful bibliography, cf. Cowling (2017, Chap. 2).

sible, that mathematical objects are *types*.[38] The essential properties of the type would be those that figured in the explanatory principle connected to the type, in such a way that its other properties somehow followed from it. Thus, the naturalness of some concepts could be thought of as following from the fact that such concepts captured the *essence* of the type in question. The continuous refinement of mathematical definitions, which I briefly described a couple of paragraphs back, would be the search for *real definitions*. But this is all speculation. As I said in the previous paragraph, there is a lot of work to do before this is even plausible.

Third, there are questions related to Klein's own framework. In one direction, even in the 19th century, it was already noted that Klein's proposal was limited, since it couldn't deal with Riemannian geometry. There are ways of extending his approach to include such cases,[39] and an investigation into those approaches could reveal something philosophically fruitful. On another direction, Marquis (2009) contains some intriguing suggestions, especially on the connections between logic, geometry, and category theory that I unfortunately could not investigate further here.

On a more technical level, I mentioned in the last chapter that Bonnay's work couldn't be included in this study; still, his work does raise some interesting mathematical questions, in particular about the connections between logic and Galois theory that I believe could be fruitfully pursued. Moreover, due to a mistake, his work still leaves open the question of which logic corresponds exactly to the homomorphism invariant operations. There is some rather strong evidence that $\mathcal{L}_{\infty,\infty}$ *without* identity is such a logic, but no proof of this has yet been published.

Finally, there is also the work of van Benthem mentioned by Dutilh Novaes. Roderick Batchelor has in recent years conducted extensive research into extending the theory of logical operations to a modal framework. Another fruitful direction of research would be to extend Batchelor's work and compare it with earlier results obtained by van Benthem, in particular in connection with the latter's use of the bisimulation technique.

[38] This would be a position close to some forms of structuralism, in the sense that the "positions" in the structure would be considered as types. In fact, structures themselves could be considered as isomorphism types.

[39] Notably, Cartan's proposal. Cf. Sharpe (1997) for a detailed mathematical treatment of Cartan's ideas.

References

Amiot, Emmanuel. *Music Through Fourier Space: Discrete Fourier Transform in Music Theory*. New York: Springer, 2016.

Arnol'd, Vladimir I. "On Teaching Mathematics". In: *Russian Mathematical Surveys* 53 (1998), pp. 229–236.

Barnard, Robert and Joseph Ulatowski. "Tarski's 1944 Polemical Remarks and Naess' "Experimental Philosophy"". In: *Erkenntnis* 81.3 (2016), pp. 457–477.

Barwise, K. Jon. *Admissible Sets and Structures*. New York: Springer-Verlag, 1975.

Beaney, Michael, ed. *The Analytic Turn: Analysis in Early Analytic Philosophy and Phenomenology*. New York: Routledge, 2007.

Bell, John Lane and A. B. Slomson. *Models and Ultraproducts: An Introduction*. New York: North Holland, 1974.

Bellotti, Luca. "Tarski on Logical Notions". In: *Synthese* 135 (2003), pp. 401–413.

Beth, Evert Willem. "Reason and Intuition". In: *Aspects of Modern Logic*. Dordrecht-Holland: D. Reidel Publishing Company, 1970. Chap. 7, pp. 86–101.

Betti, Arianna. "Leśniewski's early Liar, Tarski and natural language". In: *Annals of Pure and Applied Logic* 127 (2004), pp. 267–287.

— "Polish Axiomatics and its Truth: On Tarski's Leśniewskian Background and the Ajdukiewicz Connection". In: *New Essays on Tarski and Philosophy*. Ed. by Douglas Patterson. Oxford: Oxford University Press, 2008, pp. 44–71.

Birkhoff, Garrett and M. K. Bennett. "Felix Klein and his "Erlanger Programm"". In: *History and Philosophy of Modern Mathematics*. Ed. by Philip Kitcher and William Aspray. Minneapolis: University of Minnesota Press, 1988, pp. 145–176.

Blanchette, Patricia A. *Frege's Conception of Logic*. Oxford: Oxford University Press, 2012.

Bonnay, Denis. "Logicality and Invariance". In: *Bulletin of Symbolic Logic* 13, 1 (2008), pp. 29–68.

Brandom, Robert. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge: Harvard University Press, 2000.

— *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: Oxford University

Press, 2008.

Bromberger, Sylvain. *On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. Chicago: University of Chicago Press, 1992.

Burgess, Alexis G. and John Burgess. *Truth*. Princeton: Princeton University Press, 2011.

