Rescuing Implicit Definition from Abstractionism

Daniel Waxman
Lingnan University

Abstract

Neo-Fregeans in the philosophy of mathematics hold that the key to a correct understanding of mathematics is the *implicit definition* of mathematical terms. In this paper, I discuss and advocate the rejection of abstractionism, the putative constraint (latent within the recent neo-Fregean tradition) according to which all acceptable implicit definitions take the form of abstraction principles. I argue that there is reason to think that neo-Fregean aims would be better served by construing the *axioms of mathematical theories* themselves as implicit definitions, and consider and respond to several lines of objection to this thought.

1 Introduction

Neo-Fregeans in the philosophy of mathematics hold that the key to a correct understanding of mathematics is the *implicit definition* of mathematical terms.[1] Such implicit definitions, they believe, play two roles. First, they play the *semantic* role of introducing terms which—according to a battery of background semantic and metaphysical views—successfully refer to mathematical objects. Second, they play the *epistemic* role of allowing for *a priori* justification or knowledge of at least basic propositions concerning the objects whose reference has been secured, and thereby (via plausibly *a priori* logical resources) allow for justification or knowledge of a substantial range of non-basic propositions too. If neo-Fregeans are right, implicit definition holds out the prospect of providing a fully satisfying solution to the famous challenge raised by Benacerraf (1973), who saw the philosophy of mathematics pulled in mutually incompatible directions by the desire, on the one hand, to give mathematical sentences a face value reading and the need, on the other, to explain how a tractable epistemology of mathematical belief is possible.

[1] Many have contributed to the neo-Fregean programme in one way or another. See Wright (1983) for the *locus classicus*, Hale and Wright (2000) for their conception of implicit definition and the role it plays, and the essays in Hale and Wright (2001).

Consider arithmetic, around which much of the discussion has revolved. According to neo-Fregeans, arithmetic is grounded (in a sense to be explained) in the stipulation of what has come to be known as Hume's Principle:

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

where $\approx$ abbreviates the claim that there is a bijection between F and G (which can be defined in second-order logic).

The idea is that HP is to be viewed as an implicit definition of the cardinality operator "the number of...", denoted by #. Much neo-Fregean effort has been expended on developing a general account of implicit definition intended to vindicate the claim that, in the best cases—of which HP is supposed to be a representative example—implicit definitions can perform a great deal of valuable philosophical work. Here are some of the main benefits that have been claimed for HP, with directly analogous virtues carrying over to the other 'good cases' of implicit definition:

1. HP, when understood as an implicit definition, bestows a clear sense on the (previously undefined) "number of" operator "#" . This definition serves to thereby introduce a range of complex singular terms of the form "#Φ", read as "the number of Φs".

2. Granted the success of HP as an implicit definition (and the truth of relevant instances of its right-hand-side), singular terms of the newly introduced form "#Φ" are guaranteed to refer. There is no *further* possibility of reference failure, that somehow there is no object to which "#Φ" refers, above and beyond the possibility of failure of the implicit definition itself.

3. HP is supposed to introduce a concept of a distinctive sortal kind—*cardinal number*—under which the referents of singular terms of the newly introduced form fall, and it is supposed to *explain* this newly introduced sortal kind. The explanation of the sortal kind succeeds, in particular, in introducing a new category of objects in such a way that explains why cross-categorical identifications (e.g. the claim that Julius Caesar is identical with the number 2) are unproblematically false.

4. Finally, HP is supposed to realize certain considerable epistemic benefits. Most important, it promises to sustain what Hale and Wright call the "traditional connection" between implicit definitions and *a priori* knowledge: as they put it, "to know both that a meaning is indeed determined by an implicit definition, and what meaning it is, ought to suffice for *a priori* knowledge of the proposition

thereby expressed."[2] If this is right, HP is available as an item of *a priori* knowledge: it can be known, with no or at least minimal collateral epistemic work, simply via competent stipulation. It is well-known that HP, against a background of second-order logic, interprets second-order Peano Arithmetic. So perhaps second-order arithmetic in its entirety can come to be known *a priori* as well.[3]

Clearly, then, the benefits of implicit definitions—if the philosophical work in defending theses (1)-(4) can be pulled off—are vast. But although these benefits attach, allegedly, to implicit definitions in general, neo-Fregean attention has concentrated—I think it's fair to say exclusively—on implicit definitions of the following form:

$$\forall\alpha\forall\beta\,(\S\alpha = \S\beta \;\leftrightarrow\; \alpha \sim \beta)$$

where § is a term-forming operator, α and β are expressions of a certain type, and $\approx$ denotes an equivalence relation holding between entities of the relevant type. Call principles of that form *abstraction principles*. This paper will discuss, and ultimately advocate the rejection of, the view I will henceforth call *abstractionism*: that, in the good cases, abstraction principles enjoy certain metaphysical, semantic, and epistemic benefits not shared by putative implicit definitions of other forms. Building on a suggestion by John MacFarlane, my aim will be to motivate and explore the—in my view, considerable—attractions of a position that combines the neo-Fregean friendliness towards implicit definition with a broader non-abstractionism.[4] In particular, I will advocate a so-called *Hilbertian Strategy*, according to which *the axioms of mathematical theories* themselves constitute implicit definitions of the distinctive vocabulary used in their statement.

I will defend two main theses. The first is a *Parity Thesis*: if the benefits claimed for implicit definition by the neo-Fregeans are genuinely available for the taking, those benefits are equally available to proponents of the Hilbertian Strategy. My second thesis is that the Hilbertian Strategy in fact has significant advantages over the traditional approach. Unlike abstractionism, it generalizes seamlessly to the whole of mathematics. Furthermore, it opens up a plausible response to the Bad Company objection that has notoriously plagued the neo-Fregean program.

In a sense, the project can be viewed as an attempt to (perhaps subversively) appropriate the resources of implicit definition that have been defended so ardently and so arduously by the neo-Fregeans, and place them in the service of what might be called a neo-*Hilbertian* view.[5] Perhaps I should say "*re*-appropriate", for it is an irony of his-

[2] Hale and Wright (2000, 296).

[3] The transition from "a theory within which the axioms of PA can be interpreted" to "arithmetic" should not go unnoticed, as Heck (2000) has emphasized, but neo-Fregeans plausibly hold that this transition is warranted.

[4] MacFarlane (2009). See §2 for more on MacFarlane's contribution.

[5] This term is aptly used by Hale and Wright (2009a) as a description of a view they oppose.

tory that the self-described proponents of a distinctively neo-Fregean approach to the philosophy of mathematics are the chief contemporary defenders of implicit definition as a means of gaining mathematical knowledge, given the antipathy that Frege himself manifested towards the notion both in his correspondence with Hilbert and in his post-*Grundlagen* writings.[6]

The plan is as follows. In §2, I briefly recap the constraints that Hale and Wright place on legitimate implicit definitions, and present (largely following MacFarlane (2009)) a *prima facie* case that the Hilbertian Strategy satisfies each of them. In §3, I outline what I take to be the major advantages of such an approach, both in comparison to the orthdox neo-Fregean project and in its own right. I then turn in §4 to defend the view against a series of objections, the most serious of which are raised by Hale and Wright. These objections come in three varieties: semantic, epistemic, and concerns regarding the applicability of mathematics. I conclude that neo-Hilbertianism should, at the very least, be considered a serious rival to abstractionism.

2 Implicit Definition and the Hilbertian Strategy

2.1 Implicit Definition

The conception of implicit definition we'll consider is one elaborated by Hale and Wright in their essay *Implicit Definition and the A Priori*, although it was arguably latent in many earlier neo-Fregean writings. Hale and Wright (2000, 286) state the central idea as follows:

> we may fix the meaning of an expression by imposing some form of constraint on the use of longer expressions—typically, whole sentences—containing it.

More specifically [288]:

> We take some sentence containing—in the simplest case—just one hitherto unexplained expression. We stipulate that this sentence is to count as true. The effect is somehow to bring it about that the unexplained expression acquires a meaning of such a kind that a true thought is indeed expressed by the sentence—a thought which we understand and moreover know to be true, without incurring any further epistemological responsibility, just in virtue of the stipulation.

So implicit definition works by taking a sentence or sentences containing novel vocabulary to be *stipulatively true*. In the best cases, the idea runs, this will bestow the novel

[6] For the correspondence, see Frege (1982).

vocabulary with a meaning that serves to vindicate the stipulation. To return to the example of arithmetic, HP contains the novel operator "#" which is supposed to receive its meaning from the stipulation that, for all F and G, $\#F = \#G$ whenever F and G can be put into one-one correspondence.

Although it is beyond the scope of this paper to defend the merits of implicit definition in detail, it is worth saying something about the general theoretical orientation from which it arises. The view may seem alien to those who conceive of expressions having a meaning by somehow "latching on" to meaning-entities independently existing in some Fregean third realm. In contrast, the doctrine is seen naturally as arising from the combination of two commitments: (i) a generally use-based metasemantic account of how *sentential* semantic properties (in particular, truth) are fixed; and (ii) the explanatory priority of syntactic over semantic sub-sentential properties ("syntactic priority").[7]

In extremely compressed detail: the putatively mysterious idea of our "stipulating the truth" of a certain sentence is best understood as, roughly, commitment to employing that sentence in certain patterns of usage. The thought is that, against the background of a use-based metasemantic account, this will be possible to do in a way that ensures that such sentences are true. (It is helpful to compare the way in which inferentialists about logical connectives explain the semantic content of logical vocabulary). Syntactic priority is a complex package of views: in particular, it endorses three crucial moves: (i) from an expression's exhibiting the typical syntactic/inferential behaviour of a singular term to its being a *genuine* singular term; (ii) from an expression's being a genuine singular term figuring in a true sentence to its being a genuinely *referring* singular term; and (iii) from the fact that a singular term refers, to there being an *object* (as opposed to something of a different ontological category, e.g. a Fregean concept) to which it refers.[8]

Naturally, we can only hope to scratch the surface of issues concerning metasemantics, implicit definition, the syntactic priority thesis, and the connection between implicit definition and *a priori* justification here. My goal is not to defend the neo-Fregean cluster of views, but rather to argue for the conditional: *if* something in the ballpark of Hale and Wright's account of implicit definition is correct—i.e. in the best cases, implicit definitions can both (i) ensure that previously unmeaningful vocabulary can receive a meaning by appropriately figuring in sentences that are stipulated to be true, and (ii) provide a means for us to *know* the definitive sentence—then we have no reason to believe that the class of successful implicit definitions is restricted to abstraction

[7]Much more on these views can be found in Hale and Wright (2001). See also Hale and Wright (2009b) on neo-Fregean metaontology and MacBride (2003) for a helpful overview of the semantic and metasemantic commitments of neo-Fregeanism.

