

**Ontological Methodology and the Philosophy of
Arithmetic: A Critique of Abstractionism**

**A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY**

Michael Calasso

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy**

Advisor: Roy T. Cook

August, 2024

© Michael Calasso 2024
ALL RIGHTS RESERVED

Acknowledgements

Without the tremendous professional support, personal guidance, friendship, and love of many people, I would not have been able to reach this goal. So, I would like to take the time to thank those who have contributed to my success, though I am afraid no amount of praise or kind words can do justice to my feelings of appreciation.

First, my sincerest thanks to my dissertation committee: Roy T. Cook (advisor), Peter Hanks, Geoffrey Hellman, Thomas Hofweber, Michael Kac, and David E. Taylor. The importance of your guidance, patience, and expertise is immeasurable. I'd especially like to thank Roy and David for their commitment to the successful completion of my dissertation. Both of you have spent hours helping me improve my work. For this and so much more, I am very grateful.

To Jessica Gordon-Roth and Aleks Zarnitsyn: You both have kept me afloat and pointed in the right direction, especially this past year. Your kindness, generosity, and enthusiasm have meant everything. Thank you for believing in me, for your friendship, and for pushing me further than I thought I could go.

To Judy Grandbois: Thank you for all the laughter, warmth, and encouragement. During my time at the University of Minnesota, you were my favorite reminder of home.

To my closest friends Ainhoa Fernández Soutullo, Justin Ivory, Dongwoo Kim, Justin Kuster, Nathan Lackey, Qiannan Li, Tucker Marks, Chris Nagel, Sara Parhizgari, Ray Pedersen, Jack Powers, Kylie Shahar, Chris Small, Taylor Smith, Manon St. Amant, Brendan Sullivan, and Yoshinari Yoshida. Thank you for every moment filled with laughter, for teaching me so much, for caring for me and allowing me to care for you, and for eating all my pasta.

Finally, to my mother Barbara, my father Tony, and my sister Diana: You've always believed in me, even when I didn't. I couldn't have asked for a more supportive, loving,

and encouraging family. I love you much.

Dedication

This thesis is dedicated to two people that I have lost during my time at the University of Minnesota: Grace Calasso, my grandmother, and John F. Holmes, my childhood best friend.

Abstract

My dissertation is a study of Bob Hale and Crispin Wright's abstractionism, a realist philosophy of mathematics that originates from the philosophical and technical work of Gottlob Frege (1884-1925). More specifically, I am concerned with the structure and viability of their metaontology: the method by which Hale and Wright establish the existence of numbers as mind-independent objects. At the heart of their view is the claim that the truth of abstraction principles (a special type of implicit definition) and the Syntactic Priority Thesis (a special semantic principle) is enough to guarantee mathematical realism. Hence, they adopt the language-first approach, according to which facts about language can decide metaphysical questions about mathematical objects. The first chapter is introductory; it serves to situate the subject matter of this thesis and provide the necessary background information. The second chapter is historical. Therein I offer a novel interpretation of the language-first arguments for mathematical realism put forth by Frege. On this reading, his metaontology relies on the aboutness properties of arithmetical terms and sentences. This chapter serves to make explicit the mechanics of Frege's metaphysical arguments, which have hitherto remained somewhat mysterious; and to place the metaontology of abstractionism in relief. The third chapter is critical. There I present Hale and Wright's methodology and level several criticisms against it: First, I demonstrate that their argument for the truth of a given abstraction principle is unsuccessful. Second, I show that the Syntactic Priority Thesis has plausible counterexamples. Thus, a different approach must be taken if abstractionism is to count as a promising species of mathematical realism. The fourth and final chapter is constructive. I lay the foundations of a new language-first metaontology for abstractionism that is inspired by the work of Stewart Shapiro, Øystein Linnebo, and Roy T. Cook. As a species of coherentist minimalism, this approach is committed to the following claim: if a formal theory of abstraction meets stringent coherence conditions (or is coherent-plus), then the entities it purports to be about exist as mind-independent abstract objects. Lastly, I show that a particularly important theory of abstraction that grounds arithmetic does, in fact, meet said coherence conditions.

Contents

Acknowledgements	i
Dedication	iii
Abstract	iv
1 Introduction	1
1.1 Bibliography	10
2 Frege and the Language-First Approach	12
2.1 Introduction	12
2.2 About Aboutness	16
2.3 The Mind-Independence Argument	20
2.4 The Objecthood Argument	25
2.5 The Existence Argument	31
2.6 Conclusion	35
2.7 Bibliography	37
3 Abstractionism and the Language-First Approach	40
3.1 Introduction	40
3.1.1 The Epistemology of Abstractionism	42
3.1.2 The Metaphysics of Abstractionism	44
3.2 Syntactic Priority	47
3.3 Conclusion	54
3.4 Bibliography	55

4	Coherentist-Plus Minimalism	57
4.1	Introduction	57
4.2	Sufficiency Claims and Coherentist-Plus Minimalism	59
4.2.1	Auxiliary: Premise 1	61
4.2.2	Auxiliary: Premise 2	63
4.3	The Case of Abstractionism	66
4.3.1	Our Logical Apparatus	66
4.3.2	Metatheory	70
4.3.3	The Theory of Abstraction and Coherence-Plus	72
4.3.4	The Three Cs: Categoricity, Caesar, and CR	90
4.4	Possible Objections	96
4.5	Conclusion	98
4.6	Bibliography	100

Chapter 1

Introduction

The subject of this dissertation is metaontology and the philosophy of arithmetic.¹ “Metaontology,” like most terms in the philosophical lexicon, is ambiguous. Perhaps the most common understanding of the word takes metaontology to be a special branch of philosophy.² Its subject matter: ontology itself. Notice that ontology (or metaphysics, more generally) is concerned with the nature and structure of reality. The ontologist asks questions like: Do abstract objects, like numbers and universals, exist? If so, what is their relationship to concrete objects? Is the past as real as the present? And so on. *Metaontology* on the common understanding is different, of course. It is concerned mainly with the status of ontology as a domain of philosophical discourse. Thus, the metaontologist will ask questions like: Do ontological questions have answers at all? Are ontological debates substantive or merely linguistic disputes? What is the best semantic interpretation of existence claims? Etc.

Though metaontology so understood is fascinating, we will not attempt to answer such questions.³ Rather, we are interested in metaontology in a very specific sense — as a study of the *ways* in which one can decide ontological questions. Accordingly, *a metaontology* is a method (or strategy) for answering the types of questions posed by

¹“Arithmetic” is a slightly antiquated term. Contemporary mathematicians might prefer “number theory.” We assume a broad understanding of arithmetic so that it includes elementary number theory up to real analysis. But we will mainly be concerned with the subject matter of elementary number theory; i.e., the natural numbers: 0, 1, 2, ...

²Or a special branch of *metaphilosophy* if you like.

³Those readers who are interested in such questions are encouraged to read (Chalmers, Manley, and Wasserman 2009).

ontologists.

The most well-known metaontology in analytic metaphysics belonged to W.V. Quine. His particular approach starts with the thesis that our best guides to the nature and structure of the world are well-established scientific theories (Quine 1981, 21). But such theories are often imprecisely stated. Therefore, they are in need of *regimentation*:

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 bare the essence of scientific discourse, can profitably *rework* the language of science beyond what might reasonably be urged upon the practicing scientists.⁴ (Quine 1957, 7-8)

It is then the job of the ontologist to provide this regimentation. How is it to be carried out? By translating the sentences of our best scientific theories into the language of first-order logic with identity, Quine's preferred logical apparatus. Once this regimentation has taken place, Quine's *criterion of ontological commitment* is applied —

A theory is committed to those and only those entities to which the bound variables of the theory must be capable of referring in order that the affirmations made in the theory be true. (Quine 1948, 33)

— thereby determining what our best scientific theories “say” exists. And we ought to admit such entities into our ontology.

At the heart of Quine's metaontology is the link between first-order quantification and existential commitment. It is this association that has been particularly influential. In fact, the *Quinean approach* is metaontological orthodoxy, especially in the philosophy of mathematics.

Thus, the debate between the platonist and the nominalism typically takes this form:⁵ Let \mathcal{D} be some domain of mathematical discourse, e.g., arithmetic. The platonist will argue that quantification over the abstract entities that \mathcal{D} is “about” cannot be eliminated. And so, we *ought* to admit them into our ontology. The nominalist, on the other hand, will search for a way to do away with such existential commitments. Most often, this requires the nominalist to provide a translation scheme for the quantified sentences of \mathcal{D} :

⁴My emphasis.

⁵Instead of “platonist,” we will sometimes use “mathematical realist” or “realist”; instead of “nominalist,” we will sometimes use “mathematical antirealist” or “antirealist”.

$$(\exists x)\phi, \text{ or}$$

$$(\forall x)\phi$$

where ϕ might feature a one-place predicate that “carves out” the abstract ontology of \mathcal{D} .⁶ If the nominalist is successful, then we *ought* not admit said abstract entities into our ontology.

The careful reader will have noticed two things: First, there is an overt linguistic character to the Quinean approach. This is one of its methodological virtues re the philosophy of mathematics. The ontologist need not appeal to *a priori* metaphysical principles or some mystical intuition when dealing with the metaphysics of geometry, arithmetic, etc. Instead, she need only appeal to language and a wealth of logical resources to conduct her investigation. Her work is then continuous with the sciences.⁷

Second, there is an emphasis placed on the word “ought” above. This is meant to suggest that the Quinean approach does not tell us what exists or do not exist based on linguistic considerations. Its goals are more modest than that. At most, this approach gives us a principled reason for accepting or rejecting a certain kind of entity — nothing more.

There are other metaontologies that have an overt linguistic character but are more ambitious. How so? At their core is the assumption that linguistic facts can — in some way — *decide* ontological questions. (Hofweber 2016) offers a metaontology of this sort. In broad outline, Hofweber’s view is this: Natural language quantifiers are *polysemous*. Hence, they admit of two distinct readings: the *external* reading and the *internal* reading. The former is ontologically committal, the latter is not. Now, let \mathcal{D} be some domain of discourse in natural language. We say that externalism is true of \mathcal{D} when the quantifiers it employs are external and the singular terms it features aim to refer; when the quantifiers \mathcal{D} employs are internal and the singular terms that it features do not aim to refer, we say that internalism is true of \mathcal{D} . Thus, if externalism is true of \mathcal{D} , its existential quantifiers are ontologically committal and the singular terms it contains aim to refer. If internalism is true of \mathcal{D} , its existential quantifiers are not ontologically committal and the singular terms it contains do not aim to refer.

⁶We assume a familiarity with formal logic and set theory throughout.

⁷This is something that Quine, himself, was very keen on emphasizing throughout his career.

The ontologist need only show that internalism is true of \mathcal{D} (by doing some linguistics) to demonstrate that certain entities do not exist, according to this picture. Why? Because when internalism is true of \mathcal{D} its existential quantifiers have no ontological import, and its singular terms do not pick out anything in the world.

Any metaontology which assumes the thesis that linguistic facts can decide ontological questions about arithmetic is a part of, what we will call, the *language-first approach*. And it is this type of metaontology that will concern us.

It should be noted that our formulation of the language-first approach suggests at least two interpretations of it: First, that facts about the ontological character and existence of numbers is determined by, or grounded in, facts about arithmetical language. In other words, that numbers exist (or not) and are objects (or not) somehow *depends* on the words and sentences we use or might use to talk about them. This is a substantive metaphysical thesis — one that implies the primacy of language over arithmetic. At the time of writing, the author knows of only two philosophers that have adopted this position: Bob Hale and Roy T. Cook.⁸ Notice that this list is rather short. This is not too surprising, since there is a strong intuition that the world is not furnished by (actual or possible) languages. That being said, such a view is still a version of the language-first approach, and one that we ought not dismiss out of hand.

The other interpretation is epistemological: facts about language can give us privileged epistemological insight into the ontological character and existential status of numbers. Hence, we can use arithmetical discourse to learn about the nature of numbers. Notice that this formulation assumes no metaphysical dependency relation nor does it suggest the primacy of language over arithmetic. So, it is markedly different than the previous one canvassed above and arguably more philosophical palatable.

This flavor of the language-first approach has its origins in Gottlob Frege’s (1848–1925) *logicism* and finds its most pointed contemporary application in Bob Hale and Crispin Wright’s *abstractionism*⁹, two related but distinct philosophies of arithmetic. Let’s consider a brief description of each below to help motivate our project.

Two theses lie at the core of Frege’s logicism: (i) the propositions of arithmetic

⁸Hale’s position is explicated in (Hale 2013). Cook has not endorsed this view in writing as of yet, but he has admitted as much to me in private correspondence. It is with Cook’s permission that I include his name here.

⁹Abstractionism of the sort advocated by Hale and Wright is sometimes called “Neo-Fregeanism.”

are analytic a priori, and (ii) numbers exist as mind-independent abstract objects. His project was philosophically motivated by a dissatisfaction with Immanuel Kant's account of arithmetic¹⁰, which took the propositions of this branch of mathematics to be synthetic a priori.¹¹ Frege defends his position in three books mainly:

1. *Begriffsschrift*(or *Concept Script*) (1879/1972)
2. *Die Grundlagen Der Arithmetik* (or *The Foundations of Arithemtic*) (1884/1980)
3. *Grundgesetze der Arithmetik* (or *The Basic Laws of Arithmetic*) (1893/1903/2013)

In the first piece, Frege develops the first formal system of higher-order logic. In the second, he attacks unsavory philosophical accounts of mathematics and provides informal proofs of the Dedekind-Peano axioms for arithmetic. And in the third, he provides rigorous formal derivations of said axioms.

Frege's strategy for showing that the propositions of arithmetic are analytic a priori was simple. He started by defining *analyticity* and *aprioricity* like so:¹²

A proposition is analytic just in case there is a proof of it that depends only on logical laws and definitions.

A proposition is a priori just in case there is a proof of it that depends only on self-evident, general truths.¹³

Hence, to show that the statements of arithmetic are analytic a priori he needed only to demonstrate that the Dedekind-Peano axioms can be derived from true logical laws and definitions.

Unfortunately, Frege's attempt to do so failed. As is well-known, Bertrand Russell pointed out to Frege that the formal system of *The Basic Laws of Arithmetic* yields a contradiction.¹⁴ And so, it is inconsistent. As one would expect, the fault lies in one of Frege's axioms, the now infamous *Basic Law V*:

¹⁰Interestingly enough, Frege happened to agree with Kant that the statements of geometry were synthetic a priori

¹¹Frege also had a technical motivation: he was greatly troubled by the lack of rigor in the mathematics of his day. Thus, he sought to provide gapless proofs of the Dedekind-Peano axioms for arithmetic.

¹²These definitions are given in (Frege 1884/1980, 4)

¹³For Frege, definitions and logical laws count as self-evident, general truths.

¹⁴This contradiction is known as *Russell's Paradox*.

$$(\forall F)(\forall G)((\S(F) = \S(G)) = (\forall x)(f(x) = g(x))).$$

Thus, with certainty we can say that Frege's attempt to demonstrate (i) was unsuccessful.

Frege's strategy for establishing (ii) is not so simple or obvious. That being said, his language-first approach is certainly evident throughout his writings. Consider one of the most famous excerpts from the Fregean corpus:

§62. How, then, are numbers to be given to us, if we cannot have any ideas or intuitions of them? Since it is only in the context of a proposition that words have any meaning, our problem becomes this: To define the sense of a proposition in which a number word occurs. (Frege 1884/1980, 75)

Frege is saying here that, though we cannot have ideas or sense perceptions of numbers, there is a way to gain access to them. We can fix the meaning of statements in which number words occur. The kind of statements that Frege concerns himself with are identity claims, which are characteristic of arithmetic.

Frege goes on in this section to consider a special type of implicit definition that may be used to fix the meaning of identity statements, *Hume's Principle*:

$$(\forall F)(\forall G)((\#(F) = \#(G)) \leftrightarrow F \approx G).$$

Stated informally, this principle says: for any concept F and any concept G , the number of the concept F and the number of the concept G are identical if and only if there is a one-to-one correspondence between the F s and the G s. But Frege rejects it. The problem is that Hume's Principle cannot tell us which object a given number is identical to. Put more formally, it cannot settle identity statements of the form

$$t = \#(F),$$

where t is not of the form $\#(G)$. As a result, Hume's Principle fails as a definition of *cardinal number* for Frege. This issue is now known as the *Caesar Problem*, taking its name from an earlier passage in which Frege raises a similar complaint to an alternative definition of the individual nature numbers:

... we can never... decide by means of our definitions whether any concept has the number Julius Caesar belonging to it, or whether that same familiar conqueror of Gaul is a number or not. (ibid, 68)

Consequently, Hume’s Principle is not central to his arguments for the existence and ontological character of numbers. Still, the except from §62 quoted above *hints* at how Frege establishes (ii); i.e., meaning must play some role in Frege’s metaontology.

To date, there is no consensus regarding how Frege establishes (ii). Furthermore, no well-worked out description of the mechanics of Frege’s argument(s) for (ii) has been given.¹⁵ Some even claim that Frege merely assumes that numbers exist as mind-independent abstract objects.¹⁶ But I am not so sure about this. Indeed, if Frege is to be rightly regarded as the arch-platonist, there must be more to this story. And there must be some way to reconstruct his arguments faithfully.

Bob Hale and Crispin Wright’s abstractionism is a revival of Frege’s logicist project. In essence, they agree with Frege that (i) and (ii) are true. But they are more optimistic about the use of Hume’s principle as an implicit definition/explanation of *cardinal number*; i.e, as a way to ground our knowledge of arithmetic. In fact, Hale and Wright hold that many mathematical concepts can be explained by way of, what are called, *abstraction principles*.¹⁷

An abstraction principle is a universally quantified biconditional of the form

$$(\forall a)(\forall b)((@ (a) = @(b)) \leftrightarrow (a \sim b)),$$

where a and b range over entities of a given type (e.g., objects, concepts, or n -ary relations) $@$ is function from said entities to objects, and \sim is an equivalence relation¹⁸ on the entities over which a and b range. An abstraction principle serves to introduce a mathematical concept by laying down identity conditions for the abstract objects falling under said concept.

This optimism is rooted in a technical result. A form of the result was first pointed out in (Parsons 1965), and a sketch of it was provided in (Wright 1983) — the book that

¹⁵That is, to the best of this author’s knowledge.

¹⁶e.g., (Ricketts 1986).

¹⁷Hence, “abstractionism.”

¹⁸Let Γ be a set. A relation R is an *equivalence relation* on Γ if, and only if, three conditions hold:

R is *reflective*: for any $x \in \Gamma$, Rxx ,

R is *symmetric*: for any $x, y \in \Gamma$, if Rxy , then Ryx ,

R is *transitive*: for any $x, y, z \in \Gamma$, if Rxy and Ryz , then Rxz .

initiated the abstractionist program. It is known as *Frege's Theorem*: The Dedekind-Peano axioms are derivable in second-order logic with the addition of Hume's Principle and suitable definitions.¹⁹ Therefore, if a successful defense of Hume's Principle as an implicit definition can be sustained, a version of (i) follows.²⁰

Hume's Principle also features in our abstractionist's account of (ii). As a starting axiom of Hale and Wright's program, they take it to be a *conceptual truth* of some sort. Several arguments have been given for this claim.²¹ But that is not enough to secure referential success for the singular terms that Hume's Principle introduces. So, Hale and Wright adopt the *Syntactic Priority Thesis*:

When it has been established... that a given class of terms are functioning as singular terms, and when it has been verified that certain appropriate sentences containing them are, by ordinary criteria, true, then it follows that those terms do genuinely refer. (Wright 1983, 14)

Notice that this thesis reverses the usual order of explanation re truth and referential success. Typically, the truth of a sentence S is accounted for (in part) by appealing to the referential success of the singular terms that occur in it. Here the picture is flipped. The referential success of the singular terms that appear in S are accounted for by the truth of S .

Given all this, abstractionism can rightly be regarded as one of the most impressive philosophies of mathematics on offer. Not only can our abstractionists account for the existence of numbers as mind-independent abstract objects, they can also explain how we come to know arithmetical truths. In short, the platonist can have their cake and eat it too — *if*, of course, Hale and Wright are correct.

And so, we've come to the goal of this dissertation: We will conduct a case study of the metaontology of abstractionism. We start by investigating its origins in chapter 2. Therein, we will take a close look at a number of Frege's written works and reconstruct Frege's arguments for (ii). This will serve to fill the gap in the literature described above, and it will also aid in our comprehension of Hale and Wright's metaontology by putting their methodology in relief. In chapter 3, we will critically evaluate Hale

¹⁹i.e., definitions of the primitive vocabulary that occurs in the formalized versions of the Dedekind-Peano axioms.

²⁰See (Hale and Wright 2001) for a comprehensive defense.

²¹As we will see in ch.2.

and Wright's language-first approach. A number of serious problems will be discussed. Thus, we will demonstrate that their metaontology is not tenable. In chapter 4, we offer a *new* language-first metaontology for Hale and Wright's abstractionist program — one that relies on a coherentist metaontological minimalism.

1.1 Bibliography

- Chalmers, D.J., D. Manley, and R. Wasserman, eds. (2009). *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford: Oxford University Press.
- Frege, G. (1879/1972). *Conceptual Notation and Related Articles*. Translated by T. Bynum. Oxford: Oxford University Press.
- (1884/1980). *The Foundations of Arithmetic*. Translated by J.L. Austin. Evanston, Il: Northwestern University Press.
- (1893/1903/2013). *The Basic Laws of Arithmetic Vols. I & II*. Translated by P. Ebert and M. Rossberg. Oxford: Oxford University Press.
- Hale, B. (2013). *Necessary Beings: An Essay on Ontology, Modality, & the Relation Between Them*. Oxford: Oxford University Press.
- and C. Wright. (2001). *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*. Clarendon, Oxford: Oxford University Press.
- Hofweber, T. (2016). *Ontology and the Ambitions of Metaphysics*. Oxford: Oxford University Press.
- Ricketts, T. G. (1986). "Objectivity and objecthood: Frege's metaphysics of judgment." In *Frege synthesized: Essays on the philosophical and foundational work of Gottlob Frege*, edited by L. Haaparanta and J. Hintikka: 65-95. Dordrecht: Springer Netherlands.
- Parsons, C. (1965). "Frege's Theory of Number." In *Philosophy in America*, edited by Max Black: 180-203. Ithaca, NY: Cornell University Press.
- Quine, W.V. (1948). "On What There Is." *The Review of Metaphysics*, Vol.2(5): 21-38.

— (1957). “The Scope and Language of Science.” *The British Journal of the Philosophy of Science*, Vol.8(29): 1-17.

— (1981). “Theories and Things.” Cambridge, MA: Harvard University Press.

Wright, C. (1983). *Frege’s Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

Chapter 2

Frege and the Language-First Approach

2.1 Introduction

The aim of this chapter is to determine the specific language-first arguments put forth by Frege to establish his mathematical realism. I will not be concerned with the strength of his arguments. My work here is exegetical, not critical. Consequently, I will be interpreting several of Frege's written works. Of particular interest are the following texts: *The Foundations of Arithmetic* (Frege 1884/1980), "Function and Concept" (Frege 1891a), "On Sense and Reference" (Frege 1892a), "Comments on Sense and Reference" (Frege 1892b), and "Concept and Object" (Frege 1892c).¹ Out of these works, the content of *FA* deals most directly with the metaphysics of mathematics; roughly the first half of it is intended to eliminate then-popular views of the ontology of numbers and support Frege's brand of mathematical realism.² Hence, this book will serve to frame our investigation. But not all its metaphysical arguments are explicitly stated. It is then

¹After initially mentioning the name of any of Frege's works, I will use the following abbreviations: *CS* for "Concept Script", *FA* for *The Foundations of Arithmetic*, *FC* for "Function and Concept", *SR* for "On Sense and Reference", *CSR* for "Comments on Sense and Reference", *CO* for "On Concept and Object", and *BL* for *The Basic Laws of Arithmetic*.

²The latter half of *FA* is more technical: there Frege provides a sketch of a formal proof of the Dedekind-Peano axioms from his definitions of "the number of *the concept* F," "zero," the successor relation, and *cardinal number*.

Throughout this chapter, when I speak of numbers, I mean the natural numbers: 0, 1, 2, . . .

necessary to go beyond *FA* to the essays listed above.

Their content mainly highlights fundamental changes in Frege’s conception of language and logic after the publication of his *Concept Script* (Frege 1879/1972) and *FA*. To illustrate the point: In the former work, Frege employed what might be termed a *fact-ontology semantics* to interpret his formal system: judgeable contents are, accordingly, “possible circumstances” or “facts,” unjudgeable contents are — roughly put — individuals. Hence, the negation operator of the formal system of *CS* sends a fact to a fact. In the latter work, Frege mobilizes his *middle-period semantics*, which reinterprets judgeable and unjudgeable contents as something akin to his later semantic theory of sense and reference.³ Frege discards the semantic framework of *CS* and further develops that of *FA* come the publication of *FC*: there he explicitly introduces sense and reference, truth-values as referents of sentences, functions as unsaturated and in need of supplementation by arguments of the appropriate type, etc. Each of these notions finds its application in Frege’s highly technical work *The Basic Laws of Arithmetic* (1893/1903/2013) — the magnum opus of his logicist project. Therefore, the negation-stroke symbol employed in the formal system of *BL* names a unary function that maps objects to truth-values.⁴ Despite the overt logical and linguistic nature of these essays, they have great metaphysical import. So, we will restrict our attention to the following time period: 1884-1893 — the most philosophically fruitful and hopeful time for Frege’s logicism.