Burgess, John P. and Gideon Rosen. *A Subject with no Object: Strategies for Nominalist Interpretations of Mathematics*. Oxford: Oxford University Press, 1997.

Carnap, Rudolf. *Logical Syntax of Language*. London: Routledge, 1937/2001.

— "Empiricism, Semantics, and Ontology". In: *Meaning and Necessity*. 2nd ed. Chicago: University of Chicago Press, 1950/1956, pp. 205–221.

— "Meaning and Synonymy in Natural Languages". In: *Philosophical Studies* VI.3 (1955), pp. 33–47.

— "Value Concepts". In: *Synthese* 194 (1958/2015). Edited and translated by André W. Carus, pp. 185–194.

— *Logical Foundations of Probability*. 2nd ed. Chicago: The University of Chicago Press, 1962.

— "Replies and Systematic Expositions". In: *The Philosophy of Rudolf Carnap*. Ed. by Paul Arthur Schilpp. Chicago: Open Court, 1963, pp. 859–1013.

Carus, André W. *Carnap and Twentieth-Century Thought: Explication as Enlightenment*. New York: Cambridge University Press, 2007.

— "Engineers and Drifters: The Ideal of Explication and Its Critics". In: *Carnap's Ideal of Explication and Naturalism*. Ed. by Pierre Wagner. New York: Palgrave Macmillan, 2012, pp. 225–239.

— "Carnapian Rationality". In: *Synthese* 194 (2017), pp. 163–184.

Casanovas, Enrique. "Logical Operations and Invariance". In: *Journal of Philosophical Logic* 36 (2007), pp. 33–60.

Casanovas, Enrique, Pillar Dellunde, and Ramon Jansana. "On Elementary Equivalence for Equality-free Logic". In: *Notre Dame Journal of Formal Logic* 37.3 (1996), pp. 506–522.

Coffa, J. Alberto. *The Semantic Tradition from Kant to Carnap: To The Vienna Station*. Ed. by Linda Wessels. Cambridge: Cambridge University Press, 1991.

Corcoran, John. "Material Adequacy". In: *The Cambridge Dictionary of Philosophy*. Ed. by Rober Audi. Cambridge: Cambridge University Press, 1999, p. 540.

— "Review of Sinaceur 2009". In: *Mathematical Reviews* (2011). MR2509665 (2011b:03006).

Corcoran, John and José Miguel Sagüilo. "The Abscence of Multiple Universes of Discourse in the 1936 Tarski Consequence-Definition Paper". In: *History and Philosophy of Logic* 32.4 (2011), pp. 359–374.

Corcoran, John and Leonardo Weber. "Tarski's Convention T: Condition beta". In:

South American Journal of Logic 1.1 (2015), pp. 3–32.

Corcoran, John, William Frank, and Michael Maloney. "String Theory". In: *The Journal of Symbolic Logic* 39.4 (1974), pp. 625–637.

Cowling, Sam. *Abstract Entities*. London: Routledge, 2017.

Creath, Richard, ed. *Dear Carnap, Dear Van: The Carnap-Quine Correspondence and Related Work*. Berkeley, CA: University of California Press, 1990.

— "Every Dogma Has Its Day". In: *Erkenntnis* 35 (1991), pp. 347–389.

— "The Linguistic Doctrine and Conventionality: The Main Argument in "Carnap and Logical Truth"". In: *Logical Empiricism in North America*. Ed. by Gary L. Hardcastle and Alan W. Richardson. Minneapolis: University of Minnesota Press, 2003, pp. 234–256.

— "Quine's Challenge to Carnap". In: *The Cambridge Companion to Carnap*. Ed. by Michael Friedman and Richard Creath. Cambridge: Cambridge University Press, 2007, pp. 316–335.

— "Understandability". In: *Metascience* (2015). Online first.

— "The Logical and the Analytic". In: *Synthese* 194 (2017), pp. 79–96.

de Rouilhan, Philippe. "Carnap on Logical Consequence for Languages I and II". in: *Carnap's Logical Syntax of Language*. Ed. by Pierre Wagner. New York: Palgrave Macmillan, 2009, pp. 121–146.

Dedekind, Julius Wilhelm Richard. "Continuity and Irrational Numbers". In: *Essays on the Theory of Numbers*. New York: Dover, 1872/1963, pp. 1–27.

— *Gessamelte Mathematischen Werke, Volumes 1-3*. Braunschweig: Vieweg, 1932.