[8]For an account of singular terms, see Hale's Chapters 1 and 2 of Hale and Wright (2001).

principles.

What does "in the best cases" mean? Hale and Wright (2000) give five criteria they take to be individually necessary (and, tentatively, jointly sufficient) for an implicit definition to succeed:

Consistency. The sentence serving as the vehicle of the definition must be *consistent*.

Conservativeness. The sentence serving as the vehicle of the definition must be *conservative* over the base theory. Roughly, the extended theory ought not introduce new commitments concerning the ontology of the base theory.

Less roughly, say that a theory T_2 is a *conservative* extension of T_1 if for any sentence ϕ in the language of T_1, if $\mathsf{T}_1 + \mathsf{T}_2 \vdash \phi$ then $\mathsf{T}_1 \vdash \phi$. Conservativeness is too strong of a condition to place on implicit definitions, at least if neo-Fregean proposals are to have a chance of getting off the ground. The reason is that Hume's Principle has consequences, e.g. that there are infinitely many objects, that may be expressible in the base language and yet not follow from the base theory. To get around this issue, neo-Fregeans have moved to what has become known as *Field-conservativeness*, which intuitively says that the extended theory has no new consequences *for the objects spoken of by the base theory*.[9] To express this formally, we need some notation. Let $P(x)$ be a predicate that doesn't occur in T_1 or T_2 (intended to pick out all and only the "old" objects spoken of by T_1). For each sentence ϕ in the language of T_1, let ϕ^* be the result of relativizing all quantifiers occurring in ϕ to $P(x)$ and let T_1^* be the theory whose axioms are the sentences ψ^*, where ψ is an axiom of T_1.[10]

This allows the relevant notion of conservativeness to be defined: T_2 is a *Field-conservative extension* of T_1 if, for any sentence ϕ in the language of T_1, if $\mathsf{T}_1^* + \mathsf{T}_2 \vdash \phi^*$, then $\mathsf{T}_1 \vdash \phi$.

The proposal, then, is that implicit definitions must be Field-conservative over the theory to which they are added.

Harmony. In Wright and Hale's discussion, Harmony is a constraint that applies primarily to implicit definitions of logical expressions (in particular, expressions with both an "introduction rule" and an "elimination rule"), intended to rule out unharmonious connectives like Pryor's "tonk" and its dual. It is unclear that this constraint has any bearing on the mathematical cases we are concerned with here, so I will pass over it

[9] See Field (2016, 11) for this notion of conservativeness, and Hale and Wright (2001, 297) for the thought that it is the relevant notion in stating this constraint on implicit definition.

[10] I assume that the axioms of theories are *sentences* rather than open formulae.

without further comment.[11]

Generality. This condition holds that the meaning that new terms receive from implicit definitions must satisfy Gareth Evans (1982)'s *Generality Constraint:* very roughly, if new sub-sentential expressions are introduced by a definition, one who grasps the definition must thereby be provided with the means of understanding the meaning of arbitrary sentences composed of (grammatically appropriate) combination of the new expression and antecedently understood vocabulary. This is obviously connected with the Caesar problem: if, as neo-Fregeans contend, HP construed as an implicit definition is able to provide singular terms such as $\#F$ with a meaning, then the Generality Constraint requires (as a special case) that it also provide us with a means of understanding mixed identity sentences like $\#F =$ Julius Caesar.

Anti-arrogance. By contrast with the previous—logical and semantic—conditions, Anti-arrogance is an epistemic constraint. Hale and Wright put it as follows. An implicit definition is arrogant if:

> the truth of the vehicle of the stipulation is hostage to the obtaining of conditions of which it's reasonable to demand an independent assurance, so that the stipulation cannot justifiably be made in a spirit of confidence, "for free"

Here is a proposal to make this more precise. Say that a sentence (understood as the vehicle of an implicit definition) S is arrogant if (i) there is some condition C that must be satisfied for the truth of S; (ii) we require justification that C is satisfied in order to have justification in S; and (iii) we have no justification that C is satisfied.[12]

There are two points worth noting. One is that condition (ii) is somewhat schematic: to flesh it out, we need to know which conditions require antecedent justification. This involves deep questions in epistemology, which cannot fully be discussed here. But there are different possible views here, ranging from the very conservative (we must have justification that *all* conditions necessary for the truth of the stipulation are satisfied) to the very liberal (we do not need justification that *any* are). Hale and Wright are

[11] See for instance Tennant (1978) and Dummett (1991b) for discussions of Harmony in the logical setting. More recently, Wright (2016) suggests that Hume's Principle is best understood as functioning more like a rule of inference than an axiom. A full discussion of Harmony in the context of non-logical rules, however, would take us too far afield.

[12] For the sake of clarity, I understand justification in the propositional sense: roughly, what it would be appropriate for someone in one's epistemic position to believe, whether or not one in fact has the belief in question. I mean to be as neutral as possible here, and in particular not to exclude varieties of entitlement, i.e. "default" or "non-earned" forms of justification. See Wright (2016), and Hale and Wright (2001, 127) for evidence that this is how they understand our justification in the preconditions for the stipulation of HP.

not very explicit about where on this spectrum they lie, but their discussion seems to situate them towards the more conservative end. I will follow them in this respect. Dialectically, this is appropriate: the more conservative the approach, the fewer legitimate implicit definitions there will be, so proceeding this way does not give the Hilbertian unfair advantage.[13] The second point to note is that, due to condition (iii), whether an implicit definition is arrogant in the above sense will be sensitive to one's epistemic position. In other words, it is possible for a stipulation to be arrogant in the hands of one epistemic agent, and not in the hands of another, depending on which conditions they have justification to accept. This will prove later to be of significance.[14]

At this stage, it is time to introduce the proposed alternative to the neo-Fregean strategy and to evaluate it against these criteria.

2.2 The Hilbertian Strategy

Return again to the case of arithmetic. Neo-Fregean abstractionists want to explain our understanding of arithmetic basically as follows: we begin by stipulating Hume's Principle as a definition of "the number of...", and then we use (second-order) logical resources to derive certain theorems, including, most importantly, the axioms of PA. This result is now known as Frege's theorem, and is surely one of Frege's most substantial mathematical achievements.[15] According to neo-Fregeans, Frege's Theorem provides a means of recovering the PA axioms and thereby the whole of arithmetic in an epistemically responsible way, since, they claim, (i) Hume's Principle is something that we can come to know *a priori*, and (ii) knowledge can be transmitted via competently carried-out second-order logical deductions.[16]

But why the need to go via HP? What, exactly, is mandatory about taking the neo-Fregean route to the Peano Axioms, as opposed to the following procedure? Lay down

[13] Ebert and Shapiro (2009), Section 6, discuss some of the options one might adopt here, and argue that neither (what I have called) the very conservative nor the very liberal approaches are very attractive. However, their discussion of the conservative approach seems to me to be marred by a conflation of the *knowability* or *justifiability* of consistency with its *provability*, which they rightly take to be ruled out for Godelian reasons. I discuss this, and the epistemology of consistency more generally, in other papers.

[14] Anti-arrogance ought to be distinguished from other conditions that might be confused with it. One is *conditionality*; it is not equivalent to saying that implicit definitions have a conditional as their main connective. (To see that it is not sufficient, consider any arrogant stipulation A and consider $\top \rightarrow A$, where $\top$ is some logical truth. To see that it is not necessary, take Hume's Principle itself). Anti-arrogance is also importantly distinct from Conservativeness. Consider Goldbach's conjecture which (let us suppose) is a truth of arithmetic for which we lack a proof and, consequently, justification. Now consider the stipulation of the theory: PA + Goldbach's conjecture. This is, *ex hypothesi*, a conservative extension, for Goldbach's conjecture follows from PA. But it is nevertheless arrogant, for we plausibly need *independent* assurance that Goldbach's conjecture is true before we can have any justification that the theory is true, i.e. that there are entities that simultaneously satisfy PA and Goldbach's conjecture.

[15] For details of the modern rediscovery of Frege's Theorem and the contributions (in various parts) of Geach, Parsons, Tennant, Wright, Boolos, and Heck, see Burgess (2005, Ch 3).

[16] In doing so they presumably appeal to a plausible epistemic closure principle.

the (conjunction of) the Peano Axioms as a stipulation, intended to implicitly define the expression S—denoting the successor function—the numerical singular terms—0, 1, 2, etc—and a predicate $\mathbb{N}$ applying to all and only natural numbers. Instead of using HP to implicitly define the notion of natural number and to serve as the premise for a derivation of the PA axioms, the proposal is that *the axioms themselves* are understood as the definitive principles introducing that notion. This is what I will call the Hilbertian Strategy applied to arithmetic. More generally, it takes the axioms of some target theory as themselves serving as an implicit definition of the central notions involved.

It should be simple to see how the Hilbertian Strategy can, in principle, be extended to any other axiomatizable mathematical theory. To take just a few salient examples, the axiomatic theory of the real numbers (axiomatized e.g. as a complete ordered field), the complex numbers (axiomatized as e.g. the algebraic closure of the reals, or as a field of characteristic 0 with transcendence degree $\mathfrak{c}$ over $\mathbb{Q}$), set theory (in any of its usual axiomatizations, e.g. ZFC, NBG, etc). In general there is a version of the Hilbertian Strategy available for any axiomatic theory: one merely takes the axioms themselves as an implicit definition of the predicates, relations, and constants involved in the axiomatization.[17]

Orthodox neo-Fregeans will no doubt be filled with the conviction that whatever the apparent attractions of the Hilbertian Strategy, they are the result of theft over honest toil. But they must face the question, first raised by John MacFarlane: what strictures on implicit definition would this procedure violate, in the case of arithmetic in particular and for other mathematical theories more generally? Let us take the conditions in turn.[18][19] I will attempt to make a *prima facie* case that the two approaches are, at the very least, on a par. Later, in §4, I will consider and reply to various subtle arguments to the effect that, despite first appearances, neo-Fregeans are in a better position than Hilbertians to show that the relevant criteria are satisfied.

Consistency. If HP is consistent, then so is PA. This is because PA is relatively interpretably within HP. On minimal epistemic assumptions this implies that we have at least as much reason for believing that PA is consistent as we do for believing that HP is consistent. So if the consistency constraint is satisfied by the neo-Fregean stipulation,

[17] If the theory is finitely axiomatized, the definition can be given as a single conjunctive sentence. If it is schematically axiomatized, there are different options available: to use a truth predicate of some kind; to appeal to a device of infinite conjunction or some other kind of higher order resources; or to take the instances of the schema as collectively constituting the definition.

[18] Of course, this point is merely ad hominem against Hale and Wright in the absence of arguments that HP and PA *do* satisfy the requirements. I take it that the discussion to come addresses at least some worries about generality and arrogance; I discuss issues of consistency and conservativeness further in other work.