Meaning plays a key role in Frege’s metaontology, as we will see. But, given the change in semantic theory mentioned above during this time period, this can make our interpretative work unduly difficult. So, we would do well to adopt and clarify certain exegetical principles that will make our job easier and govern our investigation.

The first: *Frege’s logicism is a work in progress*. We should not approach our investigation with the view that all Frege’s views about language, reality, and mathematics hang together perfectly well or that they are completely internally consistent. That would be a mistake. We must remember that, despite Frege’s great efforts and intellect, the man undertook a difficult feat, and he was bound to slip up along the way to his ill-fated destination. Our job then is to specify Frege’s metaontology as *consistently*

³This view will be relevant for our subsequent discussion in §2.3.

⁴For an interesting discussion of the differences between Frege’s logical systems, see (Cook 2023) and (Dummett 1981).

as possible, by providing reasonable reconstructions of his arguments for mathematical realism.

The second: *Try to keep it local.* We ought not read too much of Frege’s mature semantic theory into his middle-period semantics. More specifically, we should not read his theory of sense and reference into *FA*, without *good justification*. Luckily, Frege provides us with reason for *some* interpretative liberty: He states multiple times that during the writing of *FA*, he did not draw a distinction between sense and reference. Yet, this does not imply that they were not present: the two notions were combined into one. For example, at points, thought and truth-value were mixed into judgeable content.⁵ So, we may separate sense and reference without doing violence to Frege’s reasoning. Careful attention to his dialectic will lend itself to disambiguation.

The third: *Take Frege’s philosophical work to be the handmaiden of his logical work.* In the introduction to this monograph, we noted that Frege had technical and philosophical goals. The former was to provide gapless formal derivations of the Dedekind-Peano axioms from basic laws of logic, therefore grounding arithmetic on secure foundations. The latter was to demonstrate that the truths of arithmetic were *analytic a priori*, against Kant’s view. Given his idiosyncratic understanding of these notions, Frege’s (supposed) technical achievement implies his philosophical achievement. Hence, his technical work is, in some sense, primary.⁶ If we keep this in mind, it becomes clear why Frege made certain choices. For instance, in *FA* §§82-3, Frege sketches a proof of the claim that every natural number has a successor. To do this, Frege must show that the number belonging to the concept *member of the series of natural numbers ending with n* directly follows *n*. This requires applying Hume’s Principle to first-level concepts of this type — concepts under which only objects fall. The application of this exegetical principle will be relevant to our discussion of Frege’s explicit definition of “the number of the concept F” in *FA*.

With the preliminaries out of the way, we may move forward. To give the reader a sense of where we are going, here is a bird’s-eye-view of our destination: Frege starts from

⁵In (Frege 1891b, 150), (Frege1892c, 186), and (Frege 1892b, 178).

⁶(Benacerraf 1981) argues that Frege’s technical goal is primary in a very substantive sense. But I do not agree. If Frege’s technical goal was primary in the way that Benacerraf takes it to be, then all Frege’s metaphysical work is window dressing. Clearly, this is not the case. Frege was very concerned with the metaphysics of mathematics, especially in *FA*.

an analysis of propositions expressed in natural language.⁷ The meanings of arithmetical sentences and number terms are fixed in virtue of their use, syntactic behavior, and speaker-intention.⁸ As such, they are endowed with special representational capacities or *aboutness properties*: the aboutness property of an arithmetical sentence determines that it aims only to represent the mind-independent world; the aboutness property of a number term determines that it can only stand for an object. And so, assuming Frege’s view that sentences are a special type of name⁹, any arithmetical sentence that expresses a true proposition will guarantee the *existence* of numbers as *mind-independent objects*.¹⁰ Our goal then is to reconstruct three arguments: the *mind-independence argument*, the *objecthood argument*, and the *existence argument*.

The reader will be forgiven for thinking: There are many analyses of Frege’s metaphysics and the centrality of meaning, syntax, etc., to it. What is new or interesting about this one? This is a fair question. To answer it, we note two things: First, it is true that many such analyses exist.¹¹ But to the best of this author’s knowledge, they all fall short in clearly bringing to light the *mechanics* of Frege’s arguments for mathematical realism. It remains mysterious *how* and *in what way* Frege derives his conclusions about the ontology of mathematics. Clearly, this is important for understanding Frege’s metaphysics and its contribution to philosophy. Second, we are introducing a *novel* interpretation of Frege’s metaontology here — one that relies on some special representational capacities of language. None of these analyses have mentioned

⁷A *proposition* is the meaning of a grammatically well-formed sentence expressed in the indicative mood. I assume no special technical sense of “meaning.” Hence, my use of “proposition” implies neither Frege’s early semantic theory of judgeable/unjudgeable content of *FA* nor his mature semantic theory of sense and reference. This sketch of his metaontology is intended to be broad. Also, I do not use “state of affairs” in any special way.

For the remainder of this chapter, the *referent* of an expression *e* in a language (formal or natural) \mathcal{L} is the thing named by *e*; *reference* is the relation that holds between a name and its referent; and we say that *reference obtains* for *e* when *e* has a referent. Throughout, I characterize languages syntactically, not semantically.

⁸A *number term* is any expression in a formal or natural language that is used to refer to a number; this includes numerals, definite descriptions, symbols, and conventional number names. An *arithmetical sentence* is any sentence in a formal or natural language containing number terms that are used to refer to numbers and that is expressed in the indicative mood.

⁹i.e., in Frege’s mature semantic theory, sentences are names of truth-values. Furthermore, a syntactically complex name cannot refer unless all of its constituent names refer.

¹⁰Frege uses auxiliary arguments to establish that numbers are abstract. Though they are of considerable philosophical interest, we will not be concerned with them here.

¹¹The most prominent of which are, in this author’s humble opinion: (Dummett 1973), (Dummett 1991), and (Kluge 1980).

aboutness properties to date.¹²

In §2.2, we will discuss these aboutness properties and provide a semi-formal characterization of them. Subsequent sections will treat Frege’s metaphysical arguments: §2.3 deals with the mind-independence argument, and §2.4 and §2.5 concerns the objecthood argument and the existence argument, respectively.

2.2 About Aboutness

Many things are endowed with aboutness properties. Portraits, poems, songs, and mental states are commonly taken to have representational capacities. A portrait represents some individual. A poem can be about some feeling or experience. A mental state, like desiring, is about some situation obtaining. Intuitively, each of these things represents its subject matter in different ways and for different reasons.¹³ Portraits seem to represent people in a fundamentally different way than mental states represent people. Hence, a one-size-fits-all account of aboutness properties is highly implausible or even impossible. What about individual words and sentences?

Of course, they have aboutness properties too. The sentences we utter in common discourse are always about something: people, beliefs, ideas, etc. One natural account grounds their aboutness in reference. “Jimi Hendrix was the greatest guitar player ever” is about a specific man *because* “Jimi Hendrix” refers to him. “Sirius is 8.611 light years from Earth” can only be about the brightest star in the night sky and its distance from Earth, since “Sirius” and “Earth” refer to those celestial bodies. Consider a familiar type of example discussed in (Donnellan 1966): Suppose Jones is at a party, and he notices a content looking fellow in the corner of the room drinking sparkling water. Jones says to himself, “That man over there drinking champagne looks happy.” Even though what Jones literally says is false, there is a strong intuition that Jones succeeds in referring to the man drinking sparkling water, and as a result: the sentence is about him. In each case, the aboutness of the sentence is grounded in the fact that its constituent words name certain things. It is a matter of semantic value.

There are other types of linguistic aboutness; i.e., a feature that a syntactic item has

¹²Again, to the best of my knowledge.

¹³See (Putnam 1998) and (Yablo 2014) for interesting discussions on the topic.

in virtue of the intensional item it expresses. To motivate this, consider the following examples: Suppose there is a physicist who interprets and uses scientific language with the intention of describing reality beyond observational facts. She disagrees with her radical positivist colleagues who claim that the meaning of “All electrons are negatively charged,” is really a claim about what observations we would make if certain experiments were conducted. When asked, she will sincerely answer that if she makes a scientific claim, our physicist means to talk about the world beyond observation. But this is not a statement merely describing her intentions to depict reality with language or what she hopes to do with her words. Its significance is stronger: Our physicist’s proposition, itself, is “directed at” a state of affairs that would “make it true” — namely, one in which electrons exist and are negatively charged. Our physicist and her positivist colleagues disagree on the appropriate interpretation of scientific claims. Still, even the positivists see that her proposition is legitimate *qua* meaning. And it is one that is fundamentally different from theirs in that it is *about* the world beyond observational facts if it about anything at all. So, “All electrons are negatively charged,” when uttered by her, “aims at describing” a state of affairs beyond observation.

Now, take the Higgs boson — a chargeless, massive scalar boson with zero spin and even positive parity. Any particle that satisfies this definition will correctly be categorized as a Higgs boson. Any particle that does not have one of those properties cannot serve as the referent of, say, “that Higgs boson.” Thus, this definition determines what *kind* of entity may serve as the referent of that expression.

The first example demonstrates that a proposition can constrain the class of *possible truth-makers* for itself, thereby limiting the types of states of affairs the sentence that expresses it can aim to describe. While the second example demonstrates that the class of *possible referents* of a singular term can be constrained by the meaning of that term. Pithily put: intension determines extension.

The kind of aboutness properties that feature in Frege’s arguments are of this sort. There are two of special interest to us, which are related to the examples just seen. Before we get to the details, a few things should be noted: First, a rough grasp of the properties in question is all that is needed to understand Frege’s arguments. Hence, what follows will be stated rather loosely and generally. In fact, this is necessary, given the imprecise way in which Frege presents some of his arguments. Second, the use of

“truth-maker” (above or below) should not be taken to imply any specific metaphysical or semantic view of truth-making. Our use of this term is somewhat metaphorical. Third, given the foregoing paragraph, it should be clear to the reader that the properties in question admit of a modal characterization. But it should not be assumed that Frege would have or could have done the same. Our modal characterization is meant as a heuristic — nothing more.

We start with the first aboutness property that features in Frege’s argument for the mind-independence of numbers. Let \mathcal{L} be a language and \mathcal{P} be a proposition expressed by some declarative sentence $\mathcal{S}_{\mathcal{P}}$ of \mathcal{L} ; then:

$\mathcal{S}_{\mathcal{P}}$ has (A_1) iff for any accessible possible world w in which \mathcal{P} is true, \mathcal{P} ’s truth-maker is a mind-independent state of affairs.

To get a sense of the content of this definition, let’s look at some intuition eliciting examples: Suppose Jones is a radical metaphysical idealist. He believes that all states of affairs are composed of mental entities and are caused by his own mind. In the interest of philosophical rigor, Jones provides a translation scheme for all sentences that purportedly describe mind-independent reality. “There is a tree in my front yard” abbreviates “there is an object^I, which I subsume under the concept tree^I, that is a part of my front yard^I,” where the superscript “I” is meant to indicate ontological dependency on the speaker. Jones would insist that what is required for his proposition to be true is something quite different than what would be required if, say, the direct realist Smith were to utter the same sentence. After all, Jones’ translation scheme assigns a meaning to “There is a tree in my front yard” that is markedly different than that of Smith, since the latter’s is directed at the non-phenomenal world. If Jones were to imagine all the possible worlds in which Smith’s proposition is true, he would note that its truth-makers must be mind-independent states of affairs. So, the declarative sentences uttered by Smith that aim to describe the mind-independent world have (A_1) .

Or consider a possible world containing a sole inhabitant, the direct realist Mary, in which the radical metaphysical idealist’s thesis holds. The constituents of this world are products of Mary’s mind, arranged in various states of affairs. Whenever she utters a sentence about her surroundings it has (A_1) . Can the propositions expressed by these sentences be actually true? No. Mary believes that some of them are true, surely. But,

strictly speaking, her propositions require mind-independent truth-makers. If Mary comes to learn of her lonely predicament, she will certainly agree that she has never uttered a sentence that expresses a true proposition, since she has never correctly spoken about the world in which she lives.

The reader who feels the force of these examples can see that a proposition can limit the class of its possible truth-makers to mind-independent states of affairs. Thus, any sentence that expresses such a proposition will have this aboutness property.

The second aboutness property is a feature of individual words or noun phrases, which Frege uses in his arguments for the mind-independence and objecthood of numbers. Let s be a word or noun phrase in \mathcal{L} , and let m be the unique, fixed meaning assigned to s , given some linguistic context; then:

s has (A_2) iff there is an ontological category O such that for any accessible possible world w in which s expresses m and refers to r , r belongs to O .

Recall that Spinoza defined “God” as “the unique, infinite, self-caused substance.” Holding this meaning of “God” fixed, if “God” refers, then it is impossible for anything to be identical to God that is not a substance (in the traditional sense of “substance”); otherwise, “The unique, infinite, self-caused substance is not a substance” *could* be true! Or take the generalized quantifier “every boy” in “Every boy is a troublemaker.” Assuming the standard semantic interpretation of such quantifiers, it follows that if “every boy” refers, then it refers to a property of properties or a set of sets. It would be totally absurd then to take the universal quantifier as referring to a non-property, or a non-set, like a person. Thus, with these fixed meanings, “God” and “every boy” have (A_2) .

An analogy might help here. Egon Schiele’s *Self-Portrait with Physalis* (1912) can only be said to represent Schiele and one type of thing, a person, if it represents anything at all. Once we understand the function of a portrait, this is obvious. So, if someone were to enquire as to who this portrait represents or insist that it represents the star Sirius, there is a sense in which the speaker does not understand what this portrait is about or what it *could* be about. Similarly, the meaning of a word functions to limit the set of entities to which it can refer in some cases.

One might take exception to these aboutness properties, assuming one can get past their somewhat imprecise and metaphorical articulation. But remember: we are not

interested in the correctness of Frege's arguments. We want to understand their mechanics, not critique them. With this in mind, we press forward with our exegesis.

2.3 The Mind-Independence Argument

The mind-independence argument appears in *FA* §26: *Is number something subjective?* Frege's reasoning is presented as a series of loose analogies and examples, which makes understanding his argument all the more difficult. To aid comprehension, here is a sketch of it: Take the arithmetical sentence "The number of Jupiter's moons is four." Given its typical meaning and use, it has (A_1), since the proposition it expresses is directed at the mind-independent world. By the Context Principle, the meaning of "four" is directed at the mind-independent world too. So, "four" has (A_2) — more specifically, any referent of "four" must be mind-independent. Without loss of generality, if a number word refers, its referent must be mind-independent. The reader is advised to keep this sketch in mind throughout this section.

We start by noting the three principles that Frege states at the end of the introduction to *FA* to guide his investigation:

In the enquiry that follows, I have kept to three fundamental principles:
 always to separate sharply the psychological from the logical, the subjective from the objective;
 never to ask for the meaning of a word in isolation, but only in the context of a proposition;
 never to lose sight of the distinction between concept and object. (Frege 1884/1980, x)

If an order of importance can be ascribed to these statements, it is certainly the first one that holds primacy. Frege, as is well-known, was completely allergic to any subjectivist/psychologistic conception of the ontology of mathematics. (Hence, the importance of demonstrating that numbers are mind-independent.) The first thesis also introduces an important metaphysical bifurcation for Frege: there are two very broad, disjoint ontological categories, the mind-dependent and the mind-independent. This, of course, plays a role in his argument.

The second statement is the familiar Context Principle. It should be noted that “proposition” means “sentence” here. Incidentally, Frege holds this principle, in part, as a way to retain the distinction between the subjective and the objective.¹⁴ A myriad of interpretations of it are available. Some take it to be a thesis about sense, others a thesis about reference.¹⁵ It is beyond the scope of this chapter to canvass all other interpretations of the Context Principle or to offer a new one. Still, something must be said about it, given its place in the argument sketch provided above. We assume a broad function for the Context Principle: many aspects of the meaning of a whole sentence are conferred on the meaning of its constituent words, including *being directed at the mind-independent world*. This facilitates an important move in Frege’s argument.

The final statement does not appear directly in the argument. But this warning to always separate concept and object — two radically different kinds of entity — is important for understanding the distinction between Frege’s fact-ontology semantics and the semantic framework he assumes in *FA*. This will aid our interpretation of the argument.

Frege begins with the following:

For number is no whit more an object of psychology... than the North sea is... In the same way number, too, is something objective. If we say “The North Sea is 10,000 square miles in extent” then neither “North Sea” nor by “10,000” do we refer to any state of or process in our minds: on the contrary, we assert something quite objective, which is independent of our ideas and everything of the sort. (ibid, 34)

Notice Frege’s series of claims: First, when we say “The North Sea is 10,000 square miles in extent,” “The North Sea” and “10,000” do not *refer* to anything subjective. So, with the strict distinction between the mind-dependent and the mind-independent, Frege is implying that, like “the North Sea,” “10,000” must refer to something objective.

The second claim is that *what* we are asserting when we utter that sentence is objective. That this follows immediately from the first claim suggests that it plays some explanatory role: “10,000” must refer to something objective, *because* what we

¹⁴As Frege says, “If the second principle is not observed, one is almost forced to take as the meanings of words, mental pictures or acts of the individual mind, and so to offend against the first principle as well” (ibid, x).

¹⁵See (Pelletier 2001) for an overview of such interpretations.

are asserting — some judgeable content — is objective. Thus, judgeable contents and *their* objectivity somehow guarantee that the referent of “10,000” is mind-independent. How? And how are we to understand the notion of content objectivity? Presumably the latter question will help us answer the former. So, let’s tackle it first.

Two ways to understand content objectivity come to mind: (a) We can assume that Frege still held his fact-ontology semantics of *CS* and use that to explain it. Or (b) we can take content objectivity to be strict mind-independence. Suppose we go with (a). Thus, when someone utters “The North Sea is 10,000 square miles in extent,” they assert some fact or circumstance. It is no wonder then that numbers are mind-independent — they are constituents of some objective state of affairs! But this interpretation is surely mistaken. In the introductory remarks to this chapter, I mentioned that Frege held his middle-period semantics while writing *FA* and that it is distinct from his fact-ontology framework. Why think they are different? Recall Frege’s warning: never to lose sight of the distinction between concept and object. This is a sharp metaphysical bifurcation. But in *CS* this line was not drawn. The analogous notions of function and argument were not metaphysical at all in the earlier work. Rather, the distinction reflected different ways of carving up a sentence to represent one aspect of it as a function and another as an argument. Accordingly, we may take “Jake” to be the argument in “Jake is tall” and “is tall” to be the function, or the other way around. Frege also allows for the existence of empty concepts since 1884. Up to two years prior to the publication of *FA*, he does not.¹⁶ Indeed, his definition of “zero” in *FA* relies on empty concepts. Thus, (a) is not a viable option.

Now, suppose we go with (b). This option is problematic too. In §26 (and throughout *FA*), Frege assumes that the communicability of judgeable contents is explained by, or implies, their mind-independence. Take the following passage:

Space, according to Kant, belongs to appearance. For other rational beings it might take some form quite different from that in which we know it. Indeed, we cannot even know whether it appears the same to one man as to another; for we cannot, in order to compare them, lay one man’s intuition of space beside another’s. Yet there is something objective in it all the same; everyone recognizes the same geometrical axioms, even if only by his behavior, and must do so if he is to find his way about the world. What is objective

¹⁶See (Frege 1882).

in it is what is subject to laws, what can be conceived and judged, what is *expressible in words*. What is purely intuitable is not *communicable*.¹⁷ (ibid, 35)

Frege is saying here that mental entities, like our perception of space, cannot be communicated. But our propositions about space *can* be communicated. Again, what is explained by this, or is implied by this, is the mind-independence of such propositions. If this is so, we have a problem: Consider the sentence “My sense perception of this glass of wine is subjective.” Its content is communicable. Hence, its content is mind-independent. But assuming (b) is the correct interpretation, Frege would have to admit that the referent of “My sense perception of this glass of wine” is mind-independent. Naturally, Frege would not accept this for obvious reasons. So, (b) is off the table.

Perhaps a sensible option can be found if we read on. Following his remarks regarding the referent of “10,000,” Frege writes:

The botanist means to assert something just as factual when he gives the Number of a flower’s pedals as when he gives their color. . . There does, therefore, exist a certain similarity between Number and color; it consists. . . in their being both objective. (ibid, 35)

This short passage lends itself to a suggestion. Ordinarily, when someone makes a statement of number, what they mean to assert, the content of their sentences, is *factual* in the sense that it is about the world beyond the subjective. In other words, typical numerical claims are *world-directed*; i.e., what would make them true is some mind-independent state of affairs. Thus, a case can be made that the role of sentential content at play here involves representational capacities of a particular type. That is to say: Frege is tacitly appealing to some form of (A_1).

With this, sense can be made of the passage regarding the referent of “10, 000.” This term must refer to something mind-independent, because the sentence it appears in has (A_1). Now, though what Frege literally says in that passage features content objectivity, the subsequent passage just cited seems to suggest a nonstandard interpretation of it.

¹⁷My emphasis. (Ricketts 1986) takes passages like this to suggest that Frege was primarily concerned with the objectivity of propositions, not the objectivity of numbers; and that objectivity is really a logical notion, not a metaphysical one for Frege. Hence, he cannot have been a traditional mathematical realist. I disagree, of course. Throughout §26 of *FA*, Frege continually comes back to the mind-independence of numbers, themselves, contrasting their existence with that of mental entities; e.g., the *idea* of the number 2.

If not, it remains a mystery how and in what way Frege concludes that numbers are mind-independent.

There is a logical gap here, though. How is it that just because the content of “The North Sea is 10,000 square miles in extent,” is world-directed, some of its constituent terms can only refer to mind-independent things? The most likely explanation is that Frege must be implicitly using the Context Principle to make this move. After all, he is inferring that a property of the content of a whole sentence can tell us something about that sentence’s constituent parts. But, of course, the Context Principle concerns the meaning of words in the context of a sentence, not the words themselves. So, the inference involves two steps: First, some property of the proposition expressed by “The North Sea is 10,000 square miles in extent” is conferred on the meaning of “10,000”. Second, the fact that the meaning of “10,000” has this property guarantees that it refers to a mind-independent thing if it refers at all.

The property in question must be world-directedness. Given that the proposition expressed by “The North Sea is 10,000 square miles in extent” is about the mind-independent world, the meaning of “10,000” is about, or is directed at, some mind-independent thing. As such, “10,000” can refer only to mind-independent entities. This means that “10,000” has (A_2).

Is there any textual evidence to support this? Does Frege say anything to suggest that the meaning of a word is world-directed, and hence the possible referents of this word must be mind-independent? Yes, in fact, he does. Speaking of colors again, Frege writes:

The word ‘white’ ordinarily makes us think of a certain sensation, which is . . . entirely subjective; but even in ordinary speech, it often bears, an *objective sense*. When we call snow white, we *mean to refer* to an objective quality which we recognize, in ordinary daylight, by a certain sensation. . . Often, therefore, a color word does not signify our subjective sensation, which we cannot know to agree with anyone else’s. . . , but rather an objective quality. (ibid, 36)

The picture painted here is this: The typical meaning of “white” is fixed by its use and the intention of the speaker. When we say “Snow is white,” we mean to talk about some mind-independent state of affairs. Hence, with “white,” we intend to refer to some mind-independent quality of objects. Its meaning is directed at some objective thing,

as a result. So, the referent of “white” is mind-independent.

It appears that our way of filling the gap in Frege’s reasoning is vindicated. Therefore, we may sum up his argument for the mind-independence of number like so:

1. Take some arithmetical sentence S , and let t be a number term that appears in it.
2. Given the typical usage of S , its content is world-directed — S has (A_1) .
3. It follows that the meaning of t is world-directed, too.
4. And so, if t refers to x , x must be mind-independent — t has (A_2) .

Assuming that we have a name for all numbers *and* that reference does obtain for them, it follows that numbers are mind-independent.

This concludes our treatment of the first argument. It is worth pausing for a minute to appreciate the explicit language-first methodology that Frege employs. Notice how certain properties of language, once established, can tell us something about entities to which we have no direct epistemological access. There is, therefore, a logical neatness and profundity in Frege’s reasoning — features that are exemplified in all of Frege’s language-first arguments, as we will see.