Demopoulos, William. "Carnap's Analysis of Realism". In: *Logicism and Its Philosophucal Legacy*. New York: Cambridge University Press, 2013, pp. 68–89.

Devlin, Keith J. *Constructibility*. Berlin: Springer-Verlag, 1984.

Dutilh Novaes, Catarina. "The Undergeneration of Permutation Invariance as a Criterion for Logicality". In: *Erkenntnis* 79 (2014), pp. 81–97.

Dutilh Novaes, Catarina and Edgar Andrade-Lotero. "Validity, the Squeezing Argument and Alternative Semantic Systems: the Case of Aristotelian Syllogistic". In: *Journal of Philosophical Logic* 41 (2012), pp. 387–418.

Dutilh Novaes, Catarina and Erich Reck. "Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization". In: *Synthese* (2015). Published online.

Ebbinghaus, Hans-Dieter. "Extended Logics: The General Framework". In: *Model-Theoretic Logics*. Ed. by K. Jon Barwise and Solomon Feferman. New York: Springer-Verlag, 1985, pp. 25–76.

Ebbs, Gary. "Quine's Naturalistic Explication of Carnap's Logic of Science". In: *A Com-*

panion to W. V. O. Quine. Ed. by Gilbert Harman and Ernest Lepore. West Sussex: Wiley-Blackwell, 2014, pp. 465–482.

Ereshefsky, Marc. *The Poverty of the Linnaean Hierarchy: A Philosophical Study of Biological Taxonomy*. Cambridge: Cambridge University Press, 2003.

— "Systematics and Taxonomy". In: *A Companion to the Philosophy of Biology*. Ed. by Sahotra Sarkar and Anya Plutynski. Oxford: Wiley-Blackwell, 2008, pp. 99–118.

Etchemendy, John. *The Concept of Logical Consequence*. Stanford: Center for the Study of Language and Information, 1999.

— "Reflections on Consequence". In: *New Essays on Tarski and Philosophy*. Ed. by Douglas Patterson. Oxford: Oxford University Press, 2008, pp. 263–299.

Feferman, Solomon. "Applications of many-sorted interpolation theorems". In: *Proceedings of the Tarski Symposium*. Ed. by Leon Henkin. Providence: American Mathematical Society, 1974, pp. 205–223.

— "Logic, Logics, and Logicism". In: *Notre Dame Journal of Formal Logic* 40 (1999), pp. 31–54.

— "Set-theoretical invariance criteria for logicality". In: *Notre Dame Journal of Formal Logic* 51 (2010), pp. 3–20.

Fernández-Moreno, Luis. "Tarski y la noción carnapiana de significado". In: *Revista de Filosofia* VII (1994), pp. 403–420.

Field, Hartry. "Tarski's Theory of Truth". In: *Truth and the Abscence of Fact*. Oxford: Oxford University Press, 1972/2001, pp. 1–26.

— *Saving Truth from Paradox*. Oxford: Oxford University Press, 2008.

— *Science Without Numbers*. 2nd ed. Oxford: Oxford University Press, 2016.

Fine, Kit. "Essence and Modality". In: *Philosophical Perspectives* 8 (1994), pp. 1–16.

— *The Limits of Abstraction*. Oxford: Oxford University Press, 2002.

— "Guide to Ground". In: *Metaphysical Grounding: Understanding the Structure of Reality*. Ed. by Fabrice Correia and Benjamin Schnieder. Cambridge: Cambridge University Press, 2012, pp. 37–80.

Frege, Gottlob. *The Foundations of Arithmetic: A Logico-mathematical Enquiry into the Concept of Number*. 2nd ed. Translated by J. L. Austin. New York: Harper Torchbooks, 1884/1960.

French, Christopher Forbes. "Philosophy as Conceptual Engineering: Inductive Logic in Rudolf Carnap's Scientific Philosophy". PhD thesis. University of British Columbia, 2015.

Friedman, Michael. "Geometry as a Branch of Physics". In: *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*. Ed. by David Malament. Chicago: Open Court, 2002, pp. 193–230.