[19] The points in the remainder of this section largely follow MacFarlane (2009), who first pointed out that something like the Hilbertian Strategy appears to satisfy Hale and Wright's criteria.

it is satisfied by the Hilbertian stipulation.

Conservativeness. The justification here is similar to that of consistency. If PA has any non-conservative implications over the base theory to which it is added, then so does HP, in virtue of the relative interpretability of PA within HP.

Harmony. As mentioned, Harmony is irrelevant outside of the context of logical connectives.

Generality. *Prima facie*, a stipulation of PA appears to be no better or worse off in regard to its ability to satisfy the Generality Constraint than does a stipulation of HP. It is true that a stipulation of PA does not, at least *prima facie*, fix the meaning or truth-conditions of certain grammatically appropriate sentences involving the newly introduced terms, such as "$2 =$ Julius Caesar" or "$\mathbb{N}(\text{Caesar})$" or "$S(\text{Caesar}) = (\text{Augustus})$". This is just to say that the Hilbertian faces a version of the Julius Caesar problem. However, the Caesar problem is notoriously pressing for abstractionists too – indeed, it was precisely Frege's own reason for rejecting (what in this context can only anachronistically be called) the neo-Fregean strategy for grounding arithmetic in HP. So, again *prima facie*, the two approaches are on a par. Naturally, neo-Fregeans have said much in response to the Caesar problem. As I will argue in Section 4, many of the resources to which they appeal can be equally well appropriated by the Hilbertian.

Anti-arrogance. Again *prima facie*, a stipulation of PA appears to have equally—or less—demanding conditions for its truth than a stipulation of HP. Any model of HP can be expanded a model of PA. For instance, given classical logic, both theories entail the existence of infinitely many objects, so both exclude finite domains. Consequently, it is difficult to see how PA might be *more* arrogant than HP. Certainly, one might not be justified in, e.g., believing that the universe co-operates in providing enough or the right kind of entities to allow the stipulations to be true. But if so, this is a reason to convict *both* implicit definitions of arrogance, not just PA. In Section 4, I consider and respond to some further subtle arguments from neo-Fregeans to the effect that the Hilbertian strategy is arrogant in a way that their own approach is not.

The point of the foregoing is that there is a strong *prima facie* case for the conditional claim: *if* the neo-Fregean Strategy with respect to arithmetic is successful, then so too is the acceptability of the Hilbertian Strategy. In fact, as we have seen, the Hilbertian Strategy appears if anything to be in a *better* position. Furthermore, since nothing particular

to arithmetic was appealed to in these arguments, the reasoning above can be replicated in the general case: wherever the theory generated by an acceptable abstraction principle allows for interpretation of an axiomatic theory that we find of mathematical interest—that is just to say, whenever the neo-Fregean Strategy for explaining some portion of established mathematics is successful—then so too is the analogous instance of the Hilbertian Strategy. As I will now argue, the Hilbertian Strategy is not merely on a par with the neo-Fregean strategy: in fact, it has certain definite advantages.

3 Advantages of the Hilbertian Strategy

3.1 Abstractionism and Set Theory

A major thorn in the side of abstractionist neo-Fregeans has been an inability to develop a satisfactory theory of sets. As is well known, Frege's original approach, employing Basic Law V:

(BL V) $\forall F \forall G \left(\{x : Fx\} = \{x : Gx\} \leftrightarrow (\forall x, (Fx \leftrightarrow Gx))\right)$

as the central abstraction principle governing the identity of sets, is inconsistent. The problem is simply this: no subsequent development of set theory has delivered a theory that can be plausibly considered foundational in the manner of ZFC, the most popular and widely used set theory. The best-known approach, due to George Boolos, appeals to so-called New V:

(New V) $\forall F \forall G \left(\{x : Fx\} = \{x : Gx\} \leftrightarrow ((Bad(F) \wedge (Bad(G)) \vee (\forall x \, (Fx \leftrightarrow Gx))))\right)$

where $Bad(F)$ expresses the condition that F is "large", i.e. can be placed into bijection with the entire universe. The idea is to avoid paradox by encoding a class/set distinction into the implicit defintion of the notion of set: set-sized concepts—those that are not equinumerous with the universe—form sets, while class-sized concepts do not. As Boolos showed, New V is consistent and able to obtain (with suitable definitions) many of the axioms of ZFC, including Extensionality, Empty-Set, Well-Foundedness, Pairing, Union, Separation, and Replacement. However, there is a clear sense in which this theory—call it Boolos Set Theory (BST)—lacks the ontological power of ZFC. To see this, note that BST is satisfied by the hereditarily finite sets: it follows that in it, the Axioms of Infinity and Powerset both fail, since neither hold in the hereditarily finite sets.[20] Given that the ontological power of ZFC is generated primarily by these two axioms, the resulting theory will be seen as admitting a severly impoverished ontology

[20]The hereditarily finite sets are the sets with finite transitive closures, i.e. elements of V_k for some finite k.

from the perspective of a believer in the universe of sets as standardly conceived. I do not know of any more promising abstractionist approaches to set theory.[21]

Of course, there are two reactions one might have to the seeming inability of abstractionism to recover a theory that is recognizably comparable to contemporary set theory in its scope and power. One can think, as Wright (2007, 174) tentatively advances, that if "it turns out that any epistemologically and technically well-founded abstractionist set theory falls way short of the ontological plenitude we have become accustomed to require, we should conclude that nothing in the nature of sets, as determined by their fundamental grounds of identity and distinctness, nor any uncontroversial features of other domains on which sets may be formed, underwrites a belief in the reality of that rich plenitude." But this is a radical conclusion indeed. Whatever one thinks of set theory, it is impossible to deny its importance as a foundational theory within contemporary mathematics, in effect providing the predominant ontological background and organizational framework within which mathematics is carried out. The fact that it cannot easily be seen to obviously follow from any abstraction principle governing the identity of sets is arguably more of a reason for rejecting that demanding constraint than it is for rejecting set theory itself.[22]

By contrast, compare how easily the proponent of the Hilbertian Strategy is able to respond to the challenge posed by contemporary set theory. It is unsurprisingly straightforward: merely take one's favourite set theory (say, ZFC, perhaps with some large cardinal axioms if one is feeling adventurous) and understand its axioms as implicitly defining the notions of set and membership. To be sure, for this to be a successful implicit definition of *set* and *membership*, ZFC must (like all putative implicit

[21] As Burgess (2004) has shown, by adding a reflection principle (and plural logical resources) to BST, all of the ZFC axioms are (remarkably) once again obtainable. It nevertheless ought to be clear that a theory formulated in this way will fail to satisfy abstractionist strictures, since to the best of my knowledge there is no way to straightforwardly write it as an abstraction principle. There is additionally a further problem that New V entails a *global* choice principle, and this may violate the conservativity requirement for implicit definitions (even the relevant and relatively weaker notion of Field-conservativeness as introduced above). See Shapiro and Weir (1999) for details. Fine (2002) has also worked extensively on the limits of abstraction principles; the theories he obtains, the n^{th} order "general theory of abstraction" (for finite n), are in general, equi-interpretable with $n + 1^{st}$ order arithmetic. This is certainly enough to carry out much mathematics, but it is nevertheless a far cry from an adequate replacement for orthodox set theory. See Burgess (2005) for more on the limitations of Finean arithmetical theories.

[22] In conversation, Crispin Wright has raised the interesting idea that an abstractionist treatment of set theory might have the benefit of providing a principled distinction between features of set theory that are (my word, not his) "intrinsic", flowing from the nature of sets—i.e. those that follow from the abstraction principle (whatever it may be) governing sets—and those that are "extrinsic". (For instance, if we go with Boolos Set Theory, Powerset and Infinity will count as extrinsic axioms in this sense.) While this is intriguing, I think (if it is to recapture a theory with anything like the strength of ZFC), it will end up raising more epistemological difficulties than it solves. In particular, the question of the justification of any extrinsic axioms will become urgent in the face of something like Benacerraf's original dilemma: and, of course, on an abstractionist view, the resources of implicit defintion will be unavailable to play any epistemic role here.

definitions) satisfy the previously mentioned conditions in order to succeed: ZFC must be consistent, conservative, not require any objectionably arrogant epistemic preconditions to be satisfied, etc. I do not wish to trivialize the question of whether these conditions hold. In my view, the question of whether our best theories, ZFC included, are consistent (and in particular how we manage to obtain justification to believe they are) is one of the most pressing and neglected in the philosophy of mathematics. But this is not the place to discuss the issue further; for our purposes it's enough to note that, naturally, any neo-Fregean abstraction principle purporting to recover set theory would necessarily face the same challenges.

In short: the neo-Fregean's prospects for recovering a theory capable of playing the foundational role of set theory is questionable, and there is reason to be optimistic about the Hilbertian's prospects for doing the same.

3.2 How to Rid Oneself of Bad Company

Another major issue for neo-Fregeans is what has come to be known as the Bad Company Objection. There are a great number of possible abstraction principles that one might be tempted to adopt; but—and here is the rub—some of these principles lead to unacceptable results. The most obvious kind of unacceptability is exemplified by Basic Law V, which is inconsistent. But there are other, less obvious kinds of unacceptability also. For instance, there are pairs of abstraction principles that are individually consistent (and conservative) which nevertheless, when taken together, result in inconsistency.[23] So the problem is essentially this: given that not all consistent and conservative abstraction principles are acceptable, some account needs to be provided of which principles *are* acceptable. The problem generalizes in the obvious way to implicit definitions in general (of which abstraction principles are particular examples). How might this problem be resolved?

Neo-Fregeans face a difficult technical challenge here: to investigate the logical and mathematical features of abstraction principles and hope that some criteria can be found to distinguish, in a principled way, between good and bad cases.

I don't pretend to be in a position to offer anything like this kind of solution. Nevertheless, I do want to argue that there's a good sense in which the Hilbertian Strategy allows us to sidestep these worries by showing how implicit definitions can be combined, as long as the theories in question are consistent, to yield in effect the whole of contemporary mathematics. Here's a rough idea of what I have in mind. The Appendix contains a proof of the following result:

[23]For more on Bad Company, see e.g. Wright (1999), Shapiro and Weir (1999), a special issue of *Synthese* edited by Linnebo (2009), and several papers in Cook (2016).

Let T_m (for "mathematical") and T_b (for "background") be theories whose languages share no individual constants. As before, let T_m^* be the theory that results from relativizing the quantifiers of T_m to a fresh predicate, and let T_b^* be the theory that results from relativizing the quantifiers of T_b to a fresh (and different) predicate.