2.4 The Objecthood Argument

We now turn to Frege’s argument for the objecthood of numbers. A hint of this argument appears in *FA* §55: *Every individual number is a self-subsistent object*, nothing more. So, it is necessary to go beyond this text. As before, an argument sketch is provided to aid comprehension: Consider the sentence “The number of Jupiter’s moons is four.” It is meaningful. Thus, “four” has a sense. By virtue of the syntactic form and behavior of “four,” its sense is “complete”. This implies that it can only refer to objects — “four” has (A_2) . Hence, if reference obtains for it, then four is an object.

We begin with §55 of *FA*. There Frege writes:

In the proposition [sentence] “the number 0 belongs to the concept F”, 0 is only an element in the predicate... For this reason, I have avoided calling a number such as 0 or 1 or 2 a property of a concept. Precisely because

it forms only an element in what is asserted, the individual number shows itself for what it is, a self-subsistent object. I have already drawn attention above to the fact that we speak of “the number 1”, where the definite article serves to class it as an object. In arithmetic this self-subsistence comes out at every turn, as... in the identity $1 + 1 = 2$. Now our concern here is to arrive at a concept of number useable for the purpose of science; we should not, therefore, be deterred by the fact that in the language of everyday life number appears also in attributive constructions. That can always be got round... the proposition “Jupiter has four moons” can be converted into “the number of Jupiter’s moons is four.” (ibid, 68-69)

At first glance, it appears that Frege’s reason for believing that numbers are objects is based on a totally syntactic criterion: that an expression behaves a certain way, e.g., can occur as a “stand alone” element of a predicate or whole sentence, implies that if it has a referent, that referent must be an object. For example, the expression “0” functions like this; therefore, if reference obtains for it, 0 is an object, not a property or concept. There is seemingly no room for content to play any role here.

But this reading is mistaken. Frege follows this passage with an apparently out-of-place series of objections to the claim that number terms refer to ideas. During this argument, Frege invokes the Context Principle, the main purpose of which is, he says, to guard against taking the referent of a number word or a numeral to be an idea. Yet, its use also makes room for content in his criterion:

It is enough if the proposition taken as a whole has a sense; it is this that confers on its parts their content... The self-subsistence which I am claiming for number is not to be taken to mean that a number word signifies something when removed from the content of a proposition, but only to preclude the use of such words as predicates or attributes, which appreciably alters their *meaning*.¹⁸ (ibid, 72)

Extrapolating on this passage and the previous quotation, Frege is saying something like this: Take some arithmetical sentence. It has meaning. Thus, so do the number terms that occur in it. If we look at the way that such terms behave syntactically in this sentence and others like it, we see that they are not attributive. They “stand alone” — they are singular terms. So, the meaning of a singular term is different than that of a predicate.

¹⁸My emphasis.

Notice that this hints at the place of content in Frege’s criterion of reference to objects: (a) The fact that a word is a singular term tells us something about its content, and (b) its content will play some role in establishing that it must refer to an object.

Nothing more can be gleaned from *FA*. So, we look to some of Frege’s later writings in which he develops his mature semantic theory and clarifies aspects of the logical system presented in *CS*. In the first of these essays, *FC*, Frege explicates the status of concepts and relations as a type of mathematical function, introduces the distinction between sense and reference, as well as the distinction between complete/saturated and incomplete/unsaturated expressions, truth-values as the referents of proper names, etc. *SR* provides Frege with the tools to deal with identity statements. And *CO* addresses puzzles raised by the distinction between function and argument, which was first clearly explicated in *FC*.

Singular terms are, of course, a type of complete expression for Frege.¹⁹ The use of “complete” is somewhat metaphorical. The basic idea is this: complete expressions are in no need of supplementation; when taken as a linguistic item, they contain no “empty spaces,” unlike predicates. Thus, our goal is to figure out how these types of complete expression come to refer only to objects. To accomplish this, we will first concern ourselves with (b): the role that the sense of a singular term t plays in determining that t can refer only to objects.

We start from Frege’s claim that logically simple notions, e.g., that complete expressions refer only to objects, is not a stipulation, but a *discovery* (Frege 1892c, 182).²⁰ How is this discovery made? In *CO*, Frege responds to a series of objections from Benno Kerry regarding his distinction between concept and object. One concern of Kerry’s is

¹⁹Whole sentences being the other type of complete expression. It should be noted that Frege never offers a strict definition of “complete expression.” We have to rely on our linguistic competence to identify them, which I assume is unproblematic.

²⁰As is stated in this essay, Frege is claiming that concepts as logically simple referents of incomplete expressions is a discovery. I assume this holds for objects, too.

A natural question might arise: for x to be an object, must there exist an actual (or possible) completed expression that does (or would) refer to it? This was the view presented in (Hale 2013). I think this reflects a misunderstanding of Frege, though; it assumes that he is stating a necessary condition for objecthood. He is not. That would be too much analysis for so simple an idea. Frege is merely saying that the things picked out by complete expressions are objects. We must also remember that Frege is working toward a major technical work in which he developed a fully interpreted language to deduce the Dedekind-Peano axioms. His philosophy services that, in part. Frege’s philosophical work needs only to secure as much as is needed to fulfill his mathematical goals. Securing a type of referent for number words or numerals suffices.

that logical rules, like taking the referents of complete expressions to be objects, should not be dictated by linguistic convention. Frege responds:

... but my own way of doing this is something that nobody can avoid who lays down such rules at all, for we cannot understand one another without language, and so in the end we must rely on people's understanding words, inflexions, and sentence-construction in essentially the same way as ourselves... I was not trying to give a definition, and to this end I appeal to the general feeling for the German language. (ibid, 184)

Contra Kerry's understanding, Frege is not stipulating these "logical rules." Rather, they are based on our shared linguistic competence and understanding. They must, in fact.

For Frege, understanding the meaning of words boils down to understanding their senses (Frege 1892a, 154). So, the discovery that complete expressions can only refer to objects must rely on our knowledge of their senses. Hence, there is something that we gather from the senses of complete expressions that tells us about their possible referents. It's not clear what this is.

SR opens with a puzzle about identity: Suppose we take identity to be a relation between objects. Then it would seem that there is no difference between " $\sum_{i=0}^{\infty} \frac{1}{2^i} = 2$ " and " $2 = 2$." Yet, there is a difference: the former is informative, the latter is not. What explains the difference?

A difference can arise only if the difference between the signs corresponds to a difference in the *mode of presentation* of the thing designated. (ibid, 152)

Continuing with his own example, Frege writes:

Let a, b, c be the lines connecting the vertices of a triangle with the midpoints of the opposite sides. The point of intersection of a and b is then the same as the point of intersection of b and c . So we have different designations for the same point, and these names ('point of intersection of a and b ', 'point of intersection of b and c ') likewise indicate the *mode of presentation*; and hence the statement contains actual knowledge.²¹ (ibid, 152)

The difference is then explained by the way in which the sense of a singular term t depicts the referent of t — its mode of presentation, or the way in which the sense

²¹My emphasis.

illuminates the referent. As Dummett once said, “The senses of an expression are the way that a reference is given to us” (Dummett 1988, 69).

Perhaps the mode of presentation of a sense can tell us something about possible referents then. But from what we have here, it seems that the sense of singular term t can only tell us about the *actual* referent of t . So much is suggested by some of what Frege said:

The *Bedeutung* of a proper name is the object itself which we designate by using it; the idea which we have in that case is wholly subjective; in between lies the sense, which is indeed no longer subjective like the idea, but is yet not the object itself. The following analogy will perhaps clarify these relationships. Somebody observes the Moon through a telescope. I compare the Moon itself to the *Bedeutung*; it is the object of the observation, mediated by the real image projected by the object glass in the interior of the telescope, and by the retinal image of the observer. The former I compare to the sense, the latter is like the idea of intuition. (Frege 1892a, 155)

This analogy places sense in between the singular term t that expresses it and the referent of t . Moreover, the information conveyed by t 's sense and the sense, itself, is apparently dependent upon the referent if we take this analogy to heart.

Evans has advocated this interpretation, which is known as the *mode of presentation account* of sense (Evans 1982, 12-13, 22-23). According to it, the sense of t and how it represents t 's referent is ontologically dependent on said referent. Empty meaningful names are, therefore, an impossibility. But they are not (Frege 1892a,157).²² In fact, one of Frege's favorite examples of an empty name with a sense is “Odysseus.” It follows that senses do not need an actual referent to exist or to convey information.

A more plausible understanding of how the sense of a singular term conveys information is to think of senses as (metaphorically) containing a list of instructions, or an algorithm, for identifying something. On this view, when someone grasps the sense of, e.g., “the Morning Star,” that person knows how to determine the truth (or falsity) of identity statements of the form:

x is the Morning Star.

²²In (Dummett 1988, ch.7), it is argued that there is an inconsistency between this account and the possibility of meaningful empty names. So, like myself, Dummett disagrees with Evans.

It is no stretch then to conclude that the sense of a singular term also conveys objecthood. Why must this be the case, though? An account of (a) — the fact that a word is a singular term tells us something about its sense — will help us answer this question.

After the publication of *FC*, Frege moves freely between talk of the completeness (and incompleteness) of expressions, senses, and referents. That is, complete expressions have complete senses and referents; incomplete expressions have incomplete senses and referents. To use Frege’s example, we may decompose the expression “ $2 \cdot 1^3 + 1$ ” into “ $2 \cdot x^3 + x$ ” and “1” recognizing the function in the form of the first expression; he says, “From this we may discern that it is the common element of these expressions that contain the essential peculiarity of the *function*” (Frege 1891a, 133).²³ The essential peculiarity of the function is *its* incompleteness. But, of course, the functional expression and the function are mediated by a sense. And if we were to use the functional expression as a singular term, it would “appreciably alter its meaning.” So, too, with complete expressions, like singular terms.

What is implied by this is that the syntactic behavior of a typical singular term in a meaningful sentence forces a certain kind of sense — one that is complete. In other words, one that must convey information about an *object*. Hence, a singular term t can only refer to an object if it refers at all — t has (A_2) .

Now, we may sum up Frege’s objecthood argument like this:

1. Take some meaningful arithmetical sentence S , and let t be a number term that appears in it.
2. Thus, t has a sense.
3. By virtue of the syntactic behavior of t , its sense is complete.
4. Hence, if t refers to x , x must be an object — t has (A_2) .

Assuming that we have a name for all numbers and that reference does obtain for them, it follows that numbers are objects.

This concludes our treatment of Frege’s objecthood argument. At this point, the careful reader will have noticed that the conclusions of both the mind-independence argument and the objecthood argument are conditional claims, which share the same

²³My emphasis.

antecedent: that some arithmetical term t refers to x . The next argument establishes this proposition.

2.5 The Existence Argument

We have come to Frege’s argument for the existence of numbers. More so than the objecthood argument, we must go beyond the scope of *FA*. In contrast to the two previous arguments we have looked at, Frege’s reasoning here is fairly straightforward. So, we will omit a sketch of the argument, and jump into the deep end.

§62 of *FA* begins with the following question, “How, then, are numbers to be given to us, if we cannot have any ideas or intuitions of them?” To answer it, Frege invokes the Context Principle and considers Hume’s Principle as an implicit definition of the concept *number*; put in modern notation:

$$(\forall F)(\forall G)(\#(F) = \#(G) \leftrightarrow F \approx G)^{24},$$

where F and G range over concepts, and “ \approx ” signifies equinumerosity; i.e., the existence of a one-to-one correspondence. Frege ultimately rejects Hume’s Principle, though. The main issue is one of semantic indeterminacy — the Caesar problem: Hume’s Principle cannot decide all identity statements that involve the terms it introduces. Thus, it cannot tell us whether Caesar is the number of some concept. The main issue is that, for Frege, adequate definitions must have sharp conceptual limits. So, he proposes an explicit definition of *number*:

$$\#(F) = Ext(\approx F),$$

where $Ext(\approx F)$ is the extension of the concept *equinumerous with the concept F* (Frege 1884/1980, 80).²⁵ Therefore, numbers are identified with the extensions of second-level concepts.²⁶ This definition gives rise to a host of questions. The first, and most obvious, is: are numbers just extensions of concepts? Our intuitive sense of number seems to have

²⁴Read: for any concept F and any concept G , the number of the concept F and the number of the concept G are identical if and only if there is a one-to-one correspondence between the F s and the G s

²⁵We may think of the extension of a concept F as the collection of all things that are F . But that is not to say that an extension of a concept is a set. For Frege, it is some kind of *logical object*.

²⁶That is, a concept under which first-level concepts fall. First-level concepts are concepts under which objects fall.

nothing to do with extensions or logical objects, more generally. Thus, this definition seems rather arbitrary. This is exacerbated by the fact that Frege claims to attach no importance to bringing in extensions of concepts at all (ibid, 79).

But let's remember our third exegetical principle. Frege's identification of numbers with a certain type of extension in *FA* has technical value. Recall that his goal is to prove the Dedekind-Peano axioms from higher-order logic amended with suitable definitions. The key word is "suitable." The technical worth of his definition is demonstrated in §70-83 in which he sketches some of his sought-after proofs. Frege, himself, points to two features of extensions that lend support to their use: that they have clear identity conditions and that their size is comparable.²⁷ These are technical attributes that extensions share with numbers. As such, they lend themselves to Frege's work.

The second and more important question for our purposes is: how do we know that extensions exist? Frege gives no explicit argument for their existence in *FA*. He merely assumes their existence and our knowledge of the content of identity statements that feature extension terms (ibid, 117). More specifically, what accounts for our knowledge of these logical objects is an early version of Basic Law V that Frege implicitly assumes in *FA*:

$$(\forall F)(\forall G)((Ext(F) = Ext(G)) \leftrightarrow (\forall x)(F(x) \leftrightarrow G(x))).$$

The version is markedly different from the version of Basic Law V that we find in *BL*:

$$(\forall f)(\forall g)((\S(f) = \S(g)) = (\forall x)(f(x) = g(x))).$$

The first major difference is that the former version of Basic Law V assigns extensions to concepts, while the latter version assigns *value-ranges* to functions.²⁸ The second major difference is that by *BL*, Frege took sentences to be names of truth-values. Hence, the use of the identity symbol in place of the biconditional.²⁹ In *BL*, too, Frege merely assumes the existence of value-ranges and grounds our knowledge of them in the latter

²⁷The use of "size" is my own. Frege says of extensions that one is *wider* than the other. "Size" and "wider" might seem to suggest cardinality and suppose an already accepted definition of number, which seems circular. But this is not so, since Frege's use of a one-to-one correspondence does not presuppose number.

²⁸We may think of value-range of a function as its graph. But again, they are not a type of set, for Frege.

²⁹It should be noted that modern notation is not only anachronistic, but it is not faithful to Frege's logical notation. I use it for the sake of the reader.

version of Basic Law V.³⁰ But it does not follow that an argument for their existence cannot be constructed on Frege's behalf. After all, Frege was a careful reasoner. And he would not have admitted extensions or value-ranges into his ontology if he did not believe that there was some reason, consistent with his principles, for their existence. For the remainder, I will restrict my attention to extensions. But the general form of the argument that I will construct on Frege's behalf can be used to prove the existence of value-ranges as well.

So, how are extensions given to us if we cannot have ideas or intuitions of them? You might suspect that Frege would require fully determinate truth-conditions for sentences of the form $q = Ext(F)$, where q is substituted with a non-extension term, lest we run into Caesar again. This seems like a totally plausible expectation. Yet, there is a difference here: Frege took Basic Law V to be a *true logical law*, not a definition. As such, it need not meet the same requirements that stipulative definitions must satisfy. As Frege says, "The first place where a scientific expression appears with a clear-cut *Bedeutung* is where it is required for the statement of a law" (Frege 1891a, 131). One way of reading this quote that bares on our question is that there is no need to give fully determinate truth-conditions for identity statements involving only one extension term. Their senses are completely known, and as a consequence, they have a clear-cut referent.

Still, this is unsatisfying. Does Frege give us any more reason to believe that the truth of logical laws, or truth in general, guarantees reference during this time period?

Let's consider what Frege says about fictional statements. Speaking of the sentence "Odysseus was set ashore at Ithaca while sound asleep," he writes: "Anyone who seriously took the sentence to be true or false would ascribe to the name 'Odysseus' a *Bedeutung*. . . for it is of the *Bedeutung* of the name that the predicate is affirmed or denied" (Frege 1892a, 157). The idea seems to be that when we judge a proposition to be true (or false), we recognize a relationship between (or lack thereof) two relata, a sequence of arguments and a function. Hence, no argument(s), no truth-value.

Frege held this view of fiction beyond the time period with which we are concerned. On his treatment of truth in *Thought* (Frege 1918), he makes a distinction between sentences that are expressed in the assertoric mood and sentences that are expressed in

³⁰More specifically, in §29 of *BL*, Frege assumes that value-range terms always have a referent.

fictional contexts. In the former case, we mean to assert something about the world; in the latter case, we do not. It would then not make sense to ask whether a fictional sentence is true or false, because in those contexts we are not interested in the relationship between a sequence of arguments and a function.

What all this suggests is that, though Frege held that “true” is indefinable, something can be said about the way people *use* the word and what we are doing when we ascribe truth to the sense of a sentence. In essence, truth is *about* a relationship between arguments and functions. Therefore, it applies to propositions that we use to assert something about existents. That is why Frege says, “The laws of logic are first and foremost laws in the realm of the *Bedeutungen*...If it is a question of the truth of something — and truth is the goal of logic — we also have to enquire after the *Bedeutungen*” (Frege 1892b, 178). *Truth*, for Frege, has — perhaps — a third type of aboutness property that we have not seen before.

Hence, we have a principled reason for the existence of extensions (and value-ranges), after all. Frege took Basic Law V to be a true logical law. Thus, given this understanding of the nature of truth, the early version of Basic Law V implies the existence of extensions.

There is a major issue with this interpretation: it does violence to the content-force distinction, which runs through Frege’s writing and is formally codified in *BL* by the horizontal stroke and the judgement stroke. In particular, it suggests that it is consistent with Frege’s views on language that someone can assert a thought and thereby make it truth-apt. So, there is no strict divide between force and content.

Our first exegetical principle, which requires that we construct Frege’s arguments as consistently as possible, takes this option off the table for us. Hence, we must find a different route to Frege’s existence argument. Luckily, one is available. With the advent of Frege’s mature semantic theory, he holds that for a sentence to refer to a truth-value at all, reference must obtain for its constituent singular terms (Frege 1892a, 157). Why is this so? One answer relies on Frege’s assimilation of sentence with names. Consider “John’s oldest brother”: If there is no John, then “John’s oldest brother” has no referent. If sentences are names of truth-values, they function similarly. “John’s oldest brother is tall” would name the True only if John exists.

We can apply this same reasoning to the early version of Basic Law V: Let F be

a concept. Trivially, $(\forall x)(F(x) \leftrightarrow F(x))$. Thus, $Ext(F) = Ext(F)$. And so, there is some $x = Ext(F)$. But F was arbitrarily chosen. So, for any concept F , there is some $x = Ext(F)$. Using Frege's stipulation that $\#(F) = Ext(\approx F)$, we may prove Hume's Principle:

Suppose $\#(F) = \#(G)$. Thus, $Ext(\approx F) = Ext(\approx G)$. By Basic Law V, $(\forall K)(K \approx F \leftrightarrow K \approx G)$. We know that $F \approx F$. So, $F \approx G$. (The other direction is just as simple.)

Now, let's define $0 = \#(x \neq x)$. Plug $x \neq x$ in for F and G in Hume's Principle. Since $x \neq x \approx x \neq x$, it follows that $\#(x \neq x) = \#(x \neq x)$. Hence, there is some $x = \#(x \neq x)$; i.e., 0 exists. We can continue this mode of argument to show that each individual number exists. Define $1 = \#(x = 0)$. Plug $x = 0$ in for F and G in Hume's Principle. Since $x = 0 \approx x = 0$, it follows that $\#(x = 0) = \#(x = 0)$. Therefore, there is some $y = \#(x = 0)$; i.e., 1 exists. And so on.

This suffices to establish the antecedent of the conclusions to the mind-independence argument and the objecthood argument. It follows that numbers exist as mind-independent objects - *if* Frege is correct.

2.6 Conclusion

In this chapter, we have reconstructed three arguments on Frege's behalf: the mind-independent argument, the objecthood argument, and the existence argument. Taken in conjunction, they establish Frege's central metaphysical thesis: numbers exist as mind-independent objects. Our goal was to make clear the mechanics of Frege's reasoning, therefore demonstrating how and in what way he establishes his thesis. We saw that the first two arguments relied on aboutness properties of arithmetical sentences and terms. These aboutness properties bridge that gap between language and extralinguistic reality for Frege. And they exemplify some of the more idiosyncratic elements of his language-first approach. The last argument functioned differently. It makes no appeal to aboutness properties. Rather, it relies on Frege's thesis that sentences are names of truth-values, securing referents for number terms.

Before we move on to the next chapter, let's remind ourselves of the place and importance of what we have seen in the overall scheme of this monograph: We are

conducting an investigation into the metaontology of a particular philosophy of mathematics; i.e., Bob Hale and Crispin Wright's abstractionism — which has its origins in Frege's logicism. In order to better understand and appreciate the views of Frege's heirs, we took to studying the language-first approach of Frege himself. By doing so, we developed a novel interpretation of Frege's arguments for the existence and ontological character of numbers. One hopes that our interpretation demonstrates two things: first, the profundity and creativity of Frege as a metaphysician, and second, his importance as a contributor of a unique approach to the philosophy of mathematics.

2.7 Bibliography

Frege's Works

- (1879/1972). *Conceptual Notation and Related Articles*. Translated by T. Bynum. Oxford: Oxford University Press.
- (1882). “Letter to Marty, 29.8.1882.” Reprinted in (Beaney 1997): 79-83.
- (1884/1980). *The Foundations of Arithmetic*. Translated by J.L. Austin. Evanston, Il: Northwestern University Press.
- (1893/1903/2013). *The Basic Laws of Arithmetic Vols. I & II*. Translated by P. Ebert and M. Rossberg. Oxford: Oxford University Press.
- (1891a). “Function and Concept.” Reprinted in (Beaney 1997): 130-148.
- (1891b). “Letter to Husserl.” Reprinted in (Beaney 1997): 149-150.
- (1892a). “On Sense and Reference.” Reprinted in (Beaney 1997): 151-170.
- (1892b). “Comments on Sense and Reference.” Reprinted in (Beaney 1997): 171-180.
- (1892c). “On Concept and Object.” Reprinted in (Beaney 1997): 181-193.
- (1918). “Thought.” Reprinted in (Beaney 1997): 325-345.

Secondary Literary and Other Materials

Beaney, M., ed. (1997). *The Frege Reader*. Oxford: Blackwell.

Benacerraf, P. (1981). “Frege: The Last Logician.” *Midwest Studies in Philosophy*, 6(1): 17-36.

Cook, Roy. (2024). “Frege’s Logic.” *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), edited by Edward N. Zalta and Uri Nodelman.
<https://plato.stanford.edu/archives/sum2024/entries/frege-logic/>.

Donnellan, K. (1966). “Reference and Definite Descriptions.” *The Philosophical Review*, vol.75(3): 281-304.

Dummett, M. (1973). *Frege: Philosophy of Language*. London: Duckworth.

— (1981). *The Interpretations of Frege’s Philosophy*. London: Duckworth.

— (1988). *Origins of Analytical Philosophy*. Cambridge, MA: Harvard University Press.

— (1991). *Frege: Philosophy of Mathematics*. Cambridge, MA: Harvard University Press.

Evans, G. (1982). *The Varieties of Reference*. Edited by John McDowell. Oxford: Clarendon Press.

Kluge, E.W.H. (1980) *The Metaphysics of Gottlob Frege: An Essay in Ontological Reconstruction*, vol.5. Dordrecht: Springer Science+Business Media.

Pelletier, F.J. (2001). “Did Frege Believe Frege’s Principle?” *Journal of Logic, Language, and Information*, vol.10: 87-114.

Putnam, H. (1988). *Representation and Reality*. Cambridge, Massachusetts: The MIT Press.

Ricketts, T. G. (1986). "Objectivity and objecthood: Frege's metaphysics of judgment."
In *Frege synthesized: Essays on the philosophical and foundational work of Gottlob Frege*, edited by L. Haaparanta and J. Hintikka: 65-95. Dordrecht: Springer Netherlands.

Yablo, S. (2014). *Aboutness*, vol.3. Princeton: Princeton University Press.

Chapter 3

Abstractionism and the Language-First Approach

3.1 Introduction

In this chapter, we concern ourselves with the direct philosophical descendants of Frege, Bob Hale and Crispin Wright, and their *abstractionism*. As was noted in the introduction to this monograph, Hale and Wright agree with Frege’s epistemological and metaphysical theses: (i) the truths of arithmetic are analytic a priori, and (ii) numbers exists as mind-independent objects. They differ with Frege on several points, the most substantial of which is this: Hale and Wright are more optimistic about the possibility of using certain implicit definitions as explanations of the fundamental concepts of arithmetic. To this end, our abstractionists employ a special type of implicit definition: *abstraction principles*.¹ To remind the reader, an abstraction principle is a universally quantified biconditional of the form:

$$(\forall a)(\forall b)((@a) = @(b)) \leftrightarrow (a \sim b),$$

where a and b range over entities of a given type (e.g., objects, concepts, or n -ary relations), $@$ is function from said entities to objects, and \sim is an equivalence relation on the entities over which a and b range. An abstraction principle endows the name

¹Hence, “abstractionism.”

of @ with meaning. As such, it serves to introduce a mathematical concept by laying down identity conditions for the abstract objects falling under said concept.