— "Carnap and Quine: Twentieth-Century Echoes of Kant and Hume". In: *Philosophical Topics* 34 (2006), pp. 35–58.
— "Carnap's Philosophical Neutrality Between Realism and Instrumentalism". In: *Analysis and Interpretation in the Exact Sciences: Essays in Honour of William Demopoulos*. Ed. by Melanie Frappier, Derek Brown, and Robert DiSalle. Dordrecht: Springer, 2012, pp. 95–114.
— "Scientific Philosophy from Helmholtz to Carnap and Quine". In: *Rudolf Carnap and the Legacy of Logical Empiricism*. Ed. by Richard Creath. Vienna Circle Institute Yearbook 16. Dordrecht: Springer, 2012, pp. 1–11.
Frost-Arnold, Greg. *Carnap, Tarski, and Quine at Harvard: Conversations on Logic, Mathematics, and Science*. Chicago, Illinois: Open Court, 2013.
— "Replies to Creath, Ebbs, and Lavers". In: *Metascience* (2015). Online first.
Gawroński, Alfred. "Psychologism and the principle of relevance in semantics". In: *Kotarbiński: Logic, Semantics and Ontology*. Ed. by Jan Woleński. Dordrecht: Kluwer Academic Publishers, 1990, pp. 23–29.
Goodman, Nelson. *Fact, Fiction, and Forecast*. 4th ed. Cambridge, Massachusetts: Harvard University Press, 1983.
Grabiner, Judith V. *The Origin's of Cauchy's Rigorous Calculus*. New York: Dover, 1981/2005.
Gray, Jeremy J. "Poincaré and Klein - Groups and Geometries". In: *1830-1930: A Century of Geometry*. Ed. by Luciano Boi, Dominique Flament, and Jean-Michel Salanskis. Berlin: Springer, 1992, pp. 35–44.
— *Worlds Out of Nothing: A Course in the History of Geometry in the 19th Century*. London: Springer, 2011.
Gruber, Monika. *Alfred Tarski and the "Concept of Truth in Formalized Languages": A Running Commentary of the Polish Original and the German Translation*. New York: Springer, 2016.
Guigon, Ghislain and Gonzalo Rodriguez-Pereyra, eds. *Nominalism About Properties: New Essays*. New York: Routledge, 2015.
Gustafsson, Martin. "Quine's Concept of Explication —and Why It Isn't Carnap's". In: *A Companion to W. V. O. Quine*. Ed. by Gilbert Harman and Ernest Lepore. West Sussex: Wiley Blackwell, 2014. Chap. 24, pp. 508–525.
Haddock, Guillermo E. Rosado. *Against the Current: Selected Philosophical Papers*. Frankfurt a. M.: Ontos Verlag, 2012.
Halbach, Volker. *Axiomatic Theories of Truth*. New York: Cambridge University Press, 2011.
Hale, Bob. *Abstract Objects*. Oxford: Basil Blackwell, 1987.
Hamkins, Joel David, David Linetsky, and Jonas Reitz. "Pointwise Definable Models of

Set Theory". In: (2012). arXiv:1105.4597 [math.LO].

Hawkins, Thomas. "The *Erlanger Programm* of Felix Klein: Reflections on Its Place in the History of Mathematics". In: *Historia Mathematica* 11 (1984), pp. 442–470.

— *Emergence of the Theory of Lie Groups: An Essay in the History of Mathematics, 1869-1926*. New York: Springer, 2000.

Heck Jr., Richard G. "Syntactic Reductionism". In: *Frege's Theorem*. Oxford: Oxford University Press, 2011, pp. 180–199.

— "The Existence (and Non-existence) of Abstract Objects". In: *Abstractionism: Essays in the Philosophy of Mathematics*. Ed. by Philip A. Ebert and Marcus Rossberg. Oxford: Oxford University Press, 2017, pp. 50–78.

Hirsch, Eli. *Dividing Reality*. Oxford: Oxford University Press, 1993.

Hodges, Wilfrid. "Truth in a Structure". In: *Proceedings of the Aristotelian Society* 86, N. S. (1986), pp. 135–151.

— *Model Theory*. Cambridge: Cambridge University Press, 2004.

Horsten, Leon. *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, Massachusetts: MIT Press, 2011.

Hylton, Peter. *Quine*. London: Routledge, 2007.

Jané, Ignacio. "What is Tarski's Common Concept of Consequence?" In: *The Bulletin of Symbolic Logic* 12.1 (2006), pp. 1–42.

Jeffrey, Richard. "Carnap's Voluntarism". In: *Logic, Methodology, and Philosophy of Science IX*. ed. by Dag Prawitz, Brian Skyrms, and Dag Westerståhl. Amsterdam: Elsevier, 1994, pp. 847–866.

Jensen, Ronald B. and Carol Karp. "Primitive Recursive Set Functions". In: *Axiomatic Set Theory: Proceedings of Symposia in Pure Mathematics*. Ed. by Dana S. Scott. Vol. 1. Providence: American Mathematical Society, 1971, pp. 143–176.