Theorem 1. *If T_m and T_b are consistent first-order theories, then $\mathsf{T}_b^* + \mathsf{T}_m^*$ is consistent and T_m^* is Field-conservative over T_b.*

The restriction to first-order theories is important: unfortunately, the result does not hold for second-order theories.[24] However, a slightly weaker result can be shown for second-order theories. First a definition:

Definition. *Field*-conservativeness.*

T_2 is a Field-conservative extension of T_1 if, for any sentence ϕ in the language of T_1, if $\mathsf{T}_1^* + \mathsf{T}_2 \vDash \phi^*$, then $\mathsf{T}_1 \vDash \phi$.*

where $\vDash$ is understood as *full* semantic consequence.[25]
Then we have:

Theorem 2. *If T_m and T_b are satisfiable second-order theories, then $\mathsf{T}_b^* + \mathsf{T}_m^*$ is satisfiable and T_m^* is Field*-conservative over T_b.*

Take Theorem 1 first. I'd like to gloss this result as saying that, if you start with a consistent base theory, and add to that theory an arbitrarily chosen consistent mathematical theory then—after the theories are cleaned up in what I take to be a wholly philosophically defensible way—the resulting theory is going to be (i) consistent and (ii) Field-conservative over the base theory. The philosophical upshot is, I claim, that the Hilbertian Strategy can provide an operational solution to the Bad Company objection: the result shows that if we add a new mathematical theory obtained via the Hilbertian Strategy, then (as long as the theory we attempt to add is consistent, which as we have seen is a condition on its success as an implicit definition in the first place), the end result will be itself consistent and conservative over the base theory to which it is added. What is more, this is a strategy that can be extended indefinitely: as long as the new theory we add at each stage is consistent, then there is *never* any risk of ending up in inconsistency or with objectionably non-conservative consequences.[26] There is no

[24] Proof sketch: let T_b be $\mathsf{PA}_2 + Con_{\mathsf{PA}_2}$ and let T_m be $\mathsf{PA}_2 + \neg Con_{\mathsf{PA}_2}$, formulated in a disjoint language (i.e. with different arithmetical constants and predicate-symbols). Then each theory is individually consistent; but $\mathsf{T}_b^* + \mathsf{T}_m^*$ is inconsistent, since it violates Internal Categoricity in the sense of Button and Walsh (2018, Chapter 10).

[25] See Shapiro (1991) for a definition.

[26] This is true, at least, when we "only" extend our theories finitely many times; which, I submit, is more than enough in practice.

prospect (as there is with abstraction principles in general) of falling into inconsistency by way of adding individually consistent principles/theories.

Three brief clarifications are in order. First, a word on the "cleaning up" of the theories. All I really mean by this is that before the theories are added together, the quantifiers of each are relativized to a predicate that is intended to pick out all and only the entities spoken of by the theory. Suppose we have a base theory that does not make any claims whatsoever about sets—a physical theory, say—and suppose we want to implicitly define a notion of set via the Hilbertian Strategy. The idea is that we introduce new predicates "set" and "non-set" and relativize all of the quantifiers of the relevant theories to those predicates. That way, we will not be saddled with set-theoretic claims like

$$\forall x \forall y \, (x = y \leftrightarrow \forall z \, (z \in x \leftrightarrow z \in y))$$

that have the (absurd) consequence of identifying all, e.g., physical objects that are non-sets. For in its relativized form

$$\forall x \forall y \, (Set(x) \wedge Set(y) \rightarrow (x = y \leftrightarrow \forall z \, (Sz \rightarrow (z \in x \leftrightarrow z \in y))))$$

the principle will be explicitly restricted only to sets. This move seems philosophically well-motivated for reasons independent of anything to do with the present project: theories, presumably, should always be written this way when we are being fully explicit, and this is especially true when our intention is to implicitly define new predicates that are intended to range over objects to which our old language did not refer.[27][28]

Second, it's worth noting that matters are slightly more complex for second-order theories: the mere consistency of the theories in question does not suffice, and the result as stated uses a kind of generalization of Field-conservativeness that involves second-order semantic consequence. While this is certainly a qualification worth mentioning, it nevertheless seems that the results taken together give us all that we need. For first-order theories—including the foundational case of most interest (set theory as codified in ZFC)—consistency suffices. If the theories in question are second-order, then the

[27] For more see Field (2016, 12). NB: to say that mathematical theories should be *relativized* in this way (i.e. to make explicit which type of object they concern) is not to say that they should be written as *conditional* on the existence of objects of the relevant type. Thus what I say here does not take a stand on the dispute between Field (1984)—who argues that HP is only acceptable conditional on the claim that numbers exist—and Wright (in Chapter 6 of Hale and Wright (2001))—who replies that the concept of number, which of course on his view is given by HP itself, is required in order even to understand the antecedent of such a conditional. Thanks to a referee here.

[28] This relativization also allows the Hilbertian to sidestep worries about apparently incompatible theories: for instance, ZFC + CH vs ZFC + ¬CH: they will be relativized to different predicates (say, Set_1 and Set_2), avoiding any actual incompatibility.

stronger condition that they must be (individually) satisfiable—where satisfiability is, in effect, the semantic analogue of consistency—is needed. This is very much in the spirit of the consistency requirement: indeed, for first-order theories, consistency and satisfiability in the relevant sense are coextensive.

Third—and this is one reason why I do not claim to have a fully diagnostic solution to Bad Company worries in general—I concede that the result mentioned does not help much in illuminating the question of Bad Company for abstractionists, since abstraction principles cannot easily conform to the relativization procedure discussed above. That said, this last fact may be seen by some as an additional advantage of the Hilbertian Strategy over the neo-Fregean abstractionist alternative.

4 Objections and Replies

In this section I discuss and reply to a number of objections—in keeping with the spirit of the paper, primarily objections arising from a neo-Fregean perspective. These can be classified in three broad groups. On the conception considered here, implicit definition has both *semantic* and *epistemic* components. It is supposed to fix a meaning for the terms introduced, as well as to illuminate and explain our justification with respect to basic truths involving them. There are correspondingly two ways in which an implicit definition might be thought to be defective. First, it could be semantically defective and fail to establish a legitimate meaning for the terms it is intended to introduce. Second, even if it succeeds semantically, there may still be reasons why it fails to provide justification or knowledge. In addition, there is a third potential source of difficulty: whether Hilbertian theories are capable of *applications* of the sort we require from mathematics. In particular, there is a distinguished line of thought, arguably originating with Frege himself, according to which the applicability of mathematics must be placed at the center of matters; some have suggested that this is a reason to prefer the neo-Fregean Strategy. Each of these groups of complaint will be addressed in turn.

4.1 Objections to the Semantic Role of Hilbertian Definitions

4.1.1 Generality, stipulation, and the Caesar Problem

Can instances of the Hilbertian Strategy really bestow a meaning on the novel vocabulary they contain? Hale and Wright think that there are serious doubts to be had. They ask us to consider the Ramsification of the conjunction of the axioms of PA: the sentence obtained by replacing the distinctively arithmetical vocabulary—S, $\mathbb{N}$, and 0—with variables, and existentially quantifying through these variables. Then, they consider the effect of stipulating this Ramsey sentence, and make three claims. First, it

cannot be *meaning-conferring* in any plausible sense, for it contains no new vocabulary that could stand to receive a meaning. Secondly, although such a stipulation may (in a perverse way) introduce the concept ω-sequence, the stipulation of its truth is irrelevant in this respect: that concept could equally well have been introduced by saying that an ω-sequence is *whatever* satisfies the Ramsey sentence, without any claim that there are such things. Finally, they consider what the difference between a stipulation of the Ramsey sentence of PA, and PA itself would be. As they put it, a stipulation of the Ramsey sentence appears to amount to the command:

> Let there be an omega sequence!

But the stipulation of PA itself appears to amount to:

> Let there be an omega sequence whose first term is *zero*, whose every term has a unique *successor*, and all of whose terms are *natural numbers*![29]

so the complaint is:

> it is not clear whether there really is any extra content—whether anything genuinely additional is conveyed by the uses within the second injunction of the terms "zero", "successor" and "natural number". After all, in grasping the notion of an omega-sequence in the first place, a recipient will have grasped that there will be a unique first member, and a relation of succession. He learns nothing substantial by being told that, in the series whose existence has been stipulated, the first member is called "zero" and the relation of succession is called "successor"—since he does not, to all intents and purposes, know which are the objects for whose existence the stipulation is responsible. For the same reason, he learns nothing by being told that these objects are collectively the "natural numbers", since he does not know what natural numbers are. Or if he does, it's no thanks to our stipulation.[30]

The objection can be reconstructed as follows:

(1) A stipulation of the Ramsey sentence for PA is not capable of playing a meaning-conferring role;

(2) A stipulation of the Ramsey sentence for PA is, in all relevant (i.e. meaning-conferring respects) equivalent to a stipulation of PA itself;

(3) So, a stipulation of PA is not capable of playing a meaning-conferring role.

[29] Hale and Wright (2009a, 470).
[30] Hale and Wright (2009a, 471-2).

This complaint raises subtle questions. To see why, it is helpful to ask the obvious question: why, if the present complaint is successful, do Hale and Wright not feel that it has any force against their own neo-Fregean view? Although they do not explicitly address the point in their discussion, my suspicion is that the answer is intimately related to the Julius Caesar problem; seeing why will allow us to assess the objection more adequately.

One way of understanding the Julius Caesar problem for neo-Fregeanism is as an accusation that HP is vulnerable to precisely the difficulty that is currently being alleged of PA: in particular, that HP is defective as a meaning-conferring definition, because it simply does nothing to tell us what the natural numbers are supposed to be. Although HP tells us *something* about the natural numbers and the "number of..." operator, namely that two concepts have the same number iff they can be put into bijection, it does nothing to enable to us to rule out claims like $2 =$ Julius Caesar, because HP is consistent with the natural numbers being anything whatsoever—even the familiar conqueror of Gaul himself. More generally, the complaint runs, HP puts us in no position to accept or reject mixed identity sentences of the form $\alpha = \beta$ where α is a canonical name for a number and where β is not; and this signals a major defect in it, understood as an implicit definition.

It is hard not read Hale and Wright as adverting to this issue when they complain that a stipulation of PA does not put us in a position to "know which are the objects for whose existence the stipulation is responsible". In other words, it seems they suspect that PA does nothing to single out the natural numbers: they could be anything, as far as we know from the definition, as long as there are enough of them and they have the relevant properties or stand in the relevant relations.