Abstraction principles play a special role in establishing both (i) and (ii) for Hale and Wright. Since we are investigating the metaontology of abstractionism, our attention will be focused on the role they play in demonstrating (ii).

We will find that Hale and Wright's arguments for (ii) rely on one important principle: the *Syntactic Priority Thesis* (henceforth, *(SP)*). Our aim will be to level objections against it. To give the reader a sense of where we are going, here is a rough sketch of our destination: The principle *(SP)* is all important to the abstractionist's ontological program. It says, in essence, that if an expression functions as a singular term and appears in certain true sentential contexts, then that expression has a referent. The expressions that matter for our abstractionists are, of course, terms that are introduced by abstraction principles. Hence, assuming *(SP)* and the relevant abstraction principle, Hale and Wright are able to prove that numbers exist. But we will see that *(SP)* is best understood as three separate theses, two of which are problematic. The first implies a test for singular-termhood, which is unusable or moot. The second is subject to counterexample. Furthermore, Hale and Wright use *(SP)* to demonstrate that a given abstraction principle is true. However, the general form of this argument is viciously circular. Thus, Hale and Wright's language-first approach cannot decide that numbers exist as mind-independent objects.

Before one can critique this metaontology, it must be understood. I think that comprehension would best be served by offering the reader a brief sketch of the epistemology of abstractionism and — naturally — a more detailed treatment of its metaphysics. This will be the subject of the remainder of §3.1. The epistemology of Hale and Wright's abstractionism interests us insofar as it lays down a minimum condition for the acceptability of its metaontology: the latter cannot conflict with the former. More specifically, the metaontology of abstractionism cannot presuppose any assumptions or imply anything that is inconsistent with the epistemological goals of Hale and Wright. This will be known as the *consistency constraint* on abstractionism. Subsequent sections, are organized as follows: In §3.2, we examine the role that *(SP)* plays in Hale and Wright's argument for the existence of numbers and offer our objections to its subsidiary theses. Finally, in §3.3 we will attempt to answer the question: What lessons have we learned

about an acceptable metaontology for abstractionism?

3.1.1 The Epistemology of Abstractionism

We start from an epistemological problem for mathematical realism famously posed in (Benacerraf 1973). Benacerraf starts by laying down two constraints for any account of mathematical truth:

- *Semantic Constraint*: Any account of mathematical truth must parallel the semantics of the rest of our language.²
- *Epistemological Constraint*: Any account of mathematical truth must cohere with a plausible general epistemology.

The former constraint prohibits assigning truth-conditions to mathematical statements that are wildly different from the truth-conditions of grammatically similar sentences. The latter constraint prohibits accepting any account of mathematical truth that makes mathematical knowledge mysterious or unintelligible.

Now, take informal first-order Peano Arithmetic. It contains expressions that are used as names for numbers, like “0” and “the successor of 0.” If we follow the semantic constraint, we obtain the usual truth-conditions for the sentences of Peano Arithmetic. Thus:

$0 \neq s(0)$ if and only if for some $a, b \in \mathbb{N}$ such that “0” refers to a
and “ $s(0)$ ” refers to b ,
it is not the case that a is b .

Notice that this truth-condition requires the existence of two numbers. But numbers are typically taken to be mind-independent, abstract objects — entities which are causally inert and exist outside of space-time. Human beings are concrete entities. We are bounded by space-time, and our cognitive capacities are (presumably) subject to causal law. Any plausible epistemology, say, one that appeals to a causal relationship

²Benacerraf assumes that we have a well-worked out semantic theory for the rest of our language, knowing full well that this is not the case.

between the knower and the known, accordingly cannot explain mathematical knowledge. But mathematical knowledge could be explained by appeal to some mysterious intuition, or the like, which would conflict with the epistemological constraint.

Hale and Wright attempt to solve this problem by giving an account of how propositional thought about numbers³ is possible and how it can be knowledgeable. Their account explains two things: (a) how meaning is (and thus, how truth-conditions are) conferred on whole identity statements featuring number terms, and (b) how the satisfaction of said truth-conditions is cognitively accessible to us. The picture that emerges is this: thought about numbers is possible, because we already know the truth-conditions for a certain class of identity statements involving number terms, and if we know such truth-conditions, we can know which numerical identity statements are true and which are false. Hence, it is from within a language endowed with sufficient resources that we may acquire knowledge of numbers. And we need not explain our knowledge of mathematics by appeal to acquaintance or some mathematical extrasensory perception, while also holding true to the semantic constraint.

Abstraction principles play a key role in explaining both (a) and (b). In the case of number theory, (a) and (b) are explained by *Hume's Principle* (henceforth, *HP*):

$$(\forall F)(\forall G)((\#(F) = \#(G)) \leftrightarrow (F \approx G)),$$

where F and G are second-order variables ranging over concepts, " $F \approx G$ " abbreviates "There exists a one-to-one correspondence from F onto G ," and " $\#(F)$ " is read: "the number of the concept F ."

HP serves to introduce the concept *cardinal number* by providing identity conditions for the abstract objects that fall under it. Hence, as an implicit definition, *HP* serves to define " $\#$ " and the terms formed from it by fixing the truth-conditions of identity statements in which they appear. More importantly, it *explains* the truth-conditions of statements of the form $\#(F) = \#(G)$ as coincident with the truth-conditions of statements of the form $F \approx G$, which are already cognitively accessible to us.

Therefore, *HP* (purportedly) accounts for (a) and (b). If so, abstractionism answers Benacerraf decisively. But there is another, more profound implication. We know

³By "propositional thought about numbers," I take Hale and Wright to mean thought involving whole sentences that are purportedly about numbers.

that *Frege's Theorem* holds: the Dedekind-Peano axioms for number theory follow from *HP* and second-order logic plus suitable definitions. Thus, if we assume a principle of deductive epistemic closure⁴, any number theoretic statement for which there is a proof can be known, at least in principle.⁵

3.1.2 The Metaphysics of Abstractionism

If abstractionism is going to count as a viable species of mathematical realism, it must offer solid reasons for the thesis that numbers exist as mind-independent objects. Let's look at a typical abstractionist argument for the existence of zero as an object. Take the concept $x \neq x$. Universally instantiate this concept on all variables of (*HP*):

$$((\#(x \neq x) = \#(x \neq x)) \leftrightarrow (x \neq x \approx x \neq x)).$$

It is obviously true that $(x \neq x \approx x \neq x)$. Hence, by biconditional elimination, we obtain:

$$(\#(x \neq x) = \#(x \neq x)).$$

But “ $\#(x \neq x)$ ” is a singular term that appears in a true atomic sentence. Therefore, by (*SP*), we are warranted in inferring:

$$(\exists y)(y = \#(x \neq x));$$

i.e., there is an object y such that y is the number of the concept *non-self-identical*. Finally, we stipulate that $0 = \#(x \neq x)$, substitute “0” for “ $\#(x \neq x)$ ” in “ $(\exists y)(y = \#(x \neq x))$ ” and get:

$$(\exists y)(y = 0).$$

I'm sure the reader can see that we can carry on the argument that we encountered in chapter 1 to show that all finite cardinal numbers exist as mind-independent objects.

⁴We will not come back to this point later. For our purposes, it does not matter whether or not epistemic closure is viable.

⁵This section is paraphrased from (Hale and Wright 2009).

Notice that there is no appeal to any extralinguistic ontological commitments in this argument. Our abstractionists starts with, what they take to be, a meaning-constitutive truth — *(HP)* — and, using some familiar inference patterns and *(SP)*, they conclude that zero exists as a mind-independent object. If this argument works, the metaphysics of abstractionism dovetails very nicely with its epistemological goals. There is no conflict, only an easily obtainable ontology with a palatable epistemology.

A causal gloss of the argument for the existence of zero reveals that its success hinges on two things: first, the truth of *(HP)*; second, the truth of *(SP)*.

Our abstractionists begin their metaphysical arguments (and their whole program!) by laying down *(HP)*. Thus, for their metaphysical arguments to be convincing, we must have — first and foremost — some reason to countenance *(HP)*.⁶ The mere stipulation of *(HP)* cannot suffice for its truth; otherwise, it would be possible for anyone to lay down an abstraction principle and prove that their favorite abstract object exists, e.g., God. Incidentally, *(SP)* plays a central role in the argument for the truth of *(HP)*.

It may be stated as follows:

- If t is a singular term that appears in an extensionally atomic sentence, then there is some object a such that t refers to a .

By an “extensionally atomic sentence,” the abstractionists mean a sentence in which the syntactic positions held by singular terms are *reference-demanding*; i.e., these terms must refer if the sentence in which they appear in is true. Therefore, *(SP)* can be restated as:

- If t is a singular term that appears in a reference-demanding position in an atomic sentence, then there is some object a such that t refers to a .

This version of *(SP)* is, arguably, a conceptual truth. As such, it seems that there is no reasonable way to challenge it. But a closer look at *(SP)* and its role in the argument above might prove otherwise. Following (MacBride 2003), I believe that this principle

⁶Note: I did not say that we need some reason to believe that *(HP)* is a good implicit definition and can be known to be true *a priori*. That is a question about the epistemological status of *(HP)*. Hale and Wright have argued for a number of conditions, e.g., conservativeness, harmony, and generality, for an implicit definition to be meaning conferring and knowledge-underwriting. This does not concern us here. We are interested in the metaphysical status of *(HP)*.

is best understood as a collection of theses about the relationship between language and reality:

- *Syntactic Decisiveness* (SP_1): If an expression δ has the syntactic features of a singular term, then δ has the semantic function of a singular term — *reference*.
- *Referential Minimalism* (SP_2): If δ has the semantic function of a singular term and it appears in a true extensionally atomic sentence, then δ refers to something in the world.
- *Linguistic Priority* (SP_3): In the order of explanation, syntactic categories are prior to ontological categories.

(SP_1) lays down a sufficient condition for an expression δ to have the semantic function of a singular term — that of picking out some unique entity. It implies that if we have some way of telling that δ behaves syntactically like a singular-term, then there is no doubt that it functions like one. Our abstractionists do provide some syntactic/inferential criteria that is meant to decide whether an expression does, indeed, function as a singular term (and, therefore is a singular term). Hence, they are able to decide that any novel expression introduced by (HP) functions semantically as a referring expression — at least in principle.

(SP_2) states that reference obtains for any referring expression that appears in an extensionally atomic sentence. So, if we can show that the left-hand-side of any instance of (HP) is true, it follows that the novel terms featuring in that atomic sentence do pick out something in the world. Accordingly, the order of explanation re truth and reference is reversed: truth is explanatorily prior to reference.

(SP_3) is one of the most distinctive and philosophically interesting aspects of Hale and Wright's abstractionism. This principle reverses the usual order of explanation that philosophers of language have assumed. According to it, ontological categories are not to be understood first before syntactic categories. Rather, syntactic categories explain ontological categories, e.g., *object* or *property/concept*. For Hale and Wright, an object *just is* the referent of a *possible* singular term.

Taken in conjunction, these principles are at the heart of Hale and Wright's language-first approach. Without them, our abstractionists cannot bridge the gap between language and reality. Therefore, their role in the argument given above is of particular

importance to us.

3.2 Syntactic Priority

On the face of it, (SP) and its related theses, play but one role in the argument for the existence of zero that we saw above. They allow Hale and Wright to infer

$$(\exists y)(y = \#(x \neq x))$$

from

$$(\#(x \neq x) = (\#(x \neq x))).$$

But this is actually not the case. There are two places in which this class of theses shows up. We start at the beginning. In doing so, we will see that at each step, Hale and Wright run into trouble.

For Hale and Wright, the truth of (HP) , itself, relies on (SP) . Hence, in laying down (HP) as an initial premise in the argument for the existence of zero, our abstractionists tacitly appeal to (SP) . In what way does the truth of (HP) rely on (SP) ?

Take note of the fact that the argument for the existence of zero requires that any inference from the right-hand-side to the left-hand-side of (HP) must be truth-preserving. Given that Hale and Wright take (HP) at face-value, this requires the existence of a total function, named by “#,” that is defined on the field of \approx . What we need then from Hale and Wright are good reasons to believe that the Ramsay sentence for (HP) is true:

$$(\exists h)(\forall F)(\forall G)((h(F) = h(G)) \leftrightarrow (F \approx G)).$$

In (Hale and Wright 2009), our abstractionists claim that the reason for the existence of such a function can be found in the resources of (HP) , itself. And these reasons rely on what (HP) can accomplish as an implicit definition. Hence, for their argument to get off the ground, they assume that (HP) succeeds in endowing “#” with a sense. We will do the same.

Our abstractionists are not Meinongians. For them, like most of us, referential

success is over and above the possession of sense. Thus, a referent for “#” is secured a different way. Following Frege, Hale and Wright claim that “#” refers to a function only if the singular terms formed from it by *(HP)* succeed in referring. This seems more than plausible: A function f is a mapping from arguments to values; its existence requires that for each suitable argument, there exists a unique value. If this condition is not met, f does not exist. Semantically, the existence of a unique value $f(x)$, for all appropriate arguments x , is cashed out in terms of a unique referent for each term formed from the name of f and the name of x . So, the question becomes: how do we know the terms formed from “#” have referents?

One way to verify the existence of a referent for a term t is to prove that $(\exists x)(x = t)$. This is usually done by finding a term q and then demonstrating that $q = t$, where there is no doubt about the existence of q 's referent. But there is no hope of doing this here. Our abstractionists claim the *(HP)* introduces a *fundamental* means of reference to abstract objects; i.e., a means of reference that does not make appeal to the semantic values of other terms. Moreover, the epistemological goals of abstractionism require that our knowledge of numbers be totally grounded in *(HP)*. Thus, there must be some other means by which reference is secured for the terms in question.

The abstractionists appeal to an alternative method — a way to determine the truth of identity statements that feature terms formed from “#” and names for concepts. If we can do that, then we can apply *(SP)* and deduce the existence of referents for said terms. But we have a way to verify such identity statement — *(HP)*. We merely verify the right-hand-side of an instance of *(HP)*, and then use biconditional elimination to obtain the left-hand-side of that instance.

To sum up Hale and Wright's argument: We know that the Ramsay sentence for *(HP)* is true, because we know that there exists a function named by “#.” And we know that this function exists, because the terms introduced by *(HP)* refer. And we know that these terms refer, because they appear in true identity statements. And we know these identity statements are true, because we may obtain them from a true-instance of *(HP)*.

But a true instance of *(HP)* will require the existence of a function named by “#.” So, by appealing to *(SP)*, Hale and Wright, have grounded the truth of the Ramsay sentence for *(HP)* in itself. We have, therefore, a *viciously* circular argument here.

For the time being, we will disregard this circularity and move on to the next step in Hale and Wright’s argument for the existence of zero — the one that makes explicit reference to (SP) in the move to

$$(\exists y)(y = \#(x \neq x))$$

from

$$(\#(x \neq x) = \#(x \neq x)).$$

This requires, first, that we know that the expression “ $\#(x \neq x)$ ” has the semantic function of a singular term or that it is one. According to, (SP_1) this is evidenced by the syntactic behavior of “ $\#(x \neq x)$.” So, how are we to know that this expression does exhibit the requisite syntactic behavior?

Hale and Wright offer a syntactic/inferential criteria for singular-termhood, following the proposal in (Dummett (1973)).⁷ These criteria function to distinguish genuine singular terms from other types of substantival expressions, e.g., “something,” that are capable of standing *salva congruitate*, in the place of genuine singular terms, and non-substantival expression, like predicates, by administering a test. Hale and Wright’s proposal begins:

A *substantival* expression t functions as a singular term in a sentential context ‘ $A(t)$ ’ iff [a suitably competent speaker of English can recognize as valid the following inferences] (I) the inference is valid from ‘ $A(t)$ ’ to ‘Something is such that $A(it)$ ’...⁸ (Hale 2001b)

Note that Hale’s test is supposed to apply to a competent speaker of English who can recognize certain inferences as valid. It is here that this issue lies.

Suppose that we have a competent speaker of English, Jones. And that we have expanded English to English[#] by stipulating (HP) . As a stipulative implicit definition, (HP) acts as a rule that governs the behavior of, what appear to be, identity statements involving novel expressions — in fact, that’s all it does. Now, Jones either accepts the

⁷Hale and Wright’s criteria can be found in (Hale 2001a) and (Hale 2001b).

⁸What is included in the brackets is my addition. But it is one of the most important assumptions that Hale makes, though it is easy to overlook.

abstractionist's claim that the expressions introduced by (HP) are genuine singular terms or he does not. If he does, there is no need to apply Hale's test. If Jones does not accept their supposition, then ask yourself: given the rule of usage codified in (HP) , how can Jones possibly recognize the inference from, say, $(\#(x \neq x) = \#(x \neq x))$ to $(\exists y)(y = \#(x \neq x))$ as valid? The only rule of usage Jones has for these expressions is (HP) , after all. Yet, it tells us nothing about the inference patterns that are legitimate with respect to the terms it introduces.

To make this point even clearer. Suppose we expand English to English^Z with the following abstraction principle:

$$(\forall x)(\forall y)(\text{Zoink}(x) = \text{Zoink}(y) \leftrightarrow T(x, y)),$$

where T is the equivalence relation *has the same spatio-temporal location as*. According to the syntactic rules of English^Z, "Zoink(The Empire State Building)" is a legitimate expression. Suppose that Jones is a competent speaker of English^Z who does not know whether or not "Zoink(The Empire State Building)" is a singular-term, but he wants to find out. So, Jones tries to apply the first part of Hale's test. If presented with, e.g., "Zoink(The Empire State Building)=Zoink(The Empire State Building)," I highly doubt he would recognize the inference to " $(\exists y)(y = \text{Zoink}(\text{The Empire State Building}))$ " as valid, since this abstraction principle is the only rule of usage that pertains to zoink-expressions. Presumably, Jones would want to know if "Zoink(The Empire State Building)=Zoink(The Empire State Building)" is a genuine identity statement. That presupposes knowledge of the syntactic category of "Zoink(The Empire State Building)." Doesn't that put the abstract object before the horse? Hence, it seems that Hale's test is either moot or it is impossible to apply it to the case of the expressions introduced via (HP) and abstraction principles more generally.

This is highly problematic: The argument for the existence of zero requires that we know " $\#(x \neq x)$ " functions semantically like a singular term. To know this with any certainty, our abstractionists require that we must apply this test. But the test is inapplicable (unless one accepts by fiat that " $\#(x \neq x)$ " is a singular term). Therefore, the inference described above is suspect.

We move on to (SP_2) and (SP_3) . Recall (SP_2) :

- *Referential Minimalism*: If δ has the semantic function of a singular term and it appears in a true extensionally atomic sentence, then δ refers to something in the world.

On the face of it (SP_2) seems unproblematic and obviously true. But is it? One who is suspicious of (SP_2) might be tempted to offer the following counterexample from fiction:

- “Nabokov wrote of Humbert Humbert.”

Presumably, Humbert Humbert does not exist (at least not in the way he is described in the book). Furthermore, “Humbert Humbert” is a proper name — it certainly functions syntactically as such. Also, “wrote of” seems to be an atomic predicate. So, by (SP_2), “Humbert Humbert” exists. Correct?

There are several ways out of this for Hale and Wright. First, they can deny that “Humbert Humbert” functions semantically like a singular term. But on what grounds? They seemingly cannot appeal to syntax. Perhaps they can appeal to Wittgenstein’s meaning is use thesis, and argue that the way this expression is used determines that it is not referential. But if that is so, then it seems their syntactic/inferential test is doubly moot. Why bother giving a complicated criterion for singular-termhood if we can just determine its semantic function from its use?

Second, Hale and Wright can claim that this sentence is (i) false or that (ii) it does not have a truth-value at all. A way to account for (i) is to claim that Humbert Humbert does not exist. But how would they know that for sure? Hale and Wright accept numbers as abstract objects, why not fictional characters? Option (ii) is not too helpful either. It seems to fly in the face of a very, very clear intuition that this sentence has a truth-value; i.e., true. In fact, it seems absurd to deny it.

Third, our abstractionists can argue that “wrote of” is not an atomic predicate. This appears to be the best option. After all, one can analyze “Nabokov wrote of Humbert Humbert” like so: “There is a written work \mathcal{L} such that Nabokov wrote \mathcal{L} and Humbert Humbert is described in \mathcal{L} .” Hence, “Nabokov wrote of Humbert Humbert” is not atomic at all.

But it should be noted that this counterexample is no counterexample at all. The

existence of fictional entities is a contentious issue in metaphysics.⁹ And it would be, not only hasty, but unphilosophical to claim with any certainty that Humbert Humbert does not exist.

But are there any sentences that will do the job? For this, we can appeal to the history of science:¹⁰

- “Descartes posited ether to explain action-at-a-distance.”

Clearly, “ether” is a singular term, and we know that this sentence is true. Furthermore, it appears that “ x posited y to explain z ” is an atomic predicate. So, by (SP) , ether exists. Yet, given the results of the Michelson-Morley experiment, this is false.

Like before, Hale and Wright could dodge this counterexample by arguing that “ x posited y to explain z ” is not an atomic predicate. Hence, the sentence featuring “ether” can be analyzed as follows:

- “There is a time t such that in/at t , Descartes used ether to explain action-at-a-distance.”

This is one way out. But this paraphrase opens up the possibility for a more forceful counterexample, since we know *when* Descartes used ether to explain action-at-a-distance:

- “In 1664, Descartes used ether to explain action-at-a-distance.”

Notice that “In w , x used y to explain z ” is an atomic sentence — it cannot be analyzed further into a more complex predicate involving quantifiers or truth-functional connectives. If this is correct, then we have a counterexample to (SP_2) .

Finally, we have come to (SP_3) :

- *Linguistic Priority*: In the order of explanation, linguistic categories are prior to ontological categories.

⁹See (Thomasson 1999) and (Voltolini 2003). They provide interesting contemporary arguments for fictional realism; (Everett 2005) is on the other side of the philosophical fence; Everett offers an extension of Russell’s (Russell 1905) arguments against *Meinongeanism* to do away with fictional entities.

¹⁰I provide just one example, but clearly there are many different types of examples like this one.

The thought behind this principle is that ontological categorization is *dependent upon* and *prior* to syntactic categorization. Hence, x is an object just in case there is a possible singular term that refers to x . On the face of it, this view of the relationship between language and the world is troubling. This reaction springs from a common sense metaphysical intuition: extralinguistic reality is as it is, regardless of the ways in which we talk about it, think about it, or facts about the syntax of language. So, a natural series of questions arise: What is Hale and Wright's evidence for (SP_3) ? Or is it a mere stipulation/approach that they adopt as a part of their metaontology?

For Hale and Wright, (SP_3) is part of an approach — one they adopt for a simple reason: In all likelihood, there is no other way to explain what the concept *object* (or *property/relation*) that is sufficiently general enough. To contrast: Suppose we define the class of objects to be all and only those concrete entities, and we define the class of properties to be all and only those entities that are instantiated by an object. With these definitions we have ruled out the possibility of abstract objects and uninstantiated properties, unfairly.

The requirement that we do not decide, by virtue of our chosen definitions or explanations, answers to metaphysical questions is more than reasonable. And (SP_3) certainly satisfies this. Still, there might be one possible issue: (SP_3) presupposes *some* metaphysical baggage: possible linguistic items; i.e., possible singular terms. (SP_3) links the existence of possible singular terms with the existence of objects. Does this not unfairly decide what must exist?

Well, maybe. Perhaps our abstractionists could argue that their conception of possible singular terms does not drag in any unwanted metaphysical presuppositions. They do not mean that possible singular terms exist on a par with tables, chairs, etc. The mere fact that we can conceive of a singular term is enough. This seems to be implied by their use of (SP_3) . It is then my contention that, of the three subsidiary theses that comprise (SP) , (SP_3) is the only one that is unproblematic and open to use in other language-first metaontologies.

3.3 Conclusion

In this chapter, we laid out the metaontology of Bob Hale and Crispin Wright's abstractionism. We found that it relied on one important principle: (SP) . A number of objections to it were raised. We saw that our abstractionists use it, in part, to justify the truth of (HP) . Unfortunately, the argument they presented is viciously circular. Then we took to examining the subsidiary theses of (SP) . The first two faced quite separate issues: (SP_1) motivates a syntactic test for singular termhood that is ultimately unusable. And (SP_3) — the most metaphysically substantive subsidiary thesis — has plausible counterexamples taken from the history of science.