Kanamori, Akihiro. "The Empty Set, the Singleton, and the Ordered Pair". In: *The Bulletin of Symbolic Logic* 9 (2003), pp. 273–298.

Kant, Immanuel. *Critique of Pure Reason*. Translated and edited by Paul Guyer and Allen Wood. New York: Cambridge University Press, 1998.

Klein, Christian Felix. "A Comparative Review of Recent Researches in Geometry". In: *Bulletin of New York Mathematical Society* 2 (1892-1893). Translated by M. W. Haskell, pp. 215–249.

Klement, Kevin C. "A Generic Russellian Elimination of Abstract Objects". In: *Philosophia Mathematica* 25 (2017), pp. 91–115.

Kline, Morris. *Mathematical Thought from Ancient to Modern Times*. New York: Oxford University Press, 1972.

Kotarbińska, Janina. "Puzzles of Existence". In: *Kotarbiński: Logic, Semantics and Ontology*.

Ed. by Jan Woleński. Dordrecht: Kluwer Academic Publishers, 1990, pp. 53–67.

Kreisel, Georg. "Informal rigour and completeness proofs". In: *Problems in the Philosophy of Mathematics*. Ed. by Imre Lakatos. Amsterdam: North Holland, 1967, pp. 138–171.

Kripke, Saul. "Is There a Problem about Substitutional Quantification?" In: *Truth and Meaning*. Ed. by Gareth Evans and John McDowell. New York: Oxford University Press, 1976, pp. 325–419.

Lange, Marc. "Explanation, Existence, and Natural Properties in Mathematics – A Case Study: Desargues' Theorem". In: *Dialectica* 69.4 (2015), pp. 435–472.

— *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford: Oxford University Press, 2017.

Leśniewski, Stanisław. "An Attempt at a Proof of the Ontological Principle of Contradiction". In: *Collected Works*. Ed. by S. J. Surma et al. Dordrecht: Kluwer Academic Publishers, 1992, pp. 20–46.

— "Fundamentals of a New System of the Foundations of Mathematics". In: *Collected Works*. Ed. by S. J. Surma et al. Dordrecht: Kluwer Academic Publishers, 1992, pp. 410–605.

— "On the Foundations of Mathematics". In: *Collected Works*. Ed. by S. J. Surma et al. Dordrecht: Kluwer Academic Publishers, 1992, pp. 174–382.

— "The Critique of the Logical Principle of the Excluded Middle". In: *Collected Works*. Ed. by S. J. Surma et al. Dordrecht: Kluwer Academic Publishers, 1992, pp. 47–85.

Lewis, David. "General Semantics". In: *Philosophical Papers*. Vol. 1. New York: Oxford University Press, 1970/1983, pp. 189–232.

— "Language and Languages". In: *Philosophical Papers*. Vol. 1. New York: Oxford University Press, 1983, pp. 163–188.

— *On the Plurality of Worlds*. New York: Basil Blackwell, 1986.

MacFarlane, John. "What does it mean to say that logic is formal?" PhD thesis. University of Pittsburgh, 2000.

Makowski, J. A., Saharon Shelah, and Jonathan Stavi. "Δ-logics and Generalized Quantifiers". In: *Annals of Mathematical Logic* 10 (1976), pp. 155–192.

Mancosu, Paolo. "Fixed- versus Variable-domain Interpretations of Tarski's Account of Logical Consequence". In: *Philosophy Compass* 5.9 (2010), pp. 745–759.

— "Harvard 1940-1941: Tarski, Carnap, and Quine on a Finitistic Language of Mathematics for Science". In: *The Adventure of Reason: Interplay between Philosophy of Mathematics and Mathematical Logic, 1900-1940*. New York: Oxford University Press, 2010. Chap. 13, pp. 361–386.

— "Quine and Tarski on Nominalism". In: *The Adventure of Reason: Interplay between Philosophy of Mathematics and Mathematical Logic, 1900-1940*. New York: Oxford University

Press, 2010, pp. 387–409.

Mancosu, Paolo. "Tarski, Neurath, and Kokoszyńska on the Semantic Conception of Truth". In: *The Adventures of Reason: Interplay between Philosophy of Mathematics and Mathematical Logic, 1900-1940*. New York: Oxford University Press, 2010, pp. 415–439.

— *Abstraction and Infinity*. Oxford: Oxford University Press, 2016.