Hale and Wright, naturally, think that they have the resources to overcome the Caesar problem insofar as it threatens HP. In very compressed form, their solution is to appeal to the notion of a *pure sortal* predicate. Roughly, an entity falls under a pure sortal predicate if it is "a thing of a particular generic kind—a person, a tree, a river, a city or a number, for instance—such that it belongs to the essence of the object to be a thing of that kind."[31] For Hale and Wright, it is intimately part of the ideology of a pure sortal that it comes with an associated *criterion of identity*: a principle that canonically determines the truth of identity-statements that contain terms referring to entities of the relevant sort.[32] For instance, Hume's Principle can be understood as a criterion of identity for numbers (i.e. numbers are equal when they apply to equinumerous concepts); the Axiom of Extensionality can be understood as a criterion of identity for

[31] Hale and Wright (2001, 387).

[32] Hale and Wright are not explicit whether the notion of an identity criterion is best understood as epistemological or metaphysical.

sets (i.e. sets are equal when they have the same members); spatio-temporal continuity can be understood as a criterion of identity for physical objects across time; and so on. And here the possibility of an objection to the Hilbertian Strategy on behalf of the abstractionist opens up. For the instances of the Hilbertian Strategy that we have been considering do not—unlike instances of the abstractionist strategy—appear to come ready made with criteria of identity for the new sortal terms distinctively introduced.

This, I think, is among the strongest objections that the neo-Fregean is in a position to make. In short: abstraction principles function as criteria of identity; so, without abstraction principles, the Hilbertian does not have recourse to a criterion of identity for the objects characterized by the newly-defined vocabulary, and therefore lacks the resources to answer the Julius Caesar problem. What can be said in response?

My suggestion is to first briefly step back and examine the reasons why a need for a criterion of identity appears to arise in the first place. In their solution to the Caesar problem, Hale and Wright (2001, 389) present a picture in which:

> all objects belong to one or another of a smallish range of very general categories, each of these subdividing into its own respective more or less general pure sorts; and in which all objects have an essential nature given by the most specific pure sort to which they belong. Within a category, all distinctions between objects are accountable by reference to the criterion of identity distinctive of it, while across categories, objects are distinguished by just that—the fact that they belong to different categories.

The response I propose on behalf of the Hilbertian is to suggest that in order to sustain this picture—or at least, the part of it that involves *mathematical* categories—it is sufficient that we are provided with what we might call *theory-internal adjudications of identity*. The idea is that, for at least appropriately chosen mathematical theories, the theories themselves in some sense "tell us all we need to know" about the identity of the objects that the theories are about. Let me try and motivate this with some examples.

1. Set theory. It follows from the axioms of Zermelo Fraenkel set theory that

$$\forall x \forall y \left(Set(x) \wedge Set(y) \rightarrow (x = y \leftrightarrow \forall z \left(Sz \rightarrow (z \in x \leftrightarrow z \in y) \right)) \right)$$

   i.e. that two sets are identical if they have the same members.

2. Arithmetic. It follows from the Peano axioms that

$$\forall x \forall y \left(\mathbb{N}x \wedge \mathbb{N}y \rightarrow (x = y \leftrightarrow \forall z \left(Pxz \leftrightarrow Pyz \right)) \right)$$

i.e. that two natural numbers are identical if they have the same predecessors.[33]

3. Real numbers. It follows from the axioms of suitable treatments of the real numbers that

$$\forall x \forall y \left(\mathbb{R}x \wedge \mathbb{R}y \rightarrow (x = y \leftrightarrow \forall z \left(\mathbb{Q}z \rightarrow (z < x \leftrightarrow z < y)\right)\right)$$

i.e. that two real numbers are identical if they form the same "cut" in the rational numbers.

4. Complex numbers. It follows from the axioms of suitable treatments of the complex numbers that

$$\forall x \forall y \left(\mathbb{C}x \wedge \mathbb{C}y \rightarrow (x = y \leftrightarrow (\Re(x) = \Re(y) \wedge \Im(x) = \Im(y))\right)$$

where $\Re$ and $\Im$ are functions from $\mathbb{C} \rightarrow \mathbb{R}$ that respectively express the real and imaginary parts of a complex number.

More generally, introduce a new constraint on an acceptable theory intended to introduce some mathematical sort M—call it *Identity*—to the effect that the theory must entail a sentence of the form

$$\forall x \forall y \left(Mx \wedge My \rightarrow (x = y \leftrightarrow \Phi(x, y)\right)$$

where Φ is a formula expressing an equivalence relation on M-objects.

The idea, in brief, is that theory-internal adjudications of identity are capable of playing the role generally assigned to identity criteria. More specifically, take a theory that entails a claim of this kind. This allows us to understand an instance of the Hilbertian Strategy as taking the axioms of the theory as an implicit definition *of a sortal concept*: the concept of the relevant kind of mathematical object (perhaps sets, natural or real numbers, etc). The way in which this handles identity claims should be clear. Identity claims between objects of the relevant sort are to be handled straightforwardly in terms of the particular theory-internal adjudication; whereas the issue of cross-sortal identity claims is dealt with in much the same way as on the orthodox neo-Fregean account (Hale and Wright (2001, 389): "across categories, objects are distinguished by just that—the fact that they belong to different categories.")

More needs to be said about what, exactly, is required in order for a theory to provide an internal adjudication of identity in the relevant sense. Ideally, it would be desirable to enumerate further (necessary and sufficient) conditions. Although I cannot offer anything like a full theory here, a number of plausible conditions can be made

[33] An analogous condition could be written out, less perspicaciously, in terms of the successor function.

out, many of which can be adopted straightforwardly from the literature on criteria of identity.[34] For instance, Horsten (2010) mentions these:

1. Formal adequacy: a criterion of identity must express an equivalence relation.
2. Material adequacy: a criterion of identity must be true.
3. Necessity: a criterion of identity must follow from "theoretical principles concerning the subject matter in question".
4. Informativeness: a criterion of identity must be informative about the nature of the entities involved.
5. Non-circularity / Predicativity : a criterion of identity must not (essentially?) quantify over the entities whose identity is supposed to be established.

The last condition is somewhat tricky to formulate, if indeed it is a genuine constraint. However exactly it is done, it had presumably better not rule out the credentials of the Axiom of Extensionality, which is agreed on virtually all sides to be a paradigm case of an acceptable identity criterion.

At any rate, I do not propose that this list is exhaustive, and no doubt, more could be said. Nevertheless, I take it that it is extremely encouraging that all of the rough criteria set out above can be equally plausibly applied to theory-internal adjudications of identity. Whether or not one wishes to count them as criteria of identity proper, they each seem plausibly capable of performing the required job of introducing a genuine *sortal* concept, thereby allowing us a means of adjudicating identity claims between objects of the relevant sort and of explaining why identity claims between mathematical objects and objects of another sort—say, people—are unproblematically false.[35] If that is right, I can see no reason that they should not play roughly the role that identity-criteria play for the neo-Fregean. At the very least, we would need to hear a much more detailed case that the form of abstraction principles are uniquely suited to introducing sortal concepts than we have so far heard.

4.1.2 Does the Hilbertian Strategy attempt to stipulate truth?

It is worth briefly considering another complaint, related to the one discussed immediately above, where Hale and Wright appear to accuse the Hilbertian Strategist of

[34] I think that everything in the discussion above is consistent with the claim that what I have been calling theory-internal adjudications of identity simply are criteria of identity, though I would not like to make such a claim outright.

[35] The appeal to sortal concepts does not adjudicate the issue of identity claims between objects of putatively distinct *mathematical* sorts—e.g. the natural number 2 and the real number 2. This is a subtle issue, for neo-Fregeans as for others. See Fine (2002, I.5) for discussion. As far as I can tell, the proposal in Hale and Wright (2001, Ch 14) implies that identity claims of this kind are false. Thanks to a referee here.

attempting to *stipulate entities into existence*. Here is a representative passage:

> Before there is any question of anybody's knowing the vehicle to be true, the stipulation has first to make it true. Regrettably, we human beings are actually pretty limited in this department—in what we can make true simply by saying, and meaning: let it be so! No one can effectively make it true, just by stipulation, that there are exactly 200,473 zebras on the African continent. How is it easier to make it true, just by stipulation, that there is an ω-sequence of (abstract) objects of some so far otherwise unexplained kind? And even if we do somehow have such singular creationist powers, does anyone have even the slightest evidence for supposing that we do?... to lay down Dedekind-Peano as true is to stipulate, not truth-conditions, but truth itself.[36]

It seems to me that here, Hale and Wright are misled by their own rhetoric of "stipulation". What is going on, the Hilbertian theorist claims, is an instance of implicit definition, and it works no differently here than it does in the (putatively more favourable) case of Hume's Principle. A sentence containing previously uninterpreted vocabulary is being integrated into a pattern of usage in such a way as to give the sentence certain truth-conditions and the new vocabulary certain semantic values; no more, and no less. It is simply a mistake to suppose that anything objectionably "creationist" is under consideration.

Can anything more be said to support the accusation that Hilbertian implicit definitions involve the stipulation of *truth*, whereas abstraction principles merely assign *truth-conditions* to their left-hand-sides? On the neo-Fregean view, the success of HP (along with suitable definitions) ensures that, e.g. "0=0" has the same truth-conditions as the claim that the concept $x \neq x$ is in bijection with itself. Thus, on a coarse-grained view of "truth-conditions", HP ensures that this sentence has necessarily and always satisfied truth-conditions. If that is legitimate, however, there is no reason why the Hilbertian cannot claim the same status for the axioms of PA. Can the objection be pressed further if "truth-conditions" are understood in a more fine-grained way? The Hilbertian could, if necessary, take as the relevant implicit definition not the axioms of PA themselves but the biconditional consisting of the conjunction of the axioms on one side and some logical truth on the other. It might be objected in turn that this is inadequate as an implicit definition, since the choice of logical truth (and thus the truth-conditions assigned to the conjunction of the PA axioms) is arbitrary. Perhaps a natural candidate modification for the right-hand-side of the relevant implicit definition (inspired by the historical Hilbert himself) is the claim *that the axioms of* PA *are*

[36] Hale and Wright (2009a, 473).

logically consistent.[37]

4.2 Objections to the Epistemic Role of Hilbertian Definitions

4.2.1 Is HP objectionably arrogant?

Let me now turn to a concern, raised by Hale and Wright, to the effect that the Hilbertian Strategy is arrogant. It proceeds by way of the subtle observation that the equivalence between the (theories resulting from the) Hilbertian Strategy and the neo-Fregean Strategy is contingent upon the choice of background logic. More precisely, the situation is this. If the underlying logic is classical second-order logic, then HP and PA will have precisely the same models.[38] However, consider what happens when we work within a weaker *Aristotelian* logic in which the second-order constants denote and second-order quantifiers range over only *instantiated* concepts. In such a setting no empty concept will be countenanced. This impedes Frege's theorem from going through at a very early stage: for in order to define the number zero as the cardinality of some empty concept, one must be available. So, as Wright and Hale note, taken against a background of Aristotelian logic, HP has models of both finite and infinite cardinality, while PA has—just as in the classical case—only infinite models. As they put it:

> The least one has to conclude from this disanalogy is that, as a stipulation, Hume is considerably more modest than Dedekind-Peano: the attempted stipulation of the truth of Dedekind-Peano is effectively a stipulation of countable infinity; whereas whether or not Hume carries that consequence is a function of the character of the logic in which it is embedded—and more specifically, a function of aspects of the logic which, one might suppose, are not themselves a matter of stipulation at all but depend on the correct metaphysics of properties and concepts.[39]

While the logical difference that Hale and Wright advert to is undeniable, the question is whether it can legitimately be used to support the claim that the Hilbertian Strategy is *objectionably arrogant*. This would require the demonstration of two subclaims: (i) that due to this logical difference, a stipulation of HP really is "more modest" than a stipulation of PA, and (ii) that this difference genuinely marks a difference in the acceptability of the two kinds of definition for the epistemic purposes to which they are intended to be put. It is worth emphasizing that (i) is not enough—it is clearly consistent to hold that a stipulation of HP is more modest than a stipulation of PA while

[37] Thanks to a referee for pressing for clarity here.
[38] As usual, modulo appropriate definitions.
[39] Hale and Wright (2009a, 475).

maintaining that they both, so to speak, end up on the right side of the knowledge-conferring line.