So, what's the upshot? In the next chapter we will lay the foundations for a new language-first metaontology for abstractionism — one that we hope will meet Hale and Wright's major realist goals and does not conflict with their epistemological picture. Of course, ours will be different. It *has* to be: given the results of this chapter, we must do away with the syntactic priority thesis and develop a different way to decide ontological questions about numbers by appeal to language. More specifically, we reject (SP_1) and (SP_2) , retain (SP_3) and endorse a principle that will secure the existence of numbers as mind-independent objects, given certain linguistic facts.

3.4 Bibliography

- Benacerraf, P. (1973). "Mathematical Truth." *The Journal of Philosophy*, Vol.70(19): 661-679.
- Dummett, M. (1973). *Frege: Philosophy of Language*. London: Duckworth.
- Everett, A. (2005). "Against Fictional Realism." *Journal of Philosophy*, 102(12): 624-649.
- Hale, B. (2001a). "Singular Terms (1)." In *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*, edited by B. Hale and C. Wright: 31-47. Clarendon, Oxford: Oxford University Press.
- Hale, B. (2001b). "Singular Terms (2)." In *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*, edited by B. Hale and C. Wright: 48-71. Clarendon, Oxford: Oxford University Press.
- Hale, B., and C. Wright. (2009). "The Metaontology of Abstraction." In *Metametaphysics: New Essays on the Foundations of Ontology*, edited by D. J. Chalmers, D. Maney, and R. Wasserman: 178-212. Oxford: Oxford University Press.
- Linnebo, Ø. (2018). *Thin Objects: An Abstractionist Account*. Oxford: Oxford University Press.
- Macbride, F. (2003). "Speaking with Shadows: A Study of Neo-Logicism." *The British Journal of Philosophy*, (54): 103-163.
- Russell, B. (1905). "On Denoting." *Mind*, 14(4): 473-493.
- Thomasson, A. L. (1999). *Metaphysics and Fiction*. Cambridge: Cambridge University Press.

Voltolini, A. (2003). "How Fictional Works Are Related to Fictional Entities." *Dialectica*, 57(2): 225–238.

Chapter 4

Coherentist-Plus Minimalism

4.1 Introduction

The purpose of the present chapter is to salvage whatever can be salvaged of Hale and Wright’s metaontology, and to *lay the foundations* for a new one. This new metaontology will serve the epistemological and metaphysical goals of Hale and Wright’s abstractionism; i.e., to provide an account of arithmetical knowledge and mathematical realism based on *(HP)* that does not rely on mystical mathematical intuitions or a prior knowledge of the abstracts introduced by *(HP)*. But it will avoid the pitfalls of the old metaontology. In the main, it will retain a *prima facie* reading of *(HP)*, employ *(SP₃)* as an explanation of objecthood, and provide a non-circular argument for the truth of *(HP)*.

To be sure, a variety of abstractionist metaontologies already exist, and it is not our purpose merely to add one to their number. To take some prominent examples: (Linnebo 2018) develops an account of *dynamic abstraction* fitted with an innovative metaontology that makes use of a “new” ontological category, *thin objects*. (Rayo 2013) develops a metaontology that relies heavily on *just-is* sentences; i.e., sentences that are true just in case the two sentences flanking either side of “just is” are “full and accurate descriptions of the *same feature of reality*” (Rayo 2013, 5), which lends itself to the abstractionist. Take, e.g., “For the number of Jupiter’s moons to be four *just is* for there to be four moons of Jupiter.” According to Rayo, this statement is true, because each sentence flanking either side of “just is” describe the same state of affairs. Hence,

if the right-hand-side is true, so is the left-hand-side. There is then no need to provide an auxiliary metaphysical explanation for the existence of numbers.¹

Following a comment made in (Linnebo 2018, 7), and the example of (Shapiro 1997), we develop a version of *coherentist minimalism* — an approach that has not been implemented for the abstractionist program. A paradigmatic instance of coherentist minimalism is the view held by David Hilbert. According to it, the consistency of a mathematical theory suffices for the existence of the objects that the theory “talks about.”² As he writes in a letter to Frege (12/29/1899):

As long as I have been thinking... on these things, I have been saying... if the arbitrarily given axioms do not contradict each other with all their consequences, then they are true and the things defined by them exist. This is for me the criterion of truth and existence. (Frege 1980, 39)

Clearly, Hilbert’s metaontology is rather permissive — *too* permissive. Consistent theories of abstraction that govern moobles, goobles, monads, etc., will expand our mathematical ontology into the silly and fanciful. We need not be so liberal. More can be required for truth and existence. Hence, we term our position, *coherentist-plus minimalism*, which is committed to the following claim:

If a theory of abstraction \mathcal{T} is *coherent-plus*, then the sentences of the theory are true and the singular terms featured in the theory refer.

In what follows, we will apply a precisified version of this principle to a theory of abstraction that is of great relevance to the abstractionist program, therefore providing an account of arithmetical knowledge and mathematical realism based on (*HP*). The remaining portions of this chapter are structured as follows: §4.2 is devoted an explication of *coherence-plus*, and an argument for a version of the sufficiency claim given above. Broadly speaking, the argument is this: We start with a coherentist-plus account of justification for the acceptability of a new theory of abstraction. By noting our limited epistemological position, this account implies a coherentist-plus theory of truth, which in turn yields our sufficiency claim. In §4.3, we carry out the application

¹The import for (*HP*) should be clear to the reader.

²One should not be confused by the substitution of “consistency” for “coherence”. Any advocate of coherentist minimalism must define or explain “coherence” somehow. Accordingly, coherence just is consistency for Hilbert.

mentioned above but in stages: (i) we provide arguments for the specific conditions that the theory of abstraction in question must meet to be coherent-plus, and then (ii) we show that this theory satisfies those conditions. §4.4 deals with some possible objections and is followed by some concluding remarks in §4.5.

4.2 Sufficiency Claims and Coherentist-Plus Minimalism

The metaontology that we develop here is a form of *metaontological minimalism*: the view that substantive metaphysical questions can be answered fairly easily.³ One distinguishing feature of any metaontological minimalism is a commitment to a *sufficiency claim*, whose general form can be expressed as follows:

- ϕ only if ψ ,
- ϕ suffices for ψ ,
- what it takes for ψ to be true is exactly what it takes for ϕ to be true,
- all it takes for ψ is ϕ , etc.,

where the *prima facie* ontological commitments of ψ outstrip the ontological commitments of ϕ . Here are some examples:

- All that is required for the existence of directions, for example, is the parallelism of appropriate lines. (Linnebo 2018, 23)
- **coherence:** If Φ is a coherent formula in a second-order language, then there is a structure that satisfies Φ . (Shapiro 1997, 95)

Sufficiency claims are suspect precisely because the ontological commitments of ψ outstrip those of ϕ . Therefore, they require a solid defense and explanation. To aid in the defense of our sufficiency claim, we should first state it a little more precisely. Suppose we have higher-order formal language \mathcal{L} defined in the usual way. Thus, the set of \mathcal{L} -terms is the collection of expressions of \mathcal{L} that function “like names”. We let $\mathcal{W}_{\mathcal{L}}$ be the set of well-formed formulas of \mathcal{L} . Suppose that \mathcal{T} is a theory of \mathcal{L} :

³Metaontological minimalism is sometimes called “mild deflationism.” (Manley 2009, 4).

$$\phi \in \mathcal{T} \text{ if and only if } T \models \phi,$$

where ϕ is a sentence. We say that Σ is *the set of axioms* of \mathcal{T} just in case

$$\mathcal{T} = \{\phi \in \mathcal{W}_{\mathcal{L}} : \Sigma \models \phi\}.$$

Take $\mathcal{A}_{\mathcal{T}}$ to be the set of axioms of \mathcal{T} , which contains abstraction principles only; i.e., \mathcal{T} is a theory of abstraction.

Lastly, we say that \mathcal{T} is *true* when and only when for all $\phi \in \mathcal{A}_{\mathcal{T}}$, ϕ is true and each \mathcal{L} -term that occurs in ϕ refers. It should be noted that *truth* and *reference* are to be understood to have their pretheoretical sense, not some model-theoretic analogue.⁴

For our purposes, \mathcal{T} represents an idealized, new mathematical belief system about the entities purportedly named by the novel \mathcal{L} -terms introduced by the axioms of \mathcal{T} . The set $\mathcal{A}_{\mathcal{T}}$, therefore, represents the set of basic beliefs in this system.

Our sufficiency claim is then:

(C+) If \mathcal{T} is *coherent-plus*, then \mathcal{T} is true.

Obviously, (C+) codifies the idea that coherence-plus is sufficient for truth and referential success. As such, it is the defining principle of this metaontological minimalism.

The argument for C+, which we will call *Auxiliary*⁵, is deceptively simple:

1. If \mathcal{T} is *coherent-plus*, then \mathcal{T} is acceptable.
2. If \mathcal{T} is acceptable, then \mathcal{T} is true.
3. Hence, if \mathcal{T} is *coherent-plus*, then \mathcal{T} is true.

Undoubtedly, *Auxiliary* is valid. But is it sound? The next two subsections will be devoted to a defense of each premise. We take these in turn.

⁴One might regard the idea that elements of a formal language can be true or refer in the pretheoretical sense with suspicion. This might be due to the fact that the elements of formal languages do not have meanings in the way that natural languages do. If it helps, we can take the formulas and terms of \mathcal{L} to be abbreviations of the relevant English sentences and names. Hence, they carry the appropriate meanings by stipulation.

⁵I chose this name, because it acts as an auxiliary argument to an argument that we will give later on.

4.2.1 Auxiliary: Premise 1

What underlies the first premise in *Auxiliary* is some theory of justification for the acceptability of \mathcal{T} . Clearly, this needs to be explained. Moreover, the notion of *coherence-plus*, and its role in this theory should be explicated.

So, when is a new theory of abstraction like \mathcal{T} acceptable? We start by thinking about our own epistemic community, of which we make certain assumptions. Its members hold any number of empirical and mathematical theories (including set theory and model theory) to be true — or, at least, we have no good reason to doubt their truth. Each theory is well-established, according to some predetermined criteria: In the case of the empirical theories, significant empirical corroboration and predictive/explanatory power was enough. In the case of mathematical theories, perhaps applicability, richness of results, etc., sufficed.

Suppose further that our community has every confidence in the judgements of its scientists and mathematicians. Hence, we take scientific and mathematical statements at face value. No semantic error theory need apply. No one is wanting for some nominalistic translation scheme.

Given our scientific maturity, the community also rejects non-naturalistic explanations of knowledge acquisition. There is then no mystical mathematical intuition that we can appeal to to justify the acceptability of \mathcal{T} nor may we appeal to any antecedent a priori knowledge of the entities purportedly named by the singular terms of \mathcal{T} — they are “new,” after all. So, an intuition-based or foundationalist theory of justification is out of the question.

From this epistemological position, there is one recourse of action open to us: if our community wants to do as little linguistic and logical violence as possible to its pre-existent mathematical and scientific theories and add \mathcal{T} to its storehouse of knowledge, we can — at least — devise a set of criteria that would guarantee that \mathcal{T} is coherent, in some sense. Therefore, a coherence theory of justification for the acceptability of \mathcal{T} is adopted.

Generally speaking, what should this theory of justification require? Firstly, we would expect that \mathcal{T} “coheres with itself.” Thus, \mathcal{T} ought to be model-theoretically *satisfiable* — very roughly put, \mathcal{T} ought to be possibly true. Trivial theories are of no use to a community whose goal is genuine knowledge. One might balk at the idea

of using model theory to test for possible truth, *simpliciter*. But recall: set theory and model theory rank among our accepted theories. Both have yielded a wealth of impressive results. And we have been entitled to their use in the past. There is no reason to preclude their use in this case — especially since set theory and model theory have served this very purpose before.

Secondly, we ought to require that \mathcal{T} cohere with any satisfiable theory \mathcal{T}^* that we do hold or *would* hold. That way we may add \mathcal{T} to our storehouse of knowledge with the guarantee that there is no danger of it conflicting with any such theory. There are two ways that \mathcal{T} might conflict with \mathcal{T}^* : First, their union might not be model-theoretically satisfiable. Second, \mathcal{T} might have some bearing on the subject matter of \mathcal{T}^* that it should not have. If, say, \mathcal{T}^* is an empirical theory about electrons, \mathcal{T} should not dictate how many electrons exist.

These two broad requirements capture the idea that \mathcal{T} should be *intratheoretically* coherent and *intertheoretically* coherent, respectively. (Thus, the “coherence” in “coherence-plus.”) But this is not the full story. More should be required for acceptability lest we are willing to accept moobles and the like into our mathematical ontology.

Recall that we are interested in knowledge acquisition, and knowledge is factive. So, we ought to stipulate further conditions that \mathcal{T} must satisfy to be acceptable — conditions that would capture the “behavior” we would expect of \mathcal{T} *if* it were true. Since the axioms of \mathcal{T} are abstraction principles — a special type of implicit definition — we might expect, e.g., that \mathcal{T} specify the structure of the abstracts it governs, or that it settles all identity statements featuring the terms it introduces. Which specific conditions would be required to meet this goal do not matter at this point. What matters is that, with said conditions, we place higher demands on an acceptable theory of abstraction that characterize the behaviors we would expect it to have if it were true. (Hence, the “plus” in “coherence-plus.”)

Consequently, both types of conditions are meant to guarantee two facts: first, that \mathcal{T} is both intratheoretically and intertheoretically coherent; second, that the terms introduced by \mathcal{T} behave as if they refer. When a theory does meet these conditions, it is *coherent-plus* and well-behaved in the following sense: \mathcal{T} will have all the *linguistic* trappings of a true mathematical theory. And so, our community may regard it as

acceptable.

The first premise of *Auxiliary* is, therefore, highly plausible. What more can be demanded of \mathcal{T} ? Note that coherence-plus requires a high cost of admission: total linguistic cooperation with any theory, and (apparent) cooperation with the extralinguistic world.

4.2.2 Auxiliary: Premise 2

Now we turn to the argument for the second premise of *Auxiliary*. At first blush, it appears obviously false and raises many questions, the most obvious of which is: how could the acceptability of a theory of abstraction be sufficient for its truth? Unpacking further, how could the mere fact that such a theory is well-behaved in the the sense described above guarantee that it is true?

We might be able to bypass this troubling question easily. If we were to adopt some sort of idealism, then the second premise would not be needed. We could follow Putnam during his period of *internal realism*, and argue that truth is an epistemic notion; i.e., idealized justification. Truth is then reduced to verifiability conditions under which “true” can be applied to an indicative sentence in ideal circumstances. Furthermore, we can challenge the external realist conception of objecthood and existence; as Putnam says:

...signs do not intrinsically correspond to objects, independently of how these signs are employed and by whom. But a sign is actually employed in a particular way by a particular community of users. ‘Objects’ do not exist independently of conceptual schemes. We cut up the world into objects when we introduce one or another scheme of description. (Putnam 1981, 53)

This is an attractive option for a few reasons: First, with a little intellectual elbow grease, *Auxiliary*’s conclusion falls right out of Putnam’s picture. Just note that coherence-plus is essentially a set of verifiability conditions employed in an idealized situation.

Second, Putnam’s claim that “ ‘Objects’ do not exist independently of conceptual schemes. We *cut up the world into objects* when we introduce one or another scheme of description” is reminiscent of Frege’s *content-recarving*. Hence, this view lends itself

to an interpretation of content-recarving in terms of conceptual schemes/theories. And we can use that to secure truth for any theory of abstraction.

But this option is not open to us. The reason for this is straightforward: We are interested in providing a *realist* interpretation of abstractionism. Putnam's view, at least at this time, was decidedly antirealist across the board. However interesting an idealistic philosophy of mathematics might be, we are not in the business of providing a metaontology for one.

With this option rejected, some justification must be given for the second premise. Providing such a justification seems impossible upon reflection. Why? Because there is a strong intuition that truth and referential success require cooperation between two things: language and the world beyond language. Accordingly, reality can be divided along two lines; i.e., the linguistic and the extralinguistic. The linguistic is, in part, subject to convention and a matter of chance evolutionary development. Extralinguistic reality, on the other hand, is as it is — crystalline, fixed. The world must first supply us with entities to refer to and speak truthfully of. It follows that facts about language alone cannot guarantee truth or referential success.

Undoubtedly, this intuition is rooted in our experience of naming and interacting with physical objects. Take, for instance, Alpha Centauri A. The justification for its existence involves observation or experiment. Therefore, we must go beyond language to first learn of it. It is only *after* we have this empirical justification that we have any confidence that “Alpha Centauri A” refers, and that the sentences containing this name are either true or false.

Thinking by analogy with physical objects is a mistake, though. Physical objects make substantial demands on the world for their existence. They require arrangements of matter, energy, a spatiotemporal location, etc. So, the picture described re referential success and truth in the case of physical objects is totally sensible. Abstract objects, by definition, are different in this regard. As Shapiro writes:

To be sure, *abstracta*, mathematical objects are not located in space and time. That is at least part of what it is to be an *abstractum*. It does not follow from this, however, that these objects are located somewhere else, in a Platonic Heaven. *Abstracta*... are not located *anywhere*, whether in the world of being or in the world of becoming. (Shapiro 2011, 23).

This is to say that abstract objects exist “outside” of space and time but reside nowhere. They do not interact with the constituents of the physical universe. They do not feature in the causal traffic at all. Hence, abstract objects in no way ontologically depend on the physical world. It is reasonable to infer then that the entities purportedly named by the terms introduced by a new mathematical theory make no *typical* substantial demands on the world for their existence. These objects are, as (Linnebo 2018) puts it, *thin objects*. Consequently, any *reasonable* justification for the existence of abstracta is of a different sort.⁶

What sort? No one in our community can occupy a God’s-eye-view position by which they can check all the constituents of the world. Even the existence of certain physical objects, e.g., four-dimensional objects (if they exist), are beyond cognitive reach. Regarding abstract objects, we are even more epistemologically limited. Thus, the only reasonable justification for the existence of the objects purportedly named by the terms introduced in a new theory of abstraction available to us is found in language. What else can be appealed to? The best our community can do is to demand that the theory in question be well-behaved — that it has all the linguistic trappings of a true theory; that it be acceptable. This way we can be sure that, on the linguistic side of things, everything works out as expected. And our community has the best justification *available to them* for the truth of a new mathematical theory.

Indeed, to demand some *further* justification for the truth of such a theory is unreasonable. Requiring more evidence is tantamount to asking the proponent of this view to do the impossible; i.e., to step outside themselves and somehow check the constituents of the world. No one should be held to such a high standard of justification, including the nominalist.

We have provided good reason to accept the truth of *Auxiliary*’s first and second premise. We can, at this point, take **C+** on as the foundational principle of coherentist-plus minimalism, and move on to applying it in §4.3.

⁶Of course, Linnebo’s position suggests an ontological dependency relation: it is enough for objects of a kind *K* to exist that *K*-terms appear in true sentences that are supplied with an appropriate predicative identity criterion. We need not go so far. Whatever demands abstracta make of the world can remain unknown to us. Rather, our appeal to thin objects is different: it is used to motivate, along with Shapiro’s comments, what counts as a reasonable justification for the existence of the objects purportedly named by terms introduced by a new mathematical theory.

4.3 The Case of Abstractionism

Following the picture outlined in the previous section, this portion of the chapter will deal with abstractionism; i.e., we will consider a formal theory of abstraction that is of special importance to the program, and show that it is *coherent-plus*. Given $\mathbf{C}+$, it would then follow that any sentence in this theory is true, and the singular terms it introduces do refer. But, before we can formulate the theory of abstraction properly, we need to establish a technical framework.

4.3.1 Our Logical Apparatus

We start by building the third-order language \mathcal{L} . This is accomplished in two steps: first, we define, what we will call, the *logical base language of \mathcal{L}* ; second, we extend the base language by adding some special symbols to its vocabulary and formulating new rules for generating members of this extended language. We then stipulate that this extension just is \mathcal{L} .

Let's assume that we are given a countably infinite number of distinct objects, called *symbols*, which comprise the *logical vocabulary* of the logical base language of \mathcal{L} .⁷

- Logical Connectives: $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$.
- Identity symbol: $=$.
- Quantifier: \forall, \exists .
- First-order variables: v_1, v_2, \dots .
- Second-order n -place relation variables: $\mathbf{X}_1^1, \mathbf{X}_2^1, \dots, \mathbf{X}_1^2, \mathbf{X}_2^2, \dots$.
- Second-order n -place function variables: $\mathbf{f}_1^1, \mathbf{f}_2^1, \dots, \mathbf{f}_1^2, \mathbf{f}_2^2, \dots$.
- Third-order one-place function variables: $\mathbf{F}_1, \mathbf{F}_2, \dots$.
- Punctuation: $(,), [,]$.

⁷We will use $\mathbf{x}, \mathbf{y}, \mathbf{z}$ for distinct first-order variables, $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, for distinct second-order relation variables, $\mathbf{f}, \mathbf{g}, \mathbf{h}$ for distinct second-order function variables, and $\mathbf{F}, \mathbf{G}, \mathbf{H}$ for distinct third-order variables (with or without subscripts) when there is no danger of ambiguity.

The *expressions*, *terms*, and *atomic formulas* are defined recursively as usual, as are the *well-formed formulas*, *open formulas*, and *sentences* of this base language.

Now we extend the base language by adding some *non-logical vocabulary* to it, which will include an *abstraction operator* $@_E$ for every purely logical second-order equivalence relation E^8 that occurs in an abstraction principle that we are considering. Thus, given that we will be utilizing a formalized version of *(HP)* in this language, the third-order function symbol $\#$ is among these expressions. And the equivalence relation it is paired with is *equinumerosity*: \cong .⁹ The expressions, terms, well-formed formulas of this extended language \mathcal{L} etc., are defined as you would expect them to be, making use of the base language. And so, we can formulate *(HP)* in \mathcal{L} like so:

$$\text{(HP): } (\forall \mathbf{X})(\forall \mathbf{Y})(\#(\mathbf{X}) = \#(\mathbf{Y}) \leftrightarrow \mathbf{X} \cong \mathbf{Y}).$$

Some special subsets of \mathcal{L} will be needed to express the concepts that we will discuss later. Consequently, we adopt the following conventions:

Definition 1. $\Phi; \Gamma = \{\Phi\} \cup \Gamma$, where $\Phi \in \mathcal{L}$ and $\Gamma \subseteq \mathcal{L}$.

Definition 2. Let $\Gamma \subseteq \mathcal{L}$ and ϵ is an expression of \mathcal{L} . Then $\Gamma \setminus \epsilon$ is the set of all $\Phi \in \mathcal{L}$ such that ϵ does not occur in Φ .

We assume the full¹⁰ standard semantics: A higher-order structure $M = \langle \Delta, \mathcal{I} \rangle$ is comprised of a set $\Delta \neq \emptyset$ of objects (or *individuals*), and an *interpretation function* \mathcal{I} that assigns elements of the non-logical vocabulary to objects of the relevant type. So, once an abstraction operator $@_E$ is interpreted, $\mathcal{I}(@_E) : P(\Delta) \rightarrow \Delta$.

A *variable assignment* s (for M) is a function from the set of all variables to objects of the appropriate type; i.e.,

⁸Strictly speaking, there are no equivalence relations *in* \mathcal{L} nor are there any equivalence relation symbols. So, taken literally, what is written here is quite misleading. Instead, what we mean to say is E is a purely logical open formula — one that defines an equivalence relation once interpreted using higher-order semantics.

⁹ $\mathbf{X} \cong \mathbf{Y}$ abbreviates the purely logical second-order formula expressing the existence of a bijection between the objects assigned to \mathbf{X} and \mathbf{Y} .

¹⁰On the full semantics, the second-order *Comprehension Schema* is valid: For any open formula $\Phi(\mathbf{y})$ not containing \mathbf{X} free,

$$(\exists \mathbf{X})(\forall \mathbf{y})(\mathbf{X}(\mathbf{y}) \leftrightarrow \Phi(\mathbf{y})).$$

$$\begin{aligned}
s(\mathbf{x}) &\in \Delta; \\
s(\mathbf{X}_i^k) &\subseteq \Delta^k; \\
s(\mathbf{f}_i^k) &: \Delta^k \rightarrow \Delta; \\
s(\mathbf{F}_i) &: P(\Delta) \rightarrow \Delta.
\end{aligned}$$

The satisfaction conditions for formulas and, hence, the truth conditions for sentences of \mathcal{L} are defined in the typical way, along with *logical consequence* and other related semantic notions. In keeping with the standard notation, we adopt the following for our purposes:

Definition 3. $\models_M \Phi[s]$ iff M satisfies Φ with a variable assignment s . (In the special case where ϕ is a sentence, we omit “[s].”)

Definition 4. $\models_M A$ iff for every $\Phi \in A$, $\models_M \Phi[s]$.

Definition 5. $A \models \Phi$ iff Φ is a *logical consequence* of A .