Manders, Kenneth. "Logic and Conceptual Relationships in Mathematics". In: *Logic Collquium '85*. Amsterdam: Elsevier, 1987, pp. 193–211.

— "Domain Extensions and the Philosophy of Mathematics". In: *The Journal of Philosophy* 86 (1989), pp. 553–562.

Marquis, Jean-Pierre. *From a Geometrical Point of View: A Study of the History and Philosophy of Category Theory*. Dordrecht: Springer, 2009.

Mautner, F. I. "An Extension of Klein's Erlanger Program: Logic as Invariant-Theory". In: *American Journal of Mathematics* 68 (1946), pp. 345–384.

McGee, Vann. "Logical Operations". In: *Journal of Philosophical Logic* 25 (1996), pp. 567–80.

— "Tarski's Staggering Existential Assumptions". In: *Synthese* 142 (2004), pp. 371–387.

Mitchell, John C. "Type Systems for Programming Languages". In: *Handbook of Theoretical Computer Science*. Ed. by Jan van Leeuwen. Vol. B, Formal Models and Semantics. Amsterdam: Elsevier, 1990, pp. 365–458.

Patterson, Douglas. *Alfred Tarski: Philosophy of Language and Logic*. New York: Palgrave Macmillan, 2012.

Psillos, Stathis. *Scientific Realism: How Science Tracks Truth*. New York: Routledge, 1999.

Quine, Willard Van Orman. "Truth by Convention". In: *The Ways of Paradox and Other Essays*. New York: Random House, 1935/1966, pp. 70–99.

— "On Carnap's Views on Ontology". In: *Quintessence: Basic Readings from the Philosophy of W. V. Quine*. Ed. by Roger F. Gibson. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1951/2004, pp. 249–256.

— "Two Dogmas of Empiricism". In: *Quintessence: Basic Readings from the Philosophy of W. V. Quine*. Ed. by Roger F. Gibson. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1952/2004, pp. 31–53.

— "The Scope and Language of Science". In: *Quintessence: Basic Readings from the Philosophy of W. V. Quine*. Ed. by Roger F. Gibson. Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 1954/2001, pp. 193–209.

— "Carnap and Logical Truth". In: *The Philosophy of Rudolf Carnap*. Ed. by Paul Arthur Schilpp. Chicago: Open Court, 1963, pp. 385–406.

— "Carnap's Positivistic Travail". In: *Quine in Dialogue*. Ed. by Dagfinn Føllesdal and

Douglas B. Quine. Cambridge, Massachusetts: Harvard University Press, 1984/2008, pp. 119–128.

— *Word and Object*. 2nd ed. Cambridge, Massachusetts: MIT Press, 2013.

Quine, Willard Van Orman and Nelson Goodman. "Steps Towards a Constructive Nominalism". In: *The Journal of Symbolic Logic* 12 (1947), pp. 105–122.

Raatikainen, Panu. "More on Putnam and Tarski". In: *Synthese* 135.1 (2003), pp. 37–47.

Raven, Charles. *John Ray, Naturalist: His Life and Works*. Cambridge: Cambridge University Press, 2009.

Ray, Greg. "Tarski and the Metalinguistic Liar". In: *Philosophical Studies* 115 (2003), pp. 55–80.

Reck, Erich. "Dedekind's Structuralism: An Interpretation and Partial Defense". In: *Synthese* 137.3 (2003), pp. 369–419.

Reid, Miles and Balázs Szendrői. *Geometry and Topology*. Cambridge: Cambridge University Press, 2005.

Richardson, Alan. "Two Dogmas about Logical Empiricism: Carnap and Quine on Logic, Epistemology, and Empiricism". In: *Philosophical Topics* 25.2 (1997), pp. 145–168.

— "Taking the Measure of Carnap's Philosophical Engineering: Metalogic as Metrology". In: *The Historical Turn in Analytic Philosophy*. Ed. by Erich Reck. New York: Palgrave Macmillan, 2013, pp. 60–77.

Ricketts, Thomas. "Frege, Carnap, and Quine: Continuities and Discontinuities". In: *Carnap Brought Home: The View from Jena*. Ed. by Steve Awodey and Carsten Klein. Chicago: Open Court, 2004, pp. 181–202.

— "From Tolerance to Reciprocal Containment". In: *Carnap's Logical Syntax of Language*. Ed. by Pierre Wagner. New York: Palgrave Macmillan, 2009, pp. 217–235.