To respond, I argue that abstractionists who seek to recover arithmetic cannot feasibly make this objection.[40] For as mentioned earlier, neo-Fregeans want to use HP plus second-order logic to derive the PA axioms and thereby put the whole of classical arithmetic within reach. And for just the reasons above, *orthodox* second-order logic is required here, for in an Aristotelian setting, that derivation will not go through. So, presumably, Hale and Wright must feel themselves entitled to appeal to orthodox second-order logic and to thereby somehow discount what would otherwise be the epistemic possibility that Aristotelian logic is ultimately the correct logic. If they are correct, then the logical difference between HP and PA to which they advert is simply irrelevant, because both neo-Fregeans and Hilbertians are justified in discounting that Aristotelian logic is appropriate for reasoning in the relevant setting. To put it another way: if it is a genuine epistemic possibility that the right logic is Aristotelian, then there may well be grounds for drawing a line between HP and PA; however, this epistemic possibility would itself undermine the prospect of grounding arithmetic in the former.

Let us consider another way of pressing the arrogance objection. Recall our sharpening of the notion above: an implicit definition S is arrogant if (i) there is some condition C that must be satisfied for the truth of S; (ii) we require justification that C is satisfied in order to have justification in S; and (iii) we have no justification that C is satisfied. Is there any case to be made that PA is arrogant but HP is not, with the claim that there exist infinitely many objects as the relevant condition C?

It is certainly true (assuming the legitimacy of orthodox second-order logic) that the existence of infinitely many objects is a condition for the truth of both HP and PA, so (i) holds for both approaches. Plausibly, too, proponents of both approaches are in a similar epistemic position (prior to the laying down of any implicit definition) regarding this condition, so that the situation with respect to its justification is also symmetric. The crux of the issue is whether there any scope to argue for a difference in (ii), in particular, that justification in the existence of an infinity of objects is somehow required *in advance* of justifiably accepting PA, but not in advance of justifiably accepting HP. The only way I can see that this might be done is by arguing that the existence of infinitely many objects is a *immediate* consequence of PA, in some epistemically relevant sense of "immediate", whereas the same does not go for HP. But I cannot see any plausibility in the general principle that in order to have justification in P, one must have antecedent justification in its immediate consequences, on any precisification of the notion of immediacy. Furthermore, on this view, the arrogance objection would be easy to

[40] In keeping with the general methodology of the paper, I assume that something like the neo-Fregean account is tenable, putting aside more fundamental objections.

defeat by slightly modifying the statement of the Hilbertian Strategy. Take for instance the version of the view discussed in the last subsection, whereby the relevant implicit definition is not the axioms themselves but rather the biconditional consisting of the axioms on one side and the claim that the axioms are consistent on the other. It is very hard to see any sense in which this principle leads immediately to the consequence that there exist infinitely objects while Hume's Principle does not.

4.3 Objections Concerning the Applicability of Mathematics

A salient feature of the abstractionist strategy concerning at least arithmetic is that the applications of the theory come off-the-shelf. Hume's Principle, in effect, builds into the very identity conditions of numbers their role as, so to speak, measures of cardinality. By contrast, PA appears simply silent on the question of applicability; it says nothing, on its face, about how arithmetic can be applied. In particular, although PA delivers a body of truths concerning the natural numbers, it says nothing about how we are to decide what the number *of*, e.g., a concept might be.

With that said, it is not difficult to obtain a perfectly satisfactory "number of" operator, given the axioms of PA and the resources of implicit definition. An operator "#" can be defined with the truth-condition that

$$\#\Phi(x) = n \leftrightarrow \Phi(x) \approx x < n$$

where the less-than relation is defined in the language of PA as usual. Using standard second-order resources, this allows the derivation of Hume's Principle from PA. Consequently there should be no *technical* worries that the Hilbertian Strategy for arithmetic is somehow less able to provide for its applicability.

Nevertheless, a more fundamental objection is lurking. It might be argued that this opposition—placing the applicability-conditions of numbers at the center of the implicit definition by which they are introduced versus introducing them in some other way altogether—is *philosophically*, if not technically, significant. Many philosophers, Frege himself included, have placed the applicability of arithmetic at the heart of the subject. As Frege famously said, "...it is application alone that elevates arithmetic beyond a game to the rank of a science. So applicability necessarily belongs to it."[41]

For some clarity on the question, consider two salient families of positions that one might take concerning the applicability of mathematical theories. First, there are, following Pincock (2011, 282), what we might call *one-stage* views. On such a view, the applicability of mathematical theories is not merely a peripheral feature of them, something that arises as a happy coincidence once the theory has been formulated and

[41] Frege et al. (2013, 100).

worked out. Rather, it is *part of the content* of the theory that it can be applied in the way that it is. As Wright puts the idea—he calls it Frege's Constraint—"a satisfactory foundation for a mathematical theory must somehow build its applications, actual and potential, into its core—into the content it ascribes to the statements of the theory".[42]

By contrast, consider *two-stage* views, according to which the application of mathematics can be understood as a two-step process. The first stage is an characterization of the subject-matter of (pure) mathematics that is autonomous of, i.e. makes no reference to, its applications. Examples of such views include straightforward platonism (according to which mathematics concerns a domain of mind-independent objects, picked out in a way that does not mention their applications), modal structuralism (Hellman (1989)), *ante rem* structuralism (Shapiro (1997)), and fictionalism (Field (2016)).

The second stage of a two-stage view then explains how, in light of the characterization of mathematics at the first stage, applications are possible. For instance, at a very high level of abstraction, this could be done by beginning with a purely mathematical domain and appealing to "isomorphism" or "representation" relations of structural similarity between the mathematical domain and the non-mathematical domain to which it is to be applied, thereby allowing the complex mathematical techniques and inference patterns that have been developed in pure mathematics to be brought to bear on problems concerning the structure of, e.g., the physical world.

The difference between the Hilbertian Strategy and the neo-Fregean Strategy for arithmetic can be seen as a microcosm of the opposition between one- and two-stage views. The neo-Fregean Strategy puts the possibility of counting objects at the core of the theory of arithmetic, in the strong sense that the very criteria of identity for numbers essentially involves their role as the measures of cardinality of concepts. By contrast, the Hilbertian Strategy is silent on the question of the application—counting—until it is augmented with the Dedekind-inspired definition of the 'number of' operator, which, in effect, introduces counting by setting up a "representation" relation between the objects falling under a certain concept and the natural numbers themselves.

In light of the foregoing distinction, the possibility of an argument against the Hilbertian Strategy and in favour of the neo-Fregean opens up—*if*, that is, Frege's Constraint holds. The same goes, *mutatis mutandis*, for other abstractionist treatments of mathematical theories: if a one-stage account of the applicability of the theory is plausibly required, and if the abstraction principle generating the theory can plausibly be said to place its applicability at the core of the account (in the way that HP does), then it seems the neo-Fregean Strategy will have demonstrable advantages over the Hilbertian

[42]Wright (2000, 324). For an attempt to provide a one-stage account of the real numbers, see Hale (2000). See also Batitsky (2002) for interesting arguments that Hale's view is inferior to the orthodox representation-theoretic explanation of the applicability of the reals (an explanation closer to the two-stage accounts I discuss below).

Strategy. The task for the neo-Fregean is establishing the plausibility of these required premises. So, let us ask: is there any good *argument* for preferring a one-stage account over a two-stage account, in the case of arithmetic in particular and in other areas of mathematics in general?[43] The question is a large one, but in the remainder of this section I will consider some arguments that one-stage views are required and attempt to rebut them on behalf of two-stage views.

4.3.1 Can only one-stage accounts explain the generality of applications?

In a discussion of Frege, Dummett argues for something like a one-stage account:

> But the applicability of mathematics sets us a problem that we need to solve: what makes its applications possible, and how are they to be justified? We might seek to solve this problem piecemeal, in connection with each particular application in turn. Such an attempt will miss the mark, because what explains the applicability of arithmetic is a common pattern underlying all its applications. Because of its generality, the solution of the problem is therefore the proper task of arithmetic itself: it is this task that the formalist [i.e. the two-stage theorist], who regards each application as achieved by devising a new interpretation of the uninterpreted formal system *and as extrinsic to the manipulation system*, repudiates as no part of the duty of arithmetic[...]
>
> It is what is in common to all such uses, and only that, which must be incorporated into the characterisation of the real numbers as mathematical objects: that is how statements about them can be allotted a sense which explains their applications, without violating the generality of arithmetic by allusion to any specific type of empirical application.[44]

I take the argument in the foregoing to be this: only one-stage accounts of the applicability of various parts of mathematics (arithmetic and real analysis in particular) are able to intelligibly explain the possibility of the application of those theories in the required generality that the phenomenon requires.

In response: I want to say that it is glaringly unclear what is supposed to be lacking in rival, two-stage ("formalist") accounts. The charge is that a two-stage account can give, at best, a piecemeal explanation—missing, as Dummett puts it, the common pattern. Settle for the sake of argument on a platonist two-stage view of mathematics. (A similar account will be available, *mutatis mutandis*, for other two-stage views.) The

[43] For sophisticated recent discussions of the issues surrounding neo-Fregeanism and applications, congenial to the view proposed here, see Snyder et al. (2020), Panza and Sereni (2019), and Sereni (2019).