There is one more technical notion that will be of great import to our investigation: *relativization*; i.e., on occasion, we will need to consider formulas whose quantifiers have been *relativized* to a given predicate:

Definition 6. Let $\Phi \in \mathcal{L}$ and suppose that $\psi(\mathbf{x})$ is a unary predicate of \mathcal{L} . Then $[\Phi]^{\psi(\mathbf{x})}$ is the *relativization* of Φ , defined recursively like this:

- $[\alpha]^{\psi(\mathbf{x})} = \alpha$ if α is atomic.
- $[\neg\alpha]^{\psi(\mathbf{x})} = \neg[\alpha]^{\psi(\mathbf{x})}$.
- $[(\alpha \square \beta)]^{\psi(\mathbf{x})} = ([\alpha]^{\psi(\mathbf{x})} \square [\beta]^{\psi(\mathbf{x})})$ if $\square \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$.
- $[(\forall \mathbf{x})\alpha(\mathbf{x})]^{\psi(\mathbf{x})} = (\forall \mathbf{x})(\psi(\mathbf{x}) \rightarrow [\alpha(\mathbf{x})]^{\psi(\mathbf{x})})$.
- $[(\exists \mathbf{x})\alpha(\mathbf{x})]^{\psi(\mathbf{x})} = (\exists \mathbf{x})(\psi(\mathbf{x}) \wedge [\alpha(\mathbf{x})]^{\psi(\mathbf{x})})$.
- $[(\forall \mathbf{X}^n)\alpha(\mathbf{X}^n)]^{\psi(\mathbf{x})} = (\forall \mathbf{X}^n)(\forall \mathbf{y}_1)(\forall \mathbf{y}_2) \dots (\forall \mathbf{y}_n)((\mathbf{X}^n(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \rightarrow (\psi(\mathbf{y}_1) \wedge \psi(\mathbf{y}_2) \wedge \dots \wedge \psi(\mathbf{y}_n))) \rightarrow [\alpha(\mathbf{X}^n)]^{\psi(\mathbf{x})})$.

- $[(\exists \mathbf{X}^n)\alpha(\mathbf{X}^n)]^{\psi(\mathbf{x})} = (\forall \mathbf{X}^n)(\forall \mathbf{y}_1)(\forall \mathbf{y}_2) \dots (\forall \mathbf{y}_n)((\mathbf{X}^n(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) \rightarrow (\psi(\mathbf{y}_1) \wedge \psi(\mathbf{y}_2) \wedge \dots \wedge \psi(\mathbf{y}_n))) \wedge [\alpha(\mathbf{X}^n)]^{\psi(\mathbf{x})})$.
- $[(\forall \mathbf{f}^n)\alpha(\mathbf{f}^n)]^{\psi(\mathbf{x})} = (\forall \mathbf{f}^n)(\forall \mathbf{y}_1)(\forall \mathbf{y}_2) \dots (\forall \mathbf{y}_n)((\psi(\mathbf{y}_1) \wedge \psi(\mathbf{y}_2) \wedge \dots \wedge \psi(\mathbf{y}_n)) \rightarrow (\exists z)(\psi(\mathbf{x}) \wedge \mathbf{f}^n(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) = z)) \rightarrow [\alpha(\mathbf{f}^n)]^{\psi(\mathbf{x})}$.
- $[(\exists \mathbf{f}^n)\alpha(\mathbf{f}^n)]^{\psi(\mathbf{x})} = (\exists \mathbf{f}^n)(\forall \mathbf{y}_1)(\forall \mathbf{y}_2) \dots (\forall \mathbf{y}_n)((\psi(\mathbf{y}_1) \wedge \psi(\mathbf{y}_2) \wedge \dots \wedge \psi(\mathbf{y}_n)) \rightarrow (\exists z)(\psi(\mathbf{x}) \wedge \mathbf{f}^n(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n) = z)) \wedge [\alpha(\mathbf{f}^n)]^{\psi(\mathbf{x})}$.
- $[(\forall \mathbf{F})\alpha(\mathbf{F})]^{\psi(\mathbf{x})} = (\forall \mathbf{F})(\forall \mathbf{X}^1)((\forall \mathbf{y})(\mathbf{X}^1(\mathbf{y}) \rightarrow \psi(\mathbf{y})) \rightarrow (\exists \mathbf{z})(\psi(\mathbf{z}) \wedge \mathbf{F}(\mathbf{X}^1) = \mathbf{z})) \rightarrow [\alpha(\mathbf{F})]^{\psi(\mathbf{x})}$.
- $[(\exists \mathbf{F})\alpha(\mathbf{F})]^{\psi(\mathbf{x})} = (\exists \mathbf{F})(\forall \mathbf{X}^1)((\forall \mathbf{y})(\mathbf{X}^1(\mathbf{y}) \rightarrow \psi(\mathbf{y})) \rightarrow (\exists \mathbf{z})(\psi(\mathbf{z}) \wedge \mathbf{F}(\mathbf{X}^1) = \mathbf{z})) \wedge [\alpha(\mathbf{F})]^{\psi(\mathbf{x})}$.

Loosely put, the relativization of a formula Φ to some unary predicate $\psi(\mathbf{x})$ (which is possibly complex *or* a second-order variable) expressing the proposition that Φ is true of the elements of $\Sigma \subseteq \Delta$, where Σ is defined by $\psi(\mathbf{x})$. The clauses for the quantified formulas are, clearly, the most important. So, to aid comprehension, here are natural language formulations of some of them:

If $\Phi = (\forall \mathbf{x})\alpha(\mathbf{x})$, then $[\Phi]^{\psi(\mathbf{x})}$ says:

For every object x such that x is ψ , x is α .

If $\Phi = [(\forall \mathbf{X}^n)\alpha(\mathbf{X}^n)]^{\psi(\mathbf{x})}$, then $[\Phi]^{\psi(\mathbf{x})}$ says:

For every n -ary relation R and any objects y_1, \dots, y_n , if R holds of y_1, \dots, y_n only if each y_i is ψ , then...

If $\Phi = [(\forall \mathbf{F})\alpha(\mathbf{F})]^{\psi(\mathbf{x})}$, then $[\Phi]^{\psi(\mathbf{x})}$ says:

For every third-order function f , any concept X , and any object y , if y is X only if y is ψ , then there is some object z that is ψ and $f(X) = z$ only if...

Sometimes we will need to discuss collections of relativized formulas. So, lastly, we have:

Definition 7. $A^{\psi(\mathbf{x})} = \{\Phi^{\psi(\mathbf{x})} : \Phi \in \mathcal{A}\}$.

4.3.2 Metatheory

A word on metatheory: For the remainder of this chapter, we assume for our metatheory the class-set theory of NBG with choice. NBG is a particularly elegant conservative extension of ZFC.¹¹ So, by utilizing it, we can appeal to familiar theorems regarding ordinals, cardinals, etc., that we all know and love.

Though I assume the reader has at least a cursory knowledge of the basic set-theoretic facts we will use, I think it is best to explicitly state them for the sake of clarity, along with any linguistic conventions we will adopt.

Accordingly, the set of natural numbers $\omega = \{0, 1, 2, \dots\}$ and the class of ordinal numbers $\Theta = \{0, 1, 2, \dots, \omega, \dots\}$ are defined following Von Neumann's construction. Thus, ω is the smallest set containing \emptyset that is closed under the *successor function* (defined on all sets):

$$x + 1 = x \cup \{x\}.$$

The series of natural numbers is then:

$$\begin{aligned} 0 &= \emptyset, \\ 1 &= 0 + 1, \\ 2 &= 1 + 1, \\ 3 &= 2 + 1, \end{aligned}$$

etc. And Θ is the smallest (proper) class that includes ω and is closed under the successor function and chain unions.¹² Moreover, Θ (and consequently, ω) is *doubly class-transitive*¹³, and well ordered by inclusion. But we will use another metalinguistic symbol when speaking of this well-ordering: for each $i, k \in \Theta$,

$$i \leq y \text{ iff } \subseteq k, \text{ and}$$

$$i < k \text{ iff } i \subset k.$$

¹¹And it is my preferred set theoretic apparatus.

¹²A *chain* is a set A such that, for each $x, y \in A$, either $x \subseteq y$ or $y \subseteq x$. A set A is *closed under chain unions* just in case, for any chain $x \subseteq z, \bigcup x \in z$.

¹³A set A is *doubly class-transitive* iff for any $x \in A$ and any $y \in x$, $x \subseteq A$ and $y \subseteq x$.

Thus, we have the following:

$$\begin{aligned} i &\notin i; \\ i &< i + 1; \\ i &< k \text{ iff } i \in k; \\ i &< k, k < i, \text{ or } i = k, \end{aligned}$$

where exactly one of these disjuncts holds.

Let A and B be sets. Functions and their properties will play a big role in what is to come. So, we should get straight on some concepts and notation that will be employed when speaking of them. If $h : A \rightarrow B$, then

$$\text{Dom}(h) = \{x \in A : \text{for some } y, y = h(x)\}.$$

We stipulate that $A = \text{Dom}(h)$ always. And

$$\text{Ran}(h) = \{y \in B : \text{for some } x, h(x) = y\}$$

Obviously, $\text{Ran}(h) \subseteq B$ but it need not be identical to B . We say that A is *finite* just in case for some $i \in \omega$, there is a bijection $f : i \rightarrow A$ ¹⁴; otherwise, we say A is *infinite*. It is plain to see that when such a bijection exists, there is no bijection $g : k \rightarrow A$ if $k < i$. Now, since we are assuming the axiom of choice, it follows that there is some $k \in \Theta$ such that there is a bijection $f : A \rightarrow k$. The least such k (least w.r.t. \leq , of course) is the *cardinality of A* , denoted by “ $|A|$ ”. The (proper) class of all $|A|$ is \aleph — the class of *cardinal numbers*. Therefore, A is finite when $|A| \in \omega$ and A is infinite when $|A| \in \aleph \setminus \omega = \aleph^\infty$.

It will be helpful to simplify talk of bijections by stipulating that:

$$A \approx B \text{ iff there exists a bijection } f : A \rightarrow B;$$

$$A \preceq B \text{ iff for some } C \subseteq B, \text{ there is a bijection } f : A \rightarrow C;$$

$$A \prec B \text{ iff } A \preceq B \text{ but it is not the case that } A \approx B.$$

¹⁴Or if $f : A \rightarrow i$, since the the inverse of any bijection is a bijection.

which gives us:

$$\begin{aligned} A \approx B &\text{ iff } |A| = |B|; \\ A \prec B &\text{ iff } |A| < |B|; \\ A \preceq B &\text{ iff } |A| \leq |B|. \end{aligned}$$

And, for each $c, d \in \mathbb{N}$,

$$\begin{aligned} |c| &= c \text{ if } c \in \omega; \\ |c + 1| &= c \text{ if } c \notin \omega; \\ c \prec d &\text{ iff } c < d; \\ c \preceq d &\text{ iff } c \leq d. \end{aligned}$$

We will have some occasion to reference a special type of set “generated” by an equivalence relation. So, take \mathcal{E} to be an equivalence relation on A , and let $x \in A$. Then the *equivalence class of x* (w.r.t. A and \mathcal{E}) is

$$[x]_A^{\mathcal{E}} = \{y \in A : \langle x, y \rangle \in \mathcal{E}\}.$$

It follows that

$$\langle x, y \rangle \in \mathcal{E} \text{ iff } [x]_A^{\mathcal{E}} = [y]_A^{\mathcal{E}}.$$

The collection of all $[x]_A^{\mathcal{E}}$ is $A/\mathcal{E} — A \text{ modulo } \mathcal{E}$.

The notation for equivalence classes might strike the reader as rather cumbersome. Consequently, we write “[x]” instead of “[x] $_A^{\mathcal{E}}$ ” when context deems the latter unnecessary.

Finally, we will use “ \Rightarrow ” to abbreviate “implies” and “ \Leftrightarrow ” to abbreviate “if and only if”. Doing so will make some of the proofs to follow more readable. This concludes our detour into metatheory.

4.3.3 The Theory of Abstraction and Coherence-Plus

With our logical apparatus in hand, we can move forward with our metaontological investigation. Our focus will be on the formal theory of abstractionism: So, take $\mathcal{T}^\#$ to

be the theory of \mathcal{L} whose sole axiom is **HP**. Hence, we need some set of conditions that this theory must satisfy in order to be coherent-plus, according to this metaontology. More specifically, we need a class of conditions \mathcal{C} such that its members:

1. Secure the intratheoretical and intertheoretical coherence of $\mathcal{T}^\#$, and
2. characterize the behavior of $\mathcal{T}^\#$ if it were true.

Now, since **HP** is the sole axiom of $\mathcal{T}^\#$, we know that for any sentence $\Phi \in \mathcal{L}$,

$$\Phi \in \mathcal{T}^\# \text{ iff } \mathbf{HP} \models \Phi.$$

It then suffices to show that **HP** alone satisfies all the conditions belonging to \mathcal{C} : **HP** meets all these conditions only if all Φ such that $\mathbf{HP} \models \Phi$ do too. So, the elements of \mathcal{C} need only target **HP**. And if **HP** satisfies these conditions, $\mathcal{T}^\#$ is *coherent-plus*; i.e., true and all the singular terms that it introduces refer, by **C+**.

Let's ask ourselves the following question: What *specific* conditions should **HP** satisfy in order to be *coherent-plus*?

Incidentally, much work has been done already concerning the general acceptability of **HP** in trying to formulate a satisfactory response to the *Bad Company Problem*:

The abstractionist ought to provide a philosophically principled account that draws the line between acceptable abstraction principles and unacceptable abstraction principles.

This issue arises because two abstraction principles — **HP** and **BLV** — share the same logical form, but the former is inconsistent (in any second-order deduction system containing a sufficiently strong comprehension principle among its axioms) while the latter is not.

To deal with the Bad Company Problem, a myriad of conditions have been offered, each of which can be placed into one of four categories. We will call this delineation the *Good Company 4-fold division*:

1. Identity Conditions

- (a) The Caesar Constraint

- (b) The $\mathbb{C}\mathbb{R}$ Constraint

2. Conservativeness

- (a) Caesar-Neutral Conservativeness
- (b) Field Conservativeness

3. Modesty

- (a) Satisfiability
- (b) Strong Stability
- (c) Modest Reflection
- (d) Modest Logical Reflection

4. Categoricity

- (a) The Categoricity Constraint

This table represents the extant philosophical and technical work re the Bad Company Problem. Hence, our task has become two-fold: first, to pick conditions from this list that are appropriate for our project (if they exist) and provide philosophical justifications for our choices; second, to demonstrate that **HP** does, in fact, satisfy these conditions.

To begin, we note that **HP** must be intratheoretically coherent — in other words, satisfiable. Our justification for this is as simple as can be: No theory that we add to our storehouse of knowledge should admit falsehoods; otherwise, we are not in the business of knowledge acquisition. Moreover, any such theory should be discerning. Intuitively, an unsatisfiable theory amounts to a description of the world that licenses *all* descriptions of the world (since an unsatisfiable theory includes its base language). This will not do.

It was shown in (Boolos 1987) showed that **HP** is consistent if second-order arithmetic is. But even if **HP** were plain old consistent, semantic completeness fails for second-order logic (assuming the standard semantics).¹⁵ So, we cannot move from consistency to satisfiability. Luckily, models of **HP** are fairly easy to come by. Observe: We merely pick the following higher-order structure $M = \langle \omega + 1, I \rangle$, where

¹⁵Semantic completeness *does* hold for second-order logic paired with Henkin semantics, though.

$$\omega + 1 = \{0, 1, 2, \dots, \omega\},$$

and

$$\mathcal{I}(\#) : P(\omega + 1) \longrightarrow \omega + 1,$$

such that

$$\mathcal{I}(\#)(\alpha) = \begin{cases} n & \text{if } n \approx \alpha \text{ and } n \in \omega; \\ \omega & \text{if } \omega \approx \alpha. \end{cases}$$

With minimum tedium, one can show that $\models_M \mathbf{HP}$.

But this is not interesting. It is a well-known logical fact that **HP** is satisfiable — in fact, and as we will see, **HP** is satisfiable on all and only infinite domains. Perhaps there is a neater and more mathematically interesting way to obtain intratheoretical coherence — one that also gets us intertheoretical coherence, too.

We must ask ourselves if there are any conditions in the *Good Company 4-fold division* that resemble the intuitive notion behind intertheoretical coherence: that if $\mathcal{T}^\#$ is acceptable, it must not interfere with *any* possible theory that we have in, or might add to, our storehouse of knowledge. And we must also ask ourselves whether intertheoretical coherence suffices for intratheoretical coherence. The answer to both questions is: yes.

An initial way to technically codify intertheoretical coherence would be to require that **HP** be *inferentially conservative*: For the remainder, we let $\mathcal{A}_{@_E}$ be an abstraction principle such that $@_E$ is its abstraction operator.

Definition 8. $\mathcal{A}_{@}$ is *inferentially conservative* if and only if, for any $\Phi; \mathcal{T} \subseteq \mathcal{L} \setminus @$, where \mathcal{T} is a theory,

$$\mathcal{A}_{@}; \mathcal{T} \models \Phi \text{ only if } \mathcal{T} \models \Phi.$$

This might seem like a good idea, but it isn't. Consider the theory $\mathcal{T}^{\geq 2}$, whose only axiom is:

$$(\exists \mathbf{x})(\exists \mathbf{y}) \neg \mathbf{x} = \mathbf{y}.$$

It is easy to see that $\mathcal{T}^{\geq 2} \subseteq \mathcal{L}$, and $\Psi \in \mathcal{L} \setminus \#$, where

$$\Psi = (\exists \mathbf{x})(\exists \mathbf{y})(\exists \mathbf{z})[(\neg \mathbf{x} = \mathbf{y} \wedge \neg \mathbf{x} = \mathbf{z}) \wedge \neg \mathbf{y} = \mathbf{z}].$$

However, $\mathbf{HP} \models \Psi$ and $\mathcal{T}^{\geq 2} \not\models \Psi$: First, we observe that \mathbf{HP} is satisfiable on only infinite domains. Thus, any structure that satisfies it must have an infinite domain, which will contain at least three objects. So, any such structure will satisfy Ψ . But if we take a structure with *exactly* two objects in its domain, it will satisfy $\mathcal{T}^{\geq 2}$ (since it will satisfy $(\exists \mathbf{x})(\exists \mathbf{y})\neg \mathbf{x} = \mathbf{y}$) but not Ψ .

This counterexample demonstrates that if we adopt inferential conservativeness as an acceptability condition, we would have to rule out $\mathcal{T}^\#$, since there is a theory whose only axiom expresses the proposition that there exists at least two things. Obviously, this is a ludicrous consequence. Inferential conservativeness is, therefore, just too strong a requirement to be reasonable.

(Weir 2003) introduced several constraints on the acceptability of abstraction principles, two of which are *Caesar-Neutral Conservativeness* and *Field Conservativeness* (both can be found in the *Good Company 4-fold division*). Stated formally:

Definition 9. $\mathcal{A}_{@_E}$ is *Caesar-Neutral Conservative* if and only if, for any $\Phi; \mathcal{T} \subseteq \mathcal{L} \setminus @$, where \mathcal{T} is a theory, and any primitive unary predicate $\psi(\mathbf{x})$ not occurring in any element of $\Phi; \mathcal{T}$,

$$\mathcal{A}_{@_E}; \mathcal{T}^{\psi(\mathbf{x})} \models [\Phi]^{\psi(\mathbf{x})} \text{ only if } \mathcal{T} \models \Phi.$$

C-CON is the set of Caesar-Neutral Conservative abstraction principles.

Definition 10. $\mathcal{A}_{@}$ is *Field Conservative* if and only if, for any $\Phi; \mathcal{T} \subseteq \mathcal{L} \setminus @$, where \mathcal{T} is a theory,

$$\mathcal{A}_{@}; \mathcal{T}^{\neg(\exists \mathbf{Y})\mathbf{x}=@(\mathbf{Y})} \models [\Phi]^{\neg(\exists \mathbf{Y})\mathbf{x}=@(\mathbf{Y})} \text{ only if } \mathcal{T} \models \Phi.$$

F-CON is the set of Field Conservative abstraction principles.

These notions make different demands of acceptable abstraction principles. The first, Caesar-Neutral Conservativeness, is fairly weak. The rough idea behind it is this: Suppose we start with a theory \mathcal{T} , and we take $\psi(\mathbf{x})$ to be the predicate which specifies the “subject matter” of \mathcal{T} . The nature of the subject matter need not be known: It is possible that it overlaps with the abstracts introduced by the abstraction principle or

that the subject matter is wholly mathematical. We remain *neutral* on this question. Thus, this constraint mandates that, when an abstraction principle is combined with \mathcal{T} restricted to its subject matter, nothing is implied about the subject matter that isn't implied about the universe excluding the (possibly new) abstracts. In other words, the abstraction principle in question does not interfere with the subject matter of \mathcal{T} , whatever it might be.

The requirements of Field Conservativeness are stronger but similar. Instead of remaining neutral regarding the subject matter of the theory, we assume that there is *no* overlap with the abstracts introduced by the abstraction principle and the subject matter of the theory. Hence, the subject matter of the theory *could be* empirical. By similar reasoning, this constraint mandates that, when an abstraction principle is combined with \mathcal{T} restricted to its subject matter, nothing is implied about the subject matter that isn't implied about the universe as a whole.

Given the foregoing explanations, it should be fairly obvious that these two constraints are just what we are looking for. What's at stake is, after all, the *aprioricity* of an abstraction principle. If it is acceptable and successful as an implicit definition, it should have no bearing on the nature of physical objects or other mathematical entities. As Wright says:

A legitimate abstraction, in short, ought to do no more than introduce a concept by fixing truth conditions for statements concerning instances of that concept . . . How many sometime, someplace zebras there are is a matter between that concept and the world. No principle which merely assigns truth-conditions to statements concerning objects of a quite unrelated, abstract kind — and no legitimate second-order abstraction can do any more than that — can possibly have any bearing on the matter. What is at stake . . . is, in effect, *conservativeness* in (something close to) the sense of that notion deployed in Hartry Field's exposition of his nominalism.¹⁶ (Wright 1997, 296)

Conservativeness, of either sort, also gets us out of an initial form of Weir's *Embarrassment of Riches Objection* (ER): there are indefinitely many consistent but mutually inconsistent abstraction principles. Some of these principles place an upper bound on the size of the domain of individuals. Any *a priori* mathematical principle cannot do so. Obviously, this is an important ancillary benefit for us.

¹⁶My emphasis.

But unfortunately there are unacceptable abstraction principles that do satisfy both kinds of conservativeness, as is shown in (Cook 2012). Luckily, the *strong stability* condition blocks this problem. Therefore, we will have to adopt it too if we are to avoid further embarrassment.¹⁷

Definition 11. $\mathcal{A}_{@_E}$ is *strongly stable* if and only if there is some $\gamma \in \mathbb{N}$ such that, for all $k \geq \gamma$, $\mathcal{A}_{@_E}$ is satisfied on a structure of size k iff $k \geq \gamma$.

To further our study of these two conditions, we need to beef up our technical vocabulary. These definitions will fall into two sorts of categories: those that rely solely on cardinality, and those that rely on cardinality plus something more. The first sort will suffice for our understanding of Caesar-Neutral Conservativeness:

Definition 12. $\mathcal{A}_{@_E}$ is *k-satisfiable* if and only if there is a model $M = \langle \Delta, \mathcal{I} \rangle$, where $|\Delta| = k$ and $\models_M \mathcal{A}_{@}$.

\mathbf{SAT}^k is the set of all *k-satisfiable* abstraction principles. Clearly, $\mathbf{SAT}^k \subseteq \mathbf{SAT}$, where \mathbf{SAT} is the set of all satisfiable abstraction principles.

Definition 13. $\mathcal{A}_{@_E}$ is an *∞ -abstraction principle* if and only if, for any $k \in \mathbb{N}^\infty$, $\mathcal{A}_{@} \in \mathbf{SAT}^k$.

\mathbf{A}^∞ is the set of all ∞ -abstraction principles.

Definition 14. $\mathcal{A}_{@_E}$ is *unbounded* if and only if, for all $\gamma \in \mathbb{N}$, there is some $k \in \mathbb{N}$, such that $\gamma \leq k$ and $\mathcal{A}_{@} \in \mathbf{SAT}^k$.

\mathbf{UNB} is the set of all unbounded abstraction principles. And \mathbf{UNB}^∞ is the set of all ∞ -abstraction principles that are unbounded.