Robinson, Abraham. *Introduction to Model Theory and to the Metamathematics of Algebra*. 2nd ed. Amsterdam: North Holland, 1965.

Rodríguez-Consuegra, Francisco. "Tarski's Intuitive Notion of Set". In: *Essays on the Foundations of Mathematics and Logic*. Ed. by G. Sica. Monza: Polimerca, 2005, pp. 227–266.

Rodriguez-Pereyra, Gonzalo. *Resemblance Nominalism: A Solution to the Problem of Universals*. Oxford: Oxford University Press, 2002.

Rowe, David E. "Klein, Lie, and the "Erlanger Programm"". In: *1830-1930: A Century of Geometry*. Ed. by Luciano Boi, Dominique Flament, and Jean-Michel Salanskis. Berlin: Springer, 1992, pp. 45–54.

Shapiro, Stewart. "Computability, Proof, and Open-Texture". In: *Church's Thesis After 70 Years*. Ed. by Adam Olszewski, Jan Woleński, and Robert Janusz. Franfurt a. M.:

Ontos Verlag, 2006, pp. 420–455.

Sharpe, Richard W. *Differential Geometry: Cartan's Generalization of Klein's Erlangen Program*. New York: Springer, 1997.

Sher, Gila. "Tarski's Thesis". In: *New Essays on Tarski and Philosophy*. Ed. by Douglas Patterson. New York: Oxford University Press, 2008, pp. 300–339.

Simpson, Stephen G. "Short Course on Admissible Recursion Theory". In: *Generalized Recursion Theory II*. ed. by J. E. Fenstad, R. O. Gandy, and G. E. Sacks. New York: North Holland, 1978, pp. 355–390.

Sinaceur, Hourya Benis. "Tarski's Practice and Philosophy: Between Formalism and Pragmatism". In: *Logicism, Intuitionism, and Formalism: What Has Become of Them?* Ed. by Sten Lindström et al. Dordrecht: Springer, 2009, pp. 357–396.

Smid, Jeroen. "Tarski's one and only concept of truth". In: *Synthese* 191 (2014), pp. 3393–3406.

Smith, Peter. "Squeezing Arguments". In: *Analysis* 71 (2011), pp. 22–30.

— *An Introduction to Gödel's Theorems*. 2nd ed. Cambridge: Cambridge University Press, 2013.

Soames, Scott. *Understanding Truth*. Oxford: Oxford University Press, 1999.

— "The Place of Quine in Analytic Philosophy". In: *A Companion to W. V. O. Quine*. Ed. by Ernest Lepore and Gilbert Harman. West Sussex: Wiley Blackwell, 2014, pp. 432–464.

Stein, Howard. "Some Philosophical Prehistory of General Relativity". In: *Foundations of Space-Time Theories*. Ed. by John Earman, Clark Glymour, and John Statchel. Minneapolis: University of Minnesota Press, 1977, pp. 3–49.

— "Was Carnap Entirely Wrong, After All?" In: *Synthese* 93.1/2 (1992), pp. 275–295.

Sundholm, Göran. "Tarski and Leśniewski on Languages with Meaning versus Languages without Use". In: *Philosophy and Logic, in Search of the Polish Tradition: Essays in Honour of Jan Woleński on the Occasion of his 60th Birthday*. Ed. by Jaakko Hintikka et al. Dordrecht: Kluwer Academic Publishers, 2003, pp. 109–128.

Tappenden, Jamie. "Mathematical Concepts and Definitions". In: *The Philosophy of Mathematical Practice*. Ed. by Paolo Mancosu. Oxford: Oxford University Press, 2008, pp. 256–275.

— "Mathematical Concepts: Fruitfulness and Naturalness". In: *The Philosophy of Mathematicl Practice*. Ed. by Paolo Mancosu. Oxford: Oxford University Press, 2008, pp. 276–301.

Tarski, Alfred. "Fundamental Concepts of the Methodology of the Deductive Sciences". In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. 2nd ed. Translated by J. H. Woodger. Indianapolis, Indiana: Hackett Pub., 1930/1983. Chap. 5, pp. 60–109.