[44] Dummett (1991a, 259).

explanation is, very roughly, that applications depend on structural iso- or homomorphisms between the mathematical objects and the non-mathematical objects that they purport to model. If successful, it explains not only why the specific non-mathematical domain in question can have mathematics applied to it, but equally well why any non-mathematical domain that is structurally similar can have mathematics applied in the same way also. What further common pattern is there to be explained?

Take the example of arithmetic once again as an illustration. Presumably, generality in Dummett's sense here is the datum that Frege himself emphasized so heavily, namely that numbers can be assigned to any objects whatsoever (falling under a particular concept). Frege seeks to explain this with an account of numbers characterizing them in terms of their role in counting. But why think that such an account is the only sort able to explain the datum? Why can a two-stage account of counting not endorse and explain such a claim just as satisfactorily as the neo-Fregean one-stage account, as follows: all objects can be counted, simply because any plurality of objects can be linearly ordered and thereby, if finite, related isomorphically to some initial segment of the natural numbers. I do not see that this explanation lacks any required generality.

4.3.2 Are one-stage accounts required to do justice to the *practice* of applications?

Dummett identifies a second, subtler Fregean argument for one-stage accounts of applicability. He writes:

> The historical genesis of the theory will furnish an indispensable clue to formulating that general principle governing all possible applications. ...Only by following this methodological precept can applications of the theory be prevented from assuming the guise of the miraculous; only so can philosophers of mathematics, and indeed students of the subject, apprehend the real content of the theory.

This line of thought has been developed and expanded by Wright (2000).[45] Against roughly what I have called a two-stage view of applications, Wright offers an intriguing objection. He first notes that it is simply a datum that it is possible to come to appreciate simple truths of (e.g.) arithmetic via their applications. His primary example is a schoolyard demonstration that $4 + 3 = 7$ by simply counting on one's fingers. It is at least plausible that he is right here; surely this is at least one perfectly acceptable route to certain arithmetical knowledge, and indeed, the sort of route originally taken by many. But, the objection continues, it is difficult to square the existence of this route with a two-stage account of applications: it is not that the arithmetical knowledge is

[45] Related issues are further discussed in Wright (2020).

obtained by *first* apprehending the structure of the natural numbers and *then* drawing an inference that the fingers in question can be isomorphically related to some initial segment thereof; rather, it seems that by going through the counting routine one *thereby* grasps the arithmetical proposition itself, and this suggests that the content of the arithmetical proposition cannot be alienated from its application conditions in the way that the two-stage theorist contends.

For the sake of clarity, it's worth emphasizing the complaint is not that two-stage accounts misrepresent the actual order of understanding.[46] Rather, it's that the two-stage account, even as an idealized, rational reconstruction of our practice, cannot allow that naive schoolyard demonstrations are demonstrations of genuinely *arithmetical* propositions. As Wright puts it, two-stage accounts of arithmetic "involve a representation of its content from which an appreciation of potential application will be an additional step, depending upon an awareness of certain structural affinities."[47] As a result, they seem open to the charge of changing the subject.[48]

I want to outline two possible lines of response here. First, and weakest, one might simply accept Wright's point and in response distinguish between different philosophical projects that one might engage in, differing over the centrality they accord to actual mathematical practice. We need to ask: what exactly do we want from a philosophical reconstruction of mathematics? There are different desiderata here, and they plausibly pull in different directions. On the one hand, one might seek an account of mathematics that is (at least approximately) faithful to the actual genealogy of mathematical belief and that, consequently, is able to explain why the schoolyard demonstration is a way of coming to know arithmetical propositions; and this might militate towards giving one-stage accounts of at least the most conceptually basic parts of mathematics. But that is not the only project one might be interested in. For on the other hand, one might instead emphasize the *uniformity* of mathematics, and seek to provide a homogeneous account that takes into account the most sophisticated modern conceptions of the various mathematical domains. This may well push us in the direction of giving two-stage accounts across the board, even at the expense of appearing to create a gulf between sophisticated mathematical beliefs and their more naive counterparts.

But there is a second line of response worth exploring, one that takes a less concessive tack and attempts to reconcile two-stage accounts with the datum that schoolyard

[46] Pincock (2011) seems to read the objection in this way. But this cannot be what is meant, for even the neo-Fregean's own approach is *highly* intellectualized, and in particular, far too intellectualized to plausibly serve as a reconstruction of the actual order of understanding.

[47] Wright (2000, 327).

[48] I should emphasize that Wright does not argue that Frege's constraint holds across the board. Rather, he thinks, it holds when our initial understanding of a mathematical domain involves applications; but, in his view, when it does *not*—as is plausible for, e.g., complex analysis—then a two-stage account may well be acceptable.

demonstrations are (or at least can be) demonstrations of mathematical propositions. The crucial point to note here is that we seem to require a good deal of collateral conceptual mastery to count these basic counting routines as genuine demonstrations of arithmetical facts. What I mean by this is simply that we seem to require the performer of the routine to demonstrate an awareness that it is in a certain sense *general*, to not merely conclude on its basis that 4 *fingers* and 3 *fingers* are 7 *fingers*, but that 4 of *anything* plus 3 more make 7 things, *whatever things they may happen to be*. Getting genuine arithmetical knowledge from this kind of demonstration requires recognition that the particular features of the objects involved, aside from their cardinality, are irrelevant. (Consider how reluctant we would be to ascribe knowledge that $4 + 3 = 7$ to someone who goes through the routine but only seems to appreciate the identity as applied to the number of *fingers*.) I think this strongly suggests, *contra* Wright, that we *do* require an awareness of structural affinities—between e.g. the fingers, the numerals used, the natural numbers they refer to, and indeed any other objects of the same cardinality—in order for instances of the finger-counting routine to succeed in bestowing *arithmetical* knowledge (as opposed to merely knowledge about the particular fingers).[49] And if this is all right, then it is a non-sequitur to think that the two-stage explanation of applications cannot do justice to the phenomena.[50]

5 Conclusion

The main thrust of this essay has been that many of the subtle and ingenious resources devised by abstractionist neo-Fregeans to elaborate and defend their view can, with equal justice, be appropriated by neo-Hilbertians, who construe the axioms of mathematical theories themselves as implicit definitions. Furthermore, I've argued that such an approach has certain definite advantages over neo-Fregean abstractionism, and that it can plausibly respond to the main criticisms that have been raised against it. Perhaps these criticisms can be developed further, or additional ones can be pressed. But I hope that at the very least, the Hilbertian approach is seen as a plausible contender, worthy

[49] I am not saying that the *only* way of making sense of this generality requirement involves a recognition of structural affinities: indeed, neo-Fregeans offer an alternative explanation, since on their view the concept of number is given by HP, which itself provides the resources to attribute number to any concept. The dialectical point still stands, however: once it is recognized that schoolyard derivations involve further conceptual mastery, the immediate objection against two-stage accounts is substantially weakened. Thanks to a referee for discussion here.

[50] To avoid misunderstanding, my claim that this additional step—the recognition of structural affinities—is required is not a claim that the resulting belief that $4 + 3 = 7$ needs to be *inferred* from the schematic application to fingers. I am not making any claims about the architecture of inference at all. Rather, I am making claims about the architecture of *justification*, and it is this that I think supports, or at least is consistent with, the two-stage account.

of further consideration for those who are tempted by the allure of neo-Fregeanism. [51]

Appendix

We prove a result, Lemma 1, which entails both Theorem 1 and Theorem 2.

We work with second-order languages without function symbols and only monadic higher-order variables, for simplicity. Moreover, we assume the only logical symbols are $=$, $\neg$, $\vee$, and $\exists$. First-order languages are understood as fragments of a second-order language.

If $\mathscr{L}$ is a second-order language, model/interpretation $\mathcal{M}$ of $\mathscr{L}$ is just a model $\mathcal{M}$ of $\mathscr{L}$'s first-order fragment in which the higher-order quantifiers range over the power set of $\mathcal{M}$'s domain, $|\mathcal{M}|$.

Let $\mathscr{L}_a$ and $\mathscr{L}_b$ be second-order languages with no individual constants in common. Let P_a and P_b be monadic predicate symbols not occurring in $\mathscr{L}_a$ or $\mathscr{L}_b$. For each formula ϕ of $\mathscr{L}_a$, ϕ^* is the result of relativizing all quantifiers in ϕ to P_a, i.e.

$$\phi^* = \begin{cases} \phi & \text{if } \phi \text{ is atomic} \\ \neg\psi^* & \text{if } \phi := \neg\psi \\ \psi^* \vee \chi^* & \text{if } \phi := \psi \vee \chi \\ \exists v\,(P_a(v) \wedge \psi^*) & \text{if } \phi := \exists v\,\psi \\ \exists V\,(\forall v\,(Vv \rightarrow P_a(v)) \wedge \psi^*) & \text{if } \phi := \exists V\,\psi \end{cases}$$

In the last clause, v is an individual variable not occurring in ϕ. Analogously, if ϕ is a formula of $\mathscr{L}_b$, ϕ^* is the result of relativizing all quantifiers in ϕ to P_b. Let $\mathscr{L}^*$ extend $\mathscr{L}_a \cup \mathscr{L}_b$ with P_a and P_b. If Γ_a is a set of sentences of $\mathscr{L}_a$ and Γ_b a set of sentences of $\mathscr{L}_b$, then Γ_a^* is the set of sentences ϕ^* of $\mathscr{L}^*$ such that $\phi \in \Gamma_a$, and similarly for Γ_b^*.

Lemma 1. *If Γ_b is satisfiable and, for some $\phi \in \mathscr{L}_a$ ($\phi \in \mathscr{L}_b$), $\Gamma_a^* \cup \Gamma_b^* \vDash \phi^*$, then $\Gamma_a \vDash \phi$ ($\Gamma_b \vDash \phi$).*

Proof. Assume Γ_b is satisfiable, so it must have a model, $\mathcal{M}_b$. Assume, for reductio, that for some sentence $\pi \in \mathscr{L}_a$, $\Gamma_a^* \cup \Gamma_b^* \vDash \pi^*$ but $\Gamma_a \nvDash \pi$. Thus, $\Gamma_a \cup \{\neg\pi\}$ must have a model $\mathcal{M}_a$ which is, trivially, a model of Γ_a too. We establish the result by showing that

[51] I am grateful to two anonymous referees for many constructive suggestions. Thanks to Andrea Christofidou, Hannes Leitgeb, Sabina Lovibond, Steven Methven, Beau Mount, Michail Peramatzis, Martin Pickup, Andrea Sereni, Stephen Williams, Crispin Wright, Luca Zanetti, and audiences at New York University, Worcester College, Oxford, the Munich Center for Mathematical Philosophy and IUSS-Pavia, for very helpful comments and discussion. Special thanks to Lavinia Picollo and Jared Warren for detailed comments on several drafts.

$\mathcal{M}_a$ and $\mathcal{M}_b$ can be extended to a model $\mathcal{M}^*$ of $\Gamma_a^* \cup \Gamma_b^*$; thus we have $\Gamma_a^* \cup \Gamma_b^* \nvDash \pi^*$, contradicting our assumption.

Let $\mathcal{M}^*$ be the following model of $\mathscr{L}^*$:

1. $|\mathcal{M}^*| = |\mathcal{M}_a| \cup |\mathcal{M}_b|$.
2. If c is an individual constant of $\mathscr{L}_a$, then $c^{\mathcal{M}^*} = c^{\mathcal{M}_a}$, and similarly for $\mathscr{L}_b$. This is always possible because $\mathscr{L}_a$ ad $\mathscr{L}_b$ don't share any individual constants.
3. If P is a relation symbol exclusively of $\mathscr{L}_a$, $P^{\mathcal{M}^*} = P^{\mathcal{M}_a}$, and similarly for $\mathscr{L}_b$. If P occurs both in $\mathscr{L}_a$ and in $\mathscr{L}_b$, $P^{\mathcal{M}^*} = P^{\mathcal{M}_a} \cup P^{\mathcal{M}_b}$.
4. $P_a^{\mathcal{M}^*} = |\mathcal{M}_a|$ and $P_b^{\mathcal{M}^*} = |\mathcal{M}_b|$.

If σ is an assignment over a model $\mathcal{M}$ and $d \in |\mathcal{M}|$, σ_v^d is identical to σ except it maps the individual variable v to d. Similarly, if $S \in \mathscr{P}(|\mathcal{M}|)$, σ_V^S is identical to σ except it maps the second-order variable V to S.

Note that, by 1, every assignment σ over $\mathcal{M}_a$ or $\mathcal{M}_b$ is an assignment over $\mathcal{M}^*$. We prove without loss of generality, by induction on the logical complexity of the formula $\phi \in \mathscr{L}_a$, that, for every assignment σ over $\mathcal{M}_a$, $\mathcal{M}_a, \sigma \vDash \phi$ if and only if $\mathcal{M}^*, \sigma \vDash \phi^*$.

Note that this would suffice to establish our lemma. For every $\phi \in \Gamma_a$, $\mathcal{M}_a \vDash \phi$, the induction will yield that, for every $\phi^* \in \Gamma_a^*$, $\mathcal{M}^* \vDash \phi^*$; an analogous result for Γ_b can be proved in a similar fashion, so $\mathcal{M}^* \vDash \Gamma_a^* \cup \Gamma_b^*$. Moreover, since $\mathcal{M}_a \vDash \neg\pi$, we must have that $\mathcal{M}^* \vDash \neg\pi^*$, which means that $\Gamma_a^* \cup \Gamma_b^* \nvDash \pi^*$.

- If ϕ is an atomic formula, then $\phi^* = \phi$. Since σ assigns objects taken from $|\mathcal{M}_a|$ to each variable of $\mathscr{L}^*$, by 2 and 3, $\mathcal{M}_a, \sigma \vDash \phi$ iff $\mathcal{M}^*, \sigma \vDash \phi^*$.
- Assume the claim holds of every formula of lower complexity than ϕ.
  - If $\phi := \neg\psi$, then $\phi^* = \neg\psi^*$. Thus, $\mathcal{M}_a, \sigma \vDash \phi$ iff $\mathcal{M}_a, \sigma \nvDash \psi$ iff, by inductive hypothesis, $\mathcal{M}^*, \sigma \nvDash \psi^*$ iff $\mathcal{M}^*, \sigma \vDash \phi^*$.
  - If $\phi := \psi \vee \chi$, then $\phi^* = \psi^* \vee \chi^*$. Thus, $\mathcal{M}_a, \sigma \vDash \phi$ iff $\mathcal{M}_a, \sigma \vDash \psi$ or $\mathcal{M}_a, \sigma \vDash \chi^*$ iff, by inductive hypothesis, $\mathcal{M}^*, \sigma \vDash \psi^*$ or $\mathcal{M}^*, \sigma \vDash \chi^*$ iff $\mathcal{M}^*, \sigma \vDash \phi^*$.
  - If $\phi := \exists x\, \psi$, then $\phi^* = \exists x\, (P_a(x) \wedge \psi^*)$. Thus, $\mathcal{M}_a, \sigma \vDash \phi$ iff there is an $d \in |\mathcal{M}_a|$ s.t. $\mathcal{M}_a, \sigma_x^d \vDash \psi$ iff, by inductive hypothesis, $\mathcal{M}^*, \sigma_x^d \vDash \psi^*$ iff, by 4, $\mathcal{M}^*, \sigma \vDash \phi$.
  - Let $\phi := \exists X\, \psi$, then $\phi^* = \exists X(\forall x(Xx \rightarrow P_a(x)) \wedge \psi^*)$. Thus, $\mathcal{M}_a, \sigma \vDash \phi$ iff there is an $S \subseteq |\mathcal{M}_a|$ s.t. $\mathcal{M}_a, \sigma_X^S \vDash \psi$ iff, by inductive hypothesis, $\mathcal{M}^*, \sigma_X^S \vDash \psi^*$ iff, by 4, $\mathcal{M}^*, \sigma \vDash \phi$.

□

Note that Lemma 1 directly entails Theorem 2. Moreover, Theorem 1 also follows. For if the theories in question are first-order, then by the completeness theorem, (a) both theories are satisfiable iff they are consistent, and (b) one theory is a Field*-conservative extension of the other iff it is a Field-conservative extension; note that the last case of the previous induction is not relevant for such theories.

References

Batitsky, V. (2002). Some Measurement-Theoretic Concerns About Hale's 'Reals by Abstraction'. *Philosophia Mathematica 10*(3), 286–303.

Benacerraf, P. (1973). Mathematical Truth. *The Journal of Philosophy 70*(19), 661–679.

Burgess, J. (2004). E Pluribus Unum: Plural Logic and Set Theory. *Philosophia Mathematica 12*(3), 193–221.

Burgess, J. (2005). *Fixing Frege*. Princeton University Press.

Button, T. and S. Walsh (2018). *Philosophy and Model Theory*. Oxford University Press.

Cook, R. (2016). Conservativeness, Cardinality, and Bad Company. In P. Ebert and M. Rossberg (Eds.), *Abstractionism*, pp. 223–246. Oxford University Press.

Dummett, M. (1991a). *Frege: Philosophy of Mathematics*. Harvard University Press.

Dummett, M. (1991b). *The Logical Basis of Metaphysics*. Harvard University Press.

Ebert, P. and S. Shapiro (2009). The Good, the Bad and the Ugly. *Synthese 170*(3), 415–441.

Evans, G. (1982). *The Varieties of Reference*. Oxford University Press.

Field, H. (1984). Platonism for Cheap? Crispin Wright on Frege's Context Principle. *Canadian Journal of Philosophy 14*, 637–62.

Field, H. (2016). *Science Without Numbers* (Second ed.). Oxford University Press.

Fine, K. (2002). *The Limits of Abstraction*. Oxford University Press.

Frege, G. (1982). *Philosophical and Mathematical Correspondence*. Blackwell.

Frege, G., P. A. Ebert, and R. T. Cook (2013). *Gottlob Frege: Basic Laws of Arithmetic*. Oxford University Press.

Hale, B. (2000). Reals by Abstraction. *Philosophia Mathematica 8*(2), 100–123.

Hale, B. and C. Wright (2000). Implicit Definition and the A Priori. In P. Boghossian and C. Peacocke (Eds.), *New Essays on the A Priori*, pp. 286–319. Oxford University Press.

Hale, B. and C. Wright (2001). *The Reason's Proper Study: Essays Towards a Neo-Fregean Philosophy of Mathematics*. Oxford University Press.

Hale, B. and C. Wright (2009a). Focus Restored: Comments on John MacFarlane. *Synthese 170*(3), 457–482.

Hale, B. and C. Wright (2009b). The Metaontology of Abstraction. In D. Chalmers,

D. Manley, and R. Wasserman (Eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, pp. 178–212. Oxford University Press.

Heck, R. K. (2000). Cardinality, Counting, and Equinumerosity. *Notre Dame Journal of Formal Logic 41*(3), 187–209.

Hellman, G. (1989). *Mathematics Without Numbers: Towards a Modal-structural Interpretation*. Clarendon Press.

Horsten, L. (2010). Impredicative Identity Criteria. *Philosophy and Phenomenological Research 80*(2), 411–439.

Linnebo, Ø. (2009). Introduction [special issue on the Bad Company objection]. *Synthese 170*, 321–329.

MacBride, F. (2003). Speaking with Shadows: A Study of Neo-Logicism. *The British Journal for the Philosophy of Science 54*(1), 103–163.

MacFarlane, J. (2009). Double Vision: Two Questions About the Neo-Fregean Program. *Synthese 170*(3), 443–456.

Panza, M. and A. Sereni (2019). Frege's Constraint and the Nature of Frege;s Foundational Program. *Review of Symbolic Logic 12*(1), 97–143.

Pincock, C. (2011). *Mathematics and Scientific Representation*. Oxford University Press.

Sereni, A. (2019). On the Philosophical Significance of Frege's Constraint. *Philosophia Mathematica 27*(2), 244–275.

Shapiro, S. (1991). *Foundations Without Foundationalism: A Case for Second-Order Logic*. Oxford University Press.

Shapiro, S. (1997). *Philosophy of Mathematics: Structure and Ontology*. Oxford University Press.

Shapiro, S. and A. Weir (1999). New V, ZF and Abstraction. *Philosophia Mathematica 7*(3), 293–321.

Snyder, E., R. Samuels, and S. Shapiro (2020). Neologicism, Frege's Constraint, and the Frege-Heck Condition. *Noûs 54:1*, 54–77.

Tennant, N. (1978). *Natural Logic*. Edinburgh University Press.

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

Wright, C. (1999). Is Hume's Principle Analytic? *Notre Dame Journal of Formal Logic 40*(1), 6–30.

Wright, C. (2000). Neo-Fregean Foundations for Real Analysis: Some Reflections on Frege's Constraint. *Notre Dame Journal of Formal Logic 41*(4), 317–334.

Wright, C. (2007). On Quantifying Into Predicate Position: Steps Towards a New(tralist) Perspective. In M. Leng, A. Paseau, and M. Potter (Eds.), *Mathematical Knowledge*, pp. 150–74. Oxford University Press.

Wright, C. (2016). Abstraction and Epistemic Entitlement: On the Epistemological Status of Hume's Principle. In P. Ebert and M. Rossberg (Eds.), *Abstractionism*, pp. 161