(Cook 2024) has noted that the majority of work on the Bad Company problem has focused on conditions that can be imposed on an abstraction principle by considering only the class of cardinalities at which the abstraction principle is satisfied (like those

¹⁷We will not provide a proof that demonstrates that \mathbf{HP} is strongly stable. Instead, we note that the reasoning in the proof of *unboundedness* and the fact that \mathbf{HP} is satisfied on at least one structure of size ω is enough.

above). But, as Cook points out, to understand Field Conservativeness, we need to consider the cardinals at which an abstraction principle is satisfiable *and* the number of abstracts that are guaranteed to exist on domains of those cardinalities. Hence, we have the following:

Definition 15. $\mathcal{A}_{@_E}$ is *k-full* if and only if $\mathcal{A}_{@_E} \in \mathbf{SAT}^k$ and for any higher-order structure $M = \langle \Delta, \mathcal{I} \rangle$ such that $|\Delta| = k$ and $\models_M \mathcal{A}_{@_E}$

$$|\{x \in \Delta : \text{for some } Y \subseteq \Delta, x = \mathcal{I}(@)(Y)\}| = k.$$

Loosely expressed: $\mathcal{A}_{@_E}$ is *k-full* just when it is *k-satisfiable*, and for any structure that satisfies it whose domain contains *k* many things, its domain will also contain *k* many abstracts of the kind introduced by $\mathcal{A}_{@}$ — or the number of abstracts will always match the cardinality of the domain for a *k-satisfiable* abstraction principle.

FULL^{*k*} is the set of all *k-full* abstraction principles.

Definition 16. *k* is a *critical point* of $\mathcal{A}_{@_E}$ if and only if $\mathcal{A}_{@_E} \in \mathbf{SAT}^k$, and there is a $\gamma < k$ such that, for all λ where $\gamma \leq \lambda < k$, $\mathcal{A}_{@_E} \notin \mathbf{SAT}^\lambda$.

$\text{crit}(\mathcal{A}_{@})$ is the set of critical points of $\mathcal{A}_{@}$.

Definition 17. $\mathcal{A}_{@_E}$ is *weakly critically full* if and only if for any $k \in \text{crit}(\mathcal{A}_{@_E})$, there is a $\gamma \geq k$ such that $\mathcal{A}_{@_E} \in \mathbf{FULL}^\gamma$.

WC-FULL is the set of all weakly critically full abstraction principles. And

WC-FULL ^{∞} is the set of all ∞ -abstraction principles that are weakly critically full.

With the addition of this technical vocabulary, we can get to the task at hand. The logical relationships between (some of) these conditions can now be expressed — and proved. Furthermore, we will show that **HP** satisfies the requisite conditions.

Theorem 1. $\mathbf{UNB} \subseteq \mathbf{SAT}$.

Proof. Let $\mathcal{A}_{@} \in \mathbf{UNB}$. Then for all $\gamma \in \aleph$, there is some $k \in \aleph$ such that $\gamma \leq k$ and $\mathcal{A}_{@} \in \mathbf{SAT}^k$. Thus, $\mathcal{A}_{@} \in \mathbf{SAT}$. \square

Theorem 2. $\mathbf{UNB} \subseteq \mathbf{C-CON}$.

Proof. Suppose that $\mathcal{A}_{@E} \in \mathbf{UNB}$. Let $\Phi; \mathcal{T} \subseteq \mathcal{L} \setminus @E$, where \mathcal{T} is a theory, and let $\psi(\mathbf{x})$ be a primitive unary predicate that does not occur any element of $\Phi; \mathcal{T}$.

We will prove the contrapositive. So, let's assume that $\mathcal{T} \not\equiv \Phi$. Thus, there is a structure $M_1 = \langle \Delta_1, \mathcal{I}_1 \rangle$ such that $|\Delta_1| = k$ and $\models_{M_1} \mathcal{T}$ but $\not\models_{M_1} \Phi$.

By hypothesis, there is a structure whose domain is of cardinality $\gamma \geq k$ that satisfies $\mathcal{A}_{@E}$. We pick the least such $\gamma \in \mathbb{N}^\infty$ and let $M_2 = \langle \Delta_2, \mathcal{I}_2 \rangle$ be a structure such that $|\Delta_2| = \gamma$ and $\models_{M_2} \mathcal{A}_{@E}$.

Next, we construct a special structure that will satisfy $\mathcal{A}_{@E}; \mathcal{T}^{\psi(\mathbf{x})}$ but not $\Phi^{\psi(\mathbf{x})}$. So, let $M_3 = \langle \Delta_3, \mathcal{I}_3 \rangle$, where

$$\Delta_3 = \Delta_1 \cup \Delta_2.$$

Notice that $|\Delta_3| = |\Delta_2|$, since $|\Delta_1| \leq |\Delta_2|$ and $|\Delta_1 \cup \Delta_2| = |\Delta_1| + |\Delta_2| = |\Delta_2|$. Thus,

$$\begin{aligned} |P(\Delta_3)| &= |P(\Delta_2)|; \\ P(\Delta_3) &\approx P(\Delta_2). \end{aligned}$$

Let f be a such a bijection. For each $\alpha, \beta \in P(\Delta_3)$, we define

$$\mathcal{I}_3(\#)(\alpha) = \mathcal{I}_2(\#)(f(\alpha)),$$

and

$$\langle \alpha, \beta \rangle \in E^{M_3} \Leftrightarrow \langle f(\alpha), f(\beta) \rangle \in E^{M_2},$$

where E^{M_i} is the relation assigned to E in the structure M_i . It follows that

$$\begin{aligned} \mathcal{I}_3(\#)(\alpha) = \mathcal{I}_3(\#)(\beta) &\Leftrightarrow \mathcal{I}_2(\#)(f(\alpha)) = \mathcal{I}_2(\#)(f(\beta)); \\ &\Leftrightarrow \langle f(\alpha), f(\beta) \rangle \in E^{M_2}; \\ &\Leftrightarrow \langle \alpha, \beta \rangle \in E^{M_3}. \end{aligned}$$

To complete the definition of \mathcal{I}_3 , we stipulate that $\mathcal{I}(\psi(\mathbf{x})) = \Delta_2$, and when restricted to the non-logical vocabulary of \mathcal{T} , \mathcal{I}_3 agrees with \mathcal{I}_1 completely.

Given all this, we obtain:

$$\models_{M_3} \mathcal{A}_{@_E}; \mathcal{T} \text{ and } \not\models_{M_3} \Phi.$$

Lastly, with an induction proof (which is left to the reader), the following can be demonstrated: for any sentence $\delta \in \mathcal{L} \setminus @_E$,

$$\models_{M_3} \delta \Leftrightarrow \models_{M_3} [\delta]^{\psi(x)}.$$

And so,

$$\models_{M_3} \mathcal{A}_{@_E}; \mathcal{T}^{\psi(x)} \text{ and } \not\models_{M_3} [\Phi]^{\psi(x)}.$$

That is,

$$\mathcal{A}_{@_E}; \mathcal{T}^{\psi(x)} \not\models [\Phi]^{\psi(x)}.$$

□

This proof can be amended to show that $\mathbf{UNB}^\infty \subseteq \mathbf{C-CON}$

Theorem 3. $\mathbf{A}^\infty \cap \mathbf{UNB}^\infty \cap \mathbf{WC-FULL}^\infty \subseteq \mathbf{F-CON}$.

Proof. Suppose that $\mathcal{A}_{@_E} \in \mathbf{A}^\infty \cap \mathbf{UNB}^\infty \cap \mathbf{WC-FULL}^\infty$. Let $\Phi; \mathcal{T} \subseteq \mathcal{L} \setminus @_E$, where \mathcal{T} is a theory, and let $\psi(\mathbf{x})$ be a primitive unary predicate that does not occur any element of $\Phi; \mathcal{T}$.

We will prove the contrapositive. So, let's assume that $\mathcal{T} \not\models \Phi$. Thus, there is a structure $M_1 = \langle \Delta_1, \mathcal{I}_1 \rangle$ such that $|\Delta_1| = k$ and $\models_{M_1} \mathcal{T}$ but $\not\models_{M_1} \Phi$.

Since $\mathcal{A}_{@_E} \in \mathbf{UNB}^\infty$, we know that there is a structure whose domain is of cardinality $\gamma > k$ that satisfies $\mathcal{A}_{@_E}$. We pick the least $\gamma \in \mathbb{N}^\infty$ and let $M_2 = \langle \Delta_2, \mathcal{I}_2 \rangle$ be a structure such that $|\Delta_2| = \gamma$ and $\models_{M_2} \mathcal{A}_{@_E}$.

Either $\gamma = k$ or $\gamma > k$.

Assume $\gamma = k$. Then we define the structure $M_3 = \langle \Delta_3, \mathcal{I}_3 \rangle$, where

$$\Delta_3 = \Delta_1 \cup \{x \in \Delta_2 : \text{for some } Y \subseteq \Delta_2, x = \mathcal{I}_2(@_E)(Y)\}$$

and \mathcal{I}_3 restricted to the non-logical vocabulary of \mathcal{T} is \mathcal{I}_1 and $\mathcal{I}_3(@_E)$ is any function $f; P(\Delta_3) \rightarrow \Delta_2$ such that for any $\alpha, \beta \subseteq \Delta_3$, $f(\alpha) = f(\beta) \Leftrightarrow \langle \alpha, \beta \rangle \in E^{M_3}$, which we know exists given the reasoning in the previous proof. Thus, $\models_{M_3} \mathcal{A}_{@_E}; \mathcal{T}$ and $\not\models_{M_3} \Phi$.

Now, assume $\gamma > k$. Hence, $\gamma \in \text{crit}(\mathcal{A}_{@_E})$. Let γ_2 be the least cardinal such that $\gamma_2 \geq \gamma$ and $\mathcal{A}_{@_E} \in \mathbf{FULL}^{\gamma_2}$. Let $M_2 = \langle \Delta_2, \mathcal{I}_2 \rangle$, where $|\Delta_2| = \gamma_2$ and $\models_{M_2} \mathcal{A}_{@_E}$. We take $M_3 = \langle \Delta_3, \mathcal{I}_3 \rangle$, where $\Delta_3 = \Delta_1 \cup \Delta_2$, \mathcal{I}_3 restricted to the non-logical vocabulary of \mathcal{T} is \mathcal{I}_1 , and $\mathcal{I}_3(@_E)$ is any surjective function $f : P(\Delta_3) \rightarrow \Delta_2$ such that, for any $\alpha, \beta \subseteq \Delta_3$, $f(\alpha) = f(\beta) \Leftrightarrow \langle \alpha, \beta \rangle \in E^{M_3}$. Therefore, Thus, $\models_{M_3} \mathcal{A}_{@_E}; \mathcal{T}$ and $\not\models_{M_3} \Phi$.

Again, by induction it is established that for any sentence $\delta \in \mathcal{L} \setminus @_E$,

$$\models_{M_3} \delta \Leftrightarrow \models_{M_3} [\delta]^{\psi(x)}.$$

Consequently,

$$\models_{M_3} \mathcal{A}_{@_E}; \mathcal{T}^{\psi(x)} \text{ and } \not\models_{M_3} [\Phi]^{\psi(x)}.$$

That is,

$$\mathcal{A}_{@_E}; \mathcal{T}^{\psi(x)} \not\models [\Phi]^{\psi(x)}.$$

□

Thus, to show that **HP** is satisfiable, Caesar-Neutral Conservative, and Field Conservative, it suffices to show that **HP** is an ∞ -abstraction principle that is unbounded and weakly critically full.¹⁸

Theorem 4. $\mathbf{HP} \in \mathbf{A}^\infty \cap \mathbf{UNB}^\infty \cap \mathbf{WC-FULL}^\infty$.

Proof. It is enough to show that **HP** is a member of all three of these sets, which we prove individually.

¹⁸The proofs for theorems 2 and 3 were adapted from one of Roy T. Cook's unpublished manuscripts on the Bad Company problem.

(i) **HP** $\in \mathbf{A}^\infty$: Suppose that $M = \langle \Delta, \mathcal{I} \rangle$ is a higher-order structure such that $\models_M \mathbf{HP}$. For *reductio*, we assume that $|\Delta| = k \in \omega$: $\Delta = \{\delta_0, \delta_1, \dots, \delta_{k-1}\}$. By hypothesis (and some symbol pushing), we get that for any $\alpha, \beta \subseteq \Delta$,

$$\mathcal{I}(\#)(\alpha) = \mathcal{I}(\#)(\beta) \text{ iff } \alpha \approx \beta.$$

Now notice that a contradiction is easily obtained

if

$$k + 1 = |P(\Delta)/\approx| = |\text{Ran}(\mathcal{I}(\#))|,$$

since

$$\begin{aligned} \text{Ran}(\mathcal{I}(\#)) \subseteq \Delta &\Rightarrow \text{Ran}(\mathcal{I}(\#)) \preceq \Delta; \\ &\Rightarrow |\text{Ran}(\mathcal{I}(\#))| \leq |\Delta|; \\ &\Rightarrow k + 1 \leq k. \end{aligned}$$

And $k < k + 1$ - *per impossibile*. Thus, it is sufficient to show two things:

$$\text{(ia) } k + 1 = |P(\Delta)/\approx|, \text{ and}$$

$$\text{(ib) } |P(\Delta)/\approx| = |\text{Ran}(\mathcal{I}(\#))|.$$

For (ia): Consider the function

$$g : k + 1 \longrightarrow P(\Delta)^{[\approx]},$$

defined this way:

$$g(i) = \begin{cases} [\emptyset] & \text{if } i = 0; \\ [\{d_0, d_1, \dots, d_{i-1}\}] & \text{if } 0 < i \leq k, \end{cases}$$

It is easy to show that g is a bijection:

g is surjective: Let $y \in P(\Delta)/\approx$. Therefore, for some $\alpha \in P(\Delta)$, $y = [\alpha]$. So, $\alpha \subseteq \Delta$. Notice: either $\alpha = \emptyset$ or $\alpha \neq \emptyset$. If $\alpha = \emptyset$, then $y = [\emptyset] = g(0)$. If $\alpha \neq \emptyset$, then $\alpha = \{a_0, a_1, \dots, a_{j-1}\}$, where $0 < j \leq k$. Clearly, $\alpha \approx \{d_0, d_1, \dots, d_{j-1}\}$. Hence, $[\alpha] = [\{d_0, d_1, \dots, d_{j-1}\}]$. This means that $y = [\{d_0, d_1, \dots, d_{j-1}\}] = g(j)$.

g is injective: Suppose $i \neq j$ and $i, j \in k + 1$. We know that $i = 0$ or $i \neq 0$. If $i = 0$, then $g(i) = g(0) = [\emptyset]$. And $0 \neq j$. Hence, $0 < j \leq k$ and $g(j) = [\{\delta_0, \delta_1, \dots, \delta_{j-1}\}]$. But $[\emptyset] \neq [\{\delta_0, \delta_1, \dots, \delta_{j-1}\}]$. Thus, $g(i) \neq g(j)$.

Consequently, $k + 1 \approx |P(\Delta)/\approx|$, which implies (ia).

For (ib): Take the function

$$f : P(\Delta)/\approx \longrightarrow \text{Ran}(\mathcal{I}(\#)),$$

such that

$$f([\alpha]) = \mathcal{I}(\#)(\alpha).$$

This function is a bijection, too:

f is surjective: Let $y \in \text{Ran}(\mathcal{I}(\#))$. Thus, for some $\alpha \in P(\Delta)$, $\mathcal{I}(\#)(\alpha) = y$. But this implies that $[\alpha] \in P(\Delta)/\approx$ and $f([\alpha]) = y$

f is injective: Let $[\alpha], [\beta] \in P(\Delta)/\approx$. Suppose that $f([\alpha]) = f([\beta])$. So, $\mathcal{I}(\#)(\alpha) = \mathcal{I}(\#)(\beta)$. Hence, $\alpha \approx \beta$. Therefore, $[\alpha] = [\beta]$

Finally, $|P(\Delta)/\approx| = |\text{Ran}(\mathcal{I}(\#))|$, which implies (ib).

(ii) **HP** \in **UNB** ^{∞} : Let $\lambda \in \aleph^\infty$. We first note that

$$\begin{aligned} \alpha \in P(\lambda + 1) &\Rightarrow \alpha \subseteq \lambda + 1; \\ &\Rightarrow \alpha \preceq \lambda + 1; \\ &\Rightarrow |\alpha| \leq |\lambda + 1|; \\ &\Rightarrow |\alpha| \leq \lambda; \\ &\Rightarrow |\alpha| < \lambda \text{ or } |\alpha| = \lambda; \\ &\Rightarrow |\alpha| \in \lambda \text{ or } |\alpha| = \lambda; \\ &\Rightarrow |\alpha| \in \lambda + 1. \end{aligned}$$

So, we may take the higher-order structure $M^\lambda = \langle \lambda + 1, \mathcal{I}^\lambda \rangle$, where

$$\mathcal{I}^\lambda(\#)(\alpha) = |\alpha|.$$

And let $\alpha, \beta \subseteq \lambda + 1$. It follows that

$$\begin{aligned}\mathcal{I}^\lambda(\#)(\alpha) &= \mathcal{I}^\lambda(\#)(\beta); \\ |\alpha| &= |\beta|; \\ \alpha &\approx \beta.\end{aligned}$$

Hence, $\models_M \mathbf{HP}$, and since $|\lambda + 1| = \lambda$, $\mathbf{HP} \in \mathbf{SAT}^\lambda$. Given that λ was chosen arbitrarily, we know that for any $\lambda \in \aleph^\infty$, $\mathbf{HP} \in \mathbf{SAT}^\lambda$. Now, assume that $\gamma \in \aleph$. Thus, $\gamma \leq \omega$ or $\gamma > \omega$. If $\gamma \leq \omega$, then $\mathbf{HP} \in \mathbf{SAT}^\omega$. And if $\gamma > \omega$, $\mathbf{HP} \in \mathbf{SAT}^\delta$, for the least $\delta \in \aleph$ such $\delta > \gamma$. *Ergo*, \mathbf{HP} is unbounded.

(iii) $\mathbf{HP} \in \mathbf{WC-FULL}^\infty$: Let $k \in \text{crit}(\mathbf{HP})$. Thus, $\mathbf{HP} \in \mathbf{SAT}^k$, and there is some $\gamma < k$ such that, for all λ , where $\gamma \leq \lambda < k$, $\mathbf{HP} \notin \mathbf{SAT}^\lambda$.

And so, $\mathbf{HP} \notin \mathbf{SAT}^\gamma$. But recall that $\mathbf{HP} \in \mathbf{UNB}$. That means, $\gamma \notin \aleph^\infty$; otherwise, $\mathbf{HP} \in \mathbf{SAT}^\gamma$. So, $\gamma \in \omega$. Now, suppose that $k > \omega$. Then for some λ (i.e., ω) such that $\gamma \leq \lambda < k$, $\mathbf{HP} \in \mathbf{SAT}^\lambda$, contra previous results. Thus, $k = \omega$.

We know that $\mathbf{HP} \in \mathbf{SAT}^\omega$. It is, then, enough to show that $\mathbf{HP} \in \mathbf{FULL}^\omega$. To accomplish this, we let $M = \langle \Delta, \mathcal{I} \rangle$, where $|\Delta| = \omega$ and $\models_M \mathbf{HP}$, and prove that

$$|\{x \in \Delta : \text{for some } Y \subseteq \Delta, x = \mathcal{I}(\#)(Y)\}| = \omega.$$

Notice that the demonstration of (ib) in the previous sub-proof did not utilize the cardinality of the domain in question. Therefore, it applies to any structure that satisfies \mathbf{HP} , like M . It follows that

$$|P(\Delta)/\approx| = |\text{Ran}(\mathcal{I}(\#))|.$$

Clearly though,

$$\text{Ran}(\mathcal{I}(\#)) = \{x \in \Delta : \text{for some } Y \subseteq \Delta, x = \mathcal{I}(\#)(Y)\}.$$

Thus,

$$|P(\Delta)/\approx| = |\{x \in \Delta : \text{for some } Y \subseteq \Delta, x = \mathcal{I}(\#)(Y)\}|.$$

Now, take note:

$$\begin{aligned} w + 1 \approx P(\Delta)/\approx &\Rightarrow |w + 1| = |P(\Delta)/\approx| \\ &\Rightarrow \omega = |P(\Delta)/\approx|. \end{aligned}$$

All this implies that, for our purposes, it suffices to show:

$$\omega + 1 \approx P(\Delta)/\approx.$$

Consider the function

$$h : \omega + 1 \longrightarrow P(\Delta)/\approx, \text{ where}$$

$$h(i) = \begin{cases} [\emptyset] & \text{if } i = 0; \\ [\{d_0, d_1, \dots, d_{i-1}\}] & \text{if } 0 < i < \omega; \\ [\Delta] & \text{if } i = \omega. \end{cases}$$

We now show that h is a bijection to complete the proof:

h is surjective: Let $y \in P(\Delta)/\approx$. Then for some $\alpha \in P(\Delta)$, $y = [\alpha]$. Hence, $\alpha \subseteq \Delta$ - and

$$\begin{aligned} |\alpha| \leq |\Delta| &\Rightarrow |\alpha| \leq \omega; \\ &\Rightarrow |\alpha| < \omega + 1; \\ &\Rightarrow |\alpha| \in \omega + 1. \end{aligned}$$

Either $|\alpha| = 0$, $|\alpha| = i$, where $0 < i < \omega$, or $|\alpha| = \omega$. Taking each case in turn, we get our desired result:

$$\begin{aligned} |\alpha| = 0 &\Rightarrow |\alpha| = |\emptyset|; \\ &\Rightarrow \alpha \approx \emptyset; \\ &\Rightarrow [\alpha] = [\emptyset]; \\ &\Rightarrow h(|\alpha|) = h(0) = [\emptyset] = [\alpha] = y. \end{aligned}$$

$$\begin{aligned}
|\alpha| = i &\Rightarrow |\alpha| = |\{d_0, d_1, \dots, d_{i-1}\}|; \\
&\Rightarrow \alpha \approx \{d_0, d_1, \dots, d_{i-1}\}; \\
&\Rightarrow [\alpha] = [\{d_0, d_1, \dots, d_{i-1}\}]; \\
&\Rightarrow h(|\alpha|) = h(i) \\
&= [\{d_0, d_1, \dots, d_{i-1}\}] = [\alpha] \\
&= y.
\end{aligned}$$

$$\begin{aligned}
|\alpha| = \omega &\Rightarrow |\alpha| = |\Delta|; \\
&\Rightarrow \alpha \approx \Delta; \\
&\Rightarrow [\alpha] = [\Delta]; \\
&\Rightarrow h(|\alpha|) = h(\omega) = [\Delta] = [\alpha] = y.
\end{aligned}$$

h is injective: Let $i, j \in \omega + 1$. Suppose $h(i) = h(j)$. Then either $i = 0$, $i \in \omega$, or $i = \omega$. Again, taking each in turn, we get that $i = j$:

Assume $i = 0$. Then $h(i) = h(0) = [\emptyset] = h(j)$. If $j > 0$, then $h(j) = [\alpha]$, where $\alpha \neq \emptyset$, which is impossible. So, $i = 0 = j$.

Assume $i \in \omega$. Then $h(i) = [\{d_0, d_1, \dots, d_{i-1}\}] = h(j)$. If $j = 0$ or $j = \omega$, then $h(j) = [\emptyset]$ or $h(j) = [\Delta]$. So, $[\{d_0, d_1, \dots, d_{i-1}\}] = [\emptyset]$ or $[\{d_0, d_1, \dots, d_{i-1}\}] = [\Delta]$, which cannot be.

Assume $i = \omega$. Then $h(i) = [\Delta] = h(j)$. If $j < i$, then $h(j) = [\emptyset]$ or $h(j) = [\{d_0, d_1, \dots, d_{j-1}\}]$. Hence, $[\Delta] = [\emptyset]$ or $[\Delta] = [\{d_0, d_1, \dots, d_{j-1}\}]$ - contradiction.

□

So far, we have seen that **HP** satisfies conditions for both intratheoretical coherence and intertheoretical coherence. Now, it is time to consider what other conditions can be found in the *Good Company 4-fold division*. To remind the reader, these conditions should characterize the linguistic behavior of **HP** if it were true. This is to say: we must find conditions that capture our linguistic intuitions regarding the truth of **HP**.

One natural place to begin is with the *modesty* constraint that has been informally discussed by Wright in “Is Hume’s Principle Analytic?”:

An abstraction is Modest if its addition to any theory with which it is consistent results in no consequences - whether proof-or model-theoretically established - for the ontology of the combined theory which cannot be justified by reference to the consequences for its own abstracts. And again, *justification* is the crucial point: an abstraction may fail this constraint even though every consequence it has for the ontology of the combined theory may be seen to *follow from* things it entails about its proper abstracts; in particular, it will not count if, as in the case of the Limit-accessible Distraction, a consequence for the combined ontology is needed as a lemma in the proof that the abstracts have a property from which that very consequence follows. (Wright 1999, 30)

Admittedly, it is not clear what modestly amounts to given what is said here — especially when we consider that Wright makes a distinction between what an abstraction principle *justifies* and *entails* re reference to the consequences of its own abstracts. But the general idea seems to be this (if we ignore the mysterious distinction): Take any theory with which $\mathcal{A}_{@_E}$ is consistent. And suppose we combine it with $\mathcal{A}_{@_E}$. Then whatever this combined theory “says” about its ontology is justified (in *some way*), by virtue of the ontology introduced by $\mathcal{A}_{@_E}$. Such a requirement ought to strike one as too strong. If we have a theory whose ontology does not overlap with the abstracts introduced by $\mathcal{A}_{@_E}$, then the combined theory will imply things that are not justified in virtue of the ontology of abstracts. A more reasonable interpretation of modesty is offered in (Weir 2003, 30):

Definition 18. $\mathcal{A}_{@_E}$ *reflects modestly* if and only if, for every $\Phi \in \mathcal{L} \setminus @_E$,

$$\mathcal{A}_{@_E} \models \Phi \text{ only if } \mathcal{A}_{@_E} \models \Phi^{(\exists \mathbf{Y}) (\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$$

One can think of modesty then as mandating the following: Anything that $\mathcal{A}_{@_E}$ implies about the whole domain does so just if it implies the same thing about the ontology it introduces — an abstraction principle cannot say anything more.