— "Sur les ensembles définissables de nombres réels". In: *Fundamenta Mathematicae* 17.1 (1931), pp. 210–239.
— "On Definable Sets of Real Numbers". In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. Translated by J. H. Woodge. Indianapolis, Indiana: Hackett Pub., 1931/1983. Chap. VI, pp. 110–142.
— "The Concept of Truth in Formalized Languages". In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. 2nd ed. Translated by J. H. Woodger. Indianapolis, Indiana: Hackett Pub., 1933/1983. Chap. VIII, pp. 152–278.
— "On the Concept of Following Logically". In: *History and Philosophy of Logic* 23 (1936/2002). Translated by Magda Stroińska and David Hitchcock, pp. 155–196.
— *Introduction to Logic and to the Methodology of the Deductive Sciences*. New York: Oxford University Press, 1941.
— "The Semantic Concept of Truth and the Foundations of Semantics". In: *Philosophy and Phenomenological Research* 4.3 (1944), pp. 341–376.
— "A General Theorem Concerning Primitive Notions of Euclidean Geometry". In: *Indagationes Mathematicae* 18 (1956). Proceedings Series, pp. 468–474.
— "What are Logical Notions?" In: *History and Philosophy of Logic* 7 (1966/1986). Transcription of a 1966 lecture, ed. John Corcoran, pp. 143–154.
— "Truth and Proof". In: *Scientific American* 220 (1969), pp. 63–70, 75–77.
— "On the Concept of Logical Consequence". In: *Logic, Semantics, Metamathematics*. Ed. by John Corcoran. 2nd ed. Translated by J. H. Woodger. Indianapolis, Indiana: Hackett Pub., 1983. Chap. XVI, pp. 409–420.
— "Two Unpublished Contributions by Alfred Tarski". In: *History and Philosophy of Logic* 28 (2007). Transcription of two 1965 lectures, ed. Francisco Rodriguez-Consuegra, pp. 257–264.
Tarski, Alfred and Steven Givant. *A Formalization of Set Theory Without Variables*. Providence, Rhode Island: American Mathematical Society, 1988.
Tarski, Alfred and Adolf Lindenbaum. "Sur l'indépendance des notions primitives dans les systèmes mathématiques". In: *Annales de la Societé Polonaise de Mathématique* (1926), pp. 111–113.
— "On the Limitations of the Means of Expression of Deductive Theories". In: *Logic, Semantics, Meta-mathematics*. Ed. by John Corcoran. 2nd ed. Translated by J. H. Woodger. Hackett Pub., 1935/1983. Chap. 13, pp. 384–392.
Tarski, Alfred, Adrzej Mostowski, and Raphael M. Robinson. *Undecidable Theories*. New York: Dover, 1953/2010.
Väänänen, Jouko. "Set-Theoretic Definability of Logics". In: *Model-Theoretic Logics*. Ed. by K. Jon Barwise and Solomon Feferman. New York: Springer-Verlag, 1985,

pp. 599–643.

Villegas-Forero, Luis and Janusz Maciaszek. "Tarski on Logical Entities". In: *Logica Trianguli* 1 (1997), pp. 115–141.

Vinci, Thomas C. *Space, Geometry, and Kant's Transcendetal Deduction of the Categories*. Oxford: Oxford University Press, 2015.

Wagner, Pierre, ed. *Carnap's Logical Syntax of Language*. New York: Palgrave Macmillan, 2009.

— ed. *Carnap's Ideal of Explication and Naturalism*. New York: Palgrave Macmillan, 2012.

Warren, Jared. "Internal and External Questions Revisited". In: *Journal of Philosophy* 113 (2016), pp. 177–209.

Wetzel, Linda. *Types and Tokens: On Abstract Objects*. Cambridge, Massachusetts: MIT Press, 2009.

Wilson, Mark. "Frege: The Royal Road from Geometry". In: *Frege's Philosophy of Mathematics*. Ed. by William Demopoulos. Cambridge, Massachusetts: Harvard University Press, 1995, pp. 108–159.

— "Ghost World: A Context for Frege's Context Principle". In: *Gottlob Frege: Critical Assessments of Leading Philosophers*. Ed. by Michael Beaney and Erich Reck. Vol. 3. London: Routledge, 2005, pp. 157–176.

— *Wandering Significance: An Essay on Conceptual Behavior*. New York: Oxford University Press, 2006.

— "The Perils of Polyanna". In: *Carnap's Ideal of Explication and Naturalism*. Ed. by Pierre Wagner. New York: Palgrave Macmillan, 2012, pp. 205–224.

Wussing, Hans. *The Genesis of the Abstract Group Concept: A Contribution to the History of the Origin of Abstract Group Theory*. New York: Dover Publications, 2007.

Yaglom, Isaak Moiseevich. *Felix Klein and Sophus Lie: Evolution of the Idea of Symmetry in the Nineteenth Century*. Boston: Birkhauser, 1988.