This condition, however, is too loose: Weir rejects it as a necessary and sufficient condition for the acceptability of an abstraction principle, because of, what he calls,

distractions.¹⁹

Now, (Cook 2023) offers another interpretation of modesty, one that is analogous to the foregoing constraint:

Definition 19. $\mathcal{A}_{@_E}$ *modestly \mathcal{L} -reflecting* if and only if, for every purely logical formula $\Phi \in \mathcal{L}$,

$$\mathcal{A}_{@_E} \models \Phi \text{ only if } \mathcal{A}_{@_E} \models \Phi^{(\exists \mathbf{Y})(\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$$

MLR is the set of all abstraction principles that are modestly \mathcal{L} -reflecting.

The first difference that the reader will notice is that Φ must be purely logical. Everything else is the same. Still, Cook’s interpretation captures the idea behind modesty without admitting troublesome distractions.

A formal trick used by Cook will be helpful in understanding this condition and proving that **HP** satisfies it.²⁰

Definition 20. Let M be a structure whose domain is Δ , and assume $\alpha \subseteq \Delta$ and E is a purely logical equivalence relation on Δ . Then the *cardinality profile* of $\mathcal{A}_{@_E}$ is the function:

$$Prof_{@_E} : \aleph \longrightarrow \aleph,$$

such that

$$Prof_{@_E}(k) = |P(\Delta)/E|.$$

Definition 21. $\mathcal{A}_{@_E}$ is *idempotent* if and only if, for any k ,

$$Prof_{@_E}(k) \leq k \text{ only if } Prof_{@_E}(Prof_{@_E}(k)) \leq Prof_{@_E}(k).$$

ID is the class of idempotent abstraction principles.

¹⁹A *distraction* is an abstraction principle of the form

$$(\forall \mathbf{X})(\forall \mathbf{Y})((BAD(\mathbf{X}) \wedge BAD(\mathbf{Y})) \vee (\forall \mathbf{x})(\mathbf{X}(\mathbf{x}) \leftrightarrow \mathbf{Y}(\mathbf{x}))),$$

where $BAD(\mathbf{X}) = (\exists \mathbf{x})(\exists \mathbf{y})(\neg \mathbf{x} = \mathbf{y} \wedge \mathbf{X}(\mathbf{x}) \wedge \mathbf{Y}(\mathbf{y}) \wedge S)$ and S abbreviates the purely logical second-order claim that the size of the domain is a successor cardinal. Weir demonstrates that there are a number of clearly unacceptable distractions that reflect modestly. Hence, his rejection of modest reflection.

²⁰The following proofs have been adapted from (Cook 2023).

Theorem 5. ID \subseteq MLR.

Proof. Let $\mathcal{A}_{@E} \in \mathbf{ID}$, Φ be a purely logical formula such that $\mathcal{A}_{@E} \models \Phi$. We must show: $\mathcal{A}_{@E} \models \Phi^{(\exists \mathbf{Y})(\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$.

So, let $M = \langle \Delta, \mathcal{I} \rangle$ such that $\models_M \mathcal{A}_{@E}$ and $|\Delta| = k$. It suffices to show:
 $\models_M \Phi^{(\exists \mathbf{Y})(\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$.

Thus, $Prof_{@E}(k) \leq k$. By hypothesis then $Prof_{@E}(Prof_{@E}(k)) \leq Prof_{@E}$. But $Prof_{@E} = |Ran(\mathcal{I}(\#))|$, so there is a model $M_1 = \langle Ran(\mathcal{I}(\#)), \mathcal{I} \rangle$ such that $\models_{M_1} \mathcal{A}_{@E}$. Hence, $\models_{M_1} \Phi$. Note: Φ is purely logical. And so, $\models_{M_1} \Phi$ iff $\models_M \Phi^{(\exists \mathbf{Y})(\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$. Therefore, $\models_M \Phi^{(\exists \mathbf{Y})(\mathbf{x}=\mathcal{I}(\#)(\mathbf{Y}))}$. But M was arbitrarily chosen. □

Theorem 6. HP \in MLR.

Proof. Given the preceding theorem, we need only prove that **HP** \in **ID**. Since **HP** \in **A[∞]**, we know that

$$Prof_{\#}(k) \leq k \Leftrightarrow \aleph_0 \leq k.$$

By Frege's Theorem, it follows that

$$\aleph_0 \leq Prof_{\#}(k) \Leftrightarrow \aleph_0 \leq k.$$

Now, let $\gamma \in \aleph^\infty$. Hence, $Prof_{\#}(\gamma) \leq \gamma$. And so, $\aleph_0 \leq \gamma$. That means $\aleph_0 \leq Prof_{\#}(\gamma)$. Finally,

$$Prof_{\#}(Prof_{\#}(\gamma)) \leq Prof_{\#}(\gamma).$$

But γ was arbitrarily chosen. So, **HP** \in **ID**. □

4.3.4 The Three Cs: Categoricity, Caesar, and CR

We have just seen that **HP** satisfies several constraints that were deemed reasonable conditions for the acceptability of abstraction principles. It is time to consider the constraints that are deemed unreasonable on this account. There are three: the Categoricity Constraint, the $\mathbb{C}\mathbb{R}$ Constraint, and the Caesar Constraint (or *the three Cs*).

Taken in conjunction, these conditions mandate that an acceptable abstraction principle completely “fixes” the structure of the class of abstracts that it governs. We have two primary reasons for rejecting them: (i) The first constraint is motivated by a metaphysical intuition that need not be accounted for on our view. (ii) The remaining two constraints demand too much from a successful implicit definition.

We start with the Categoricity Constraint:

Any acceptable theory of abstraction \mathcal{T} must single out a unique structure instantiated by the abstracts purportedly referred to be the terms it introduces.

There are different ways to cash out this condition formally, the most natural of which is model-theoretic *categoricity*: any acceptable theory of abstraction ought to have one model up to isomorphism. But **HP** is not categorical in this sense. And if we hope to add $\mathcal{T}^\#$ to our storehouse of knowledge, we cannot accept this constraint as formulated. Efforts have been made to develop weaker conceptions of categoricity that capture the idea behind this constraint. There are two versions of permutation invariance that have been of particular interest: internal invariance and double invariance. (Cook 2017) has shown that **HP** does, in fact, have these properties.

Regardless of this technical achievement, the Categoricity Constraint might not seem like an important condition. Yet, a case can be made that it is one of — if not the — most substantive condition that an abstractionist of the Hale and Wright variety can hope for. To really get a sense of why this is the case, we must take our time to suss out the rough guiding metaphysical intuition behind it:

Suppose we have a theory of our solar system \mathcal{T}^\odot , which correctly describes the arrangement of the planets, their moons, etc., in English. Sentences like “Saturn follows Jupiter in our solar system” and “All planets revolve around the sun” are members of \mathcal{T}^\odot . Everyone agrees that \mathcal{T}^\odot is true and the singular terms that occur in its sentences refer. However one might philosophically cash out reference and truth, it seems plain that the world must first supply us with a definite arrangement of celestial bodies to refer to and speak truthfully of if this theory is true. More to the point: this arrangement is unique and determinate — and in virtue of this fact, the truth of \mathcal{T}^\odot is secured and it’s singular terms refer. Therefore, it is reasonable to suspect that any true theory will

do two things: (i) fix a unique structure; i.e., the one that it actually instantiates. And (ii) preclude alternate arrangements of its purported subject matter.

So, theories of abstraction should do the same *if* the abstractionist is a mathematical realist. But there is a sense in which the Categoricity Constraint has special bearing on abstractionism. How? Recall that, for the abstractionist, truth is prior to referential success, and referential success is prior to *objecthood*. If the foregoing explanation rightly captures the intuition behind the Categoricity Constraint, it follows that a fixed state of affairs is prior to truth — and thus, objecthood. Naturally then, the abstractionist ought to make sure that her acceptable abstraction principles specify a unique state of affairs. Two important things are at stake for her: a commonplace intuition about truth and reference, and a central thesis to her program — the objecthood of numbers!²¹

There are other ways in which an abstraction principle $\mathcal{A}_{@_E}$ might fail to fix its structure: It may fail to determine whether any abstract it governs is identical to an individual already in the domain. Put differently, $\mathcal{A}_{@_E}$ would not settle all *cross-categorical*, or *mixed*, identity statements — claims of the form:

$$t = @_E(\mathbf{X}),$$

where t is any term that is not of the form $@_E(\mathbf{Y})$. Hopefully, it is obvious to the reader that any $\mathcal{A}_{@_E}$ does *not* settle the truth (or falsity) of such statements, since it only provides identity conditions for statements of the form:

$$@_E(\mathbf{X}) = @_E(\mathbf{Y}).$$

Why might this be an issue for the abstractionist? The first problem is metaphysical: Recall that the abstractionist advocates a form of mathematical realism. Thus, she will want to separate the entities governed by her abstraction principle(s) from concrete objects already in the domain. She ought to “carve out” her platonistic ontology, by settling mixed identity statements that feature terms for concrete objects if there really are such things. But she would also do well to *specify* her ontology. If our abstractionist is interested in accepting more than one theory of abstraction, it is reasonable to assume that she ought to be able to tell whether their ontologies are disjoint. ²²

²¹This is my formulation of a philosophical justification for the Categoricity Constraint that was given to me by Roy T. Cook in private conversation. I use it with his permission.

²²(Hale and Wright 2001, 341) present a similar worry.

The second problem — or series of problems, rather — is/are semantical: (a) An abstraction principle is supposed to explain names of the abstracts it governs. More still, our grasp of an abstraction principle is supposed to explain our *capacity to refer* to said abstracts and classify them as a type of object. If this is so, then any abstraction principle $\mathcal{A}_{@_E}$ must enable us to understand predicates of the form $\mathbf{X} = @_E(\mathbf{Y})$, which are true *of objects*. Otherwise, reference to abstracta is doubtful.²³ (b) If the abstractionist cannot settle mixed identity statements, this constitutes a rejection of bivalence, which (presumably) she will not want to do.

To guard against these worries, the following two constraints have been proposed.

The Caesar Constraint:

Any acceptable theory of abstraction, when combined with an empirical theory governing non-abstracts ought to settle identity claims of the form $t = @_E(\mathbf{X})$, where t is a term governed by said empirical theory.

And the CR Constraint:

A theory of abstraction $\mathcal{T}^{@_{E_1}}$, when combined with another theory of abstraction $\mathcal{T}^{@_{E_2}}$ ought to settle identity claims of the form $@_{E_1}(\mathbf{X}) = @_E(\mathbf{Y})$.

Given the foregoing explanations, it would seem that the three Cs are very well motivated. So, what could possibly be our reasons for rejecting them? As was mentioned above, there are two primary reasons for doing so: (i) The first condition is motivated by a metaphysical intuition that need not be account for given our coherentist minimalism. (ii) The remaining two conditions place very high demands — too high, in fact — on a successful implicit definition. Let's see why.

The metaphysical intuition was illustrated with the example of \mathcal{T}^{\odot} . At the heart of it is a common sense (let's say) observation that truth and referential success require a definite arrangement of objects to be given first.²⁴ Thus, we ought to expect any true theory to preclude any alternate arrangements of its subject matter. Now, notice something: this intuition pump makes use of an empirical theory, the subject matter of which is planets, moons, etc. These are physical objects, of course. As such, their

²³This issue is raised in (Heck 1997).

²⁴We ignore worries about ontic vagueness and metaphysical indeterminacy.

existence makes substantial demands on the world. So too do the states of affairs that are “made of” concrete things. Such states of affairs seemingly require a unique and definite arrangement to exist — a change in the arrangement gets us a change in the state of affairs. So much seems obvious. But what we are suggesting by taking the coherentist-plus approach is that states of affairs made of abstracta might not require a *unique* arrangement for their existence. They are a different sort of entity. And we can come to learn of their existence by showing that a theory of abstraction meets minimal conditions and is true of *some* arrangement; i.e., a domain ordered in the appropriate way.²⁵

Furthermore, objecthood is not threatened by the rejection of the Categoricity Constraint. In the last chapter, we noted that Hale and Wright assumed (*SP*). Some of its subsidiary theses were found to be problematic. But there was one, i.e., (*SP*₃), that was found to be the least philosophically unpalatable. Why? Because it can be used as a reasonable general account of objecthood. According to it, linguistic categories are prior to ontological categories in the order of explanation. Hence, *x* is an object just in case there is a possible singular term that refers to it for Hale and Wright. We can also assume that linguistic categories are prior to ontological categories in the order of explanation. With this, the objecthood of the abstracta introduced by an acceptable theory of abstraction is secured with no cost.

Let’s move on to dealing with the final two constraints, Caesar and $\mathbb{C}\mathbb{R}$. The motivation for these conditions came in two forms: one metaphysical, the other semantical. We first noted that if an abstractionist advocates a form of mathematical realism, then she must be able to carve out and specify her mathematical ontology; i.e., she must be able to decide the truth (or falsity) of mixed identity statements. Otherwise, it seems she has not actually provided a full account of the abstract objects she wishes to countenance.

What’s interesting about this metaphysical issue is that it is particular to the abstractionist program. No other philosophy of mathematics is required to settle all identity statements featuring names for numbers. Often, it is merely assumed that numbers

²⁵Therefore, there is some affinity between the coherentist-plus view and the kind of *ante rem structuralism* advocated in (Shapiro 1997). The obvious difference between our view and Shapiro’s is that we do not take the subject matter of mathematics to be structures instantiated by relational systems. He does. The subject matter of mathematics, of us, are special types of objects introduced via abstraction.

are not people, places, pencils, etc., and that certain types of abstracta are different than other types of abstracta. This assumption is not open to the abstractionist because — presumably — she takes an abstraction principle to be an explanation of its associated mathematical concept. So, this metaphysical issue is really rooted in an epistemological worry — that abstractionism, as a philosophy of mathematics, does not offer sufficient explanations of mathematical concepts.

But how high of a standard ought the abstractionist put on herself and her chosen implicit definitions? The fact that an abstraction principle fixes the truth-conditions of *certain* identity statements implies that it serves to *partially* explain its associated mathematical concept. Perhaps this is enough for abstraction principles to be successful. After all, arithmetic, and other branches of mathematics, do not settle all mixed identity statements on their own. Arithmetic can tell you, e.g., that $2 + 2 = 4$ but not whether the conqueror of Gaul is 2. Thus, the abstractionist should be comfortable with this kind of indeterminacy. Moreover, in the case of **HP**, this indeterminacy suggests that our understanding of *cardinal number* so introduced captures our pre-theoretical understanding of cardinal numbers. It seems that the metaphysical motivation has lost its force if this analysis is correct.

Let's consider the semantical aspect of the motivation for the Caesar Constraint and the $\mathbb{C}\mathbb{R}$ constraint. The first issue mentioned was that (a) our grasp of an abstraction principle is meant to explain our capacity to refer to abstracts objects. Given this role, abstraction principles must enable us to understand predicates of the form $\mathbf{x} = @_E(\mathbf{Y})$, which are true *of objects*.

This problem is dealt with very easily. On the coherentist-plus account, abstraction principles are not meant to, and ought not, explain our capacity to refer to abstract objects. Reference to abstracta is obtained differently. We appeal to the behavior of theories of abstraction and their “interactions” with other theories. Indeed, the capacity to refer to abstract objects on Hale and Wright view does not assign this role to abstraction principles either, strictly speaking. What explains our capacity to refer to abstract objects, for them, is (*SP*) — that's it.

The second semantic issue mentioned was that (b) if the abstractionist cannot settle mixed identity statements, this constitutes a rejection of bivalence, which she would not want to do. Our answer to it: Well, not necessarily. It is possible to retain bivalence

while remaining silent re mixed identity. Consider for a moment *Goldbach's conjecture*: for any even $n \in \mathbb{N}$ such that $n > 2$, $n = p_1 + p_2$, where p_1, p_2 are primes. Surely, it is either true or false. Yet, we do not know which one. We have no proof or refutation of this claim. Perhaps either type of demonstration is out of our epistemic reach. That being said, we can still accept that this conjecture is either true or false — that there is a fact of the matter about it, given the axioms of arithmetic.

Let's adopt a similar attitude toward mixed identity statements: We can agree that there is a fact of the matter about whether, say, $\text{Caesar} = \#(\mathbf{x} \neq \mathbf{x})$, we just cannot come to know it by way of **HP**. Instead of rejecting bivalence then, we can accept some form of *epistemicism*.

And so, we have principled reasons for rejecting the three Cs. Incidentally, there is an added benefit to cutting the Categoricity Constraint loose: If we were to accept it in, e.g., the form that Cook advocates, then any acceptable abstraction principle mapping concepts to abstract objects will generate at most one abstract for each equivalence class. Consequently, there is no way to reconstruct set theory using a conceptual abstraction principle that sends each concept to an abstract that is (or “codes”) the set of objects whose members are exactly the objects falling under that concept. Since we do not require that any acceptable abstraction principle be categorical, reconstructing set theory by way of abstraction is a possibility.

This is a very welcome result: abstractionism is, if nothing else, an account of the foundations of mathematics. Set theory is one of the most successful and theoretically useful branches of contemporary mathematics. So, a reconstruction of set theory is, in some way, philosophically required of any view that deserves to call itself an account of the foundations of mathematics.

We will not be attempting to formulate an abstractionist theory of sets. Such an endeavor is the job of a different philosopher writing a different dissertation. But before we close, let's consider a number of possible objections to our view.

4.4 Possible Objections

The following objections are placed in no order of prominence or importance. As far as the viability of this view is concerned, they all ought to be answered.

(I) It can be argued that the existence of fictional entities follows from this metaontological approach. One can reformulate and generalize the coherence-plus conditions beyond theories of abstraction to theories that “talk about” fictional entities. If that is so and some theory of fictional entities is coherent-plus, then the proponent of this view must accept the existence of fictional characters, places, what have you. Hence, there must be something wrong with coherentist-plus minimalism.

This, of course, is a possibility. But it is not clear what the specific coherent-plus conditions would look like in the case of fictional theories. What would the intertheoretical non-inference conditions look like specifically? One is at a loss. So, it is doubtful that this is a *practical* possibility. Moreover, the existence of fictional entities is a problem *only if* one regards fictional entities with suspicion or it has been demonstrated (somehow) that they do not exist. Thus, the burden of proof is on the objector to show that fictional entities do not exist. Finally, someone that is inclined to apply **C+** to justify the existence of mathematical entities will, I believe, have no real issue with accepting the existence of other types of abstract entities. If this is so, then coherent-plus minimalism is a boon.

(II) At the end of the previous section, we observed that the rejection of the categoricity constraint opened the possibility of developing an abstractionist account of set theory. But, throughout this chapter, we have relied on set theory to aid us in determining that $\mathcal{T}^\#$ is coherent-plus. If the same set-theoretic resources are used to justify the acceptability of an abstractionist formulation of set theory, then it seems that we are committing some kind of “bootstrapping” fallacy or justifying the acceptance of set theory in a circular way.

This is a troubling worry but not intractable. There are a few different responses that can be given to this objection. First, we can embrace the bootstrapping strategy as inevitable and ultimately unproblematic. Contemporary set theory is one of the best tools that we have at our disposal for testing the formal properties of theories. And we are entitled to use the best tools at hand. Without it, it is unclear how we might proceed. Can we proceed at all? Thus, if the apparent circularity is an evil, it is a necessary evil.

Second, we could insist that by providing an abstractionist account of set theory, we are really providing a *reconstruction* of set theory. In other words, we are *not*, strictly speaking, using set theory to justify the acceptance of set theory. We are using set theory to justify the acceptance of a theory that behaves just like set theory — that’s it. And so, the apparent circularity disappears. Of the two choices for dealing with this objection, I find this second option to be favorable.

(III) The final objection that we will consider, a form of which was raised by Cook in private correspondence, is that the coherentist-plus approach does not capture the form of mathematical practice; i.e., theories and their axioms are, generally speaking, not justified by appeal to coherence conditions. Rather, axioms are stipulated to be true, and we deduce theorems from them. Confidence in the truth of the axioms is obtained in different ways by mathematicians, perhaps by the wealth of interesting theorems they yield or the applicability of said theorems.

As I see it, this objection is directed at the argument for *Auxiliary*’s first premise: if \mathcal{T} is coherent-plus, then it is acceptable. The argument involved making a number of assumptions about our community. One key assumption was that we are aware of our limited epistemological position re abstract objects, yet the community is interested in adopting a theory of abstraction in a philosophically principled way. Of course, it is highly dubious that these assumptions are true of most mathematicians. But that is no matter. They are doing mathematics, we are doing philosophy and foundations of mathematics. The enterprises are fundamentally different. Therefore, we need not capture the form of their practice exactly or even approximately.

4.5 Conclusion

The purpose of this chapter was to lay the foundations for an alternative metaontology for abstractionism. In doing so, we avoided the issues that plagued the metaontology of Bob Hale and Crispin Wright. How did we accomplish this? By developing a coherentist approach to the ontology of abstractionism. At the heart of this view is the all-important sufficiency claim $\mathbf{C+}$ — if a theory of abstraction \mathcal{T} is coherent-plus, then \mathcal{T} is true — which was demonstrated by *Auxiliary*. With it, we were able to bridge the gap between language and reality. And since we have demonstrated that $\mathcal{T}^\#$ is, in fact, coherent-plus,

it follows that this theory is true.

Now, this is not the end of the story. Much more work needs to be done. But hopefully I have convinced the reader that more work *should* be done to develop this position. Assuming this is so, one suggestion for future work comes to mind: some epistemological account needs to be developed further for this view. After all, we assume a coherentist theory of justification for the acceptance of a theory of abstraction. Yet, internal to any theory of abstraction we have a foundationalist framework. Thus, it appears that we have a hybrid theory of justification operating in the background here, one that is akin to Susan Haack's (1993) *foundherentism*. I leave the rest to the reader.

4.6 Bibliography

- Boolos, G. (1987). "Saving Frege From Contradiction." *Proceedings of the Aristotelian Society*, (87): 137-151.
- Cook, R. T. (2012). "Conservativeness, Stability, and Abstraction." *The British Journal of Philosophy*, vol.63(3): 673-696.
- (2017). "Abstraction and Four Kinds of Invariance (Or: What's So Logical About Counting)." *Philosophia Mathematica*, 25(1): 3-25.
- (2023). "Abstraction and Modest Reflection." In *Themes from Weir: A Celebration of the Philosophy of Alan Weir*, edited by A. Reiger and S. Leunberger: 133-170. Synthese Library, vol.484.
- Frege, G. (1980). *Philosophical and Mathematical Correspondence.*, edited by G. Gabriel *et al.*, and translated by H. Kaal: 38-42. Chicago: University of Chicago Press.
- Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology.* Oxford: Blackwell.
- Hale, B., and C. Wright. (2001). "To Bury Caesar. . ." In *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*, edited by B. Hale and C. Wright: 335-396. Clarendon, Oxford: Oxford University Press.
- Heck, R. K. (1997). "The Julius Caesar Objection." *Language, Thought, and Logic: Essays in Honour of Michael Dummett*, edited by R. K. Heck: 273-308. Oxford: Oxford University Press.
- Linnebo, Ø. (2018). *Thin Objects: An Abstractionist Account.* Oxford: Oxford University Press.

- Manely, D. (2009). "Introduction." In *Metametaphysics: New Essays on the Foundations of Ontology*, edited by D.J. Chalmers, D. Manely, and R. Wasserman: 1-37. Oxford: Oxford University Press.
- Putnam, H. (1981). *Reason, Truth, and History*, vol. 3. Cambridge: Cambridge University Press.
- Rayo, A. (2013). *The Construction of Logical Space*. Oxford: Oxford University Press.
- Shapiro, S. (1997). *Philosophy of Mathematics: Structure and Ontology*. New York: Oxford University Press.
- (2011). "EPISTEMOLOGY OF MATHEMATICS: WHAT ARE THE QUESTIONS? WHAT COUNT AS ANSWERS?" *The Philosophical Quarterly*, Jan. 2011, vol.61(242): 130-150.
- Weir, A. (2003). "Neo-Fregeanism: An Embarrassment of Riches" *Notre Dame Journal of Formal Logic*, Nov. 2003, vol.44: 13-48.
- Wright, C. (1999). "Is Hume's Principle Analytic?" *Notre Dame Journal of Formal Logic*, (40): 6-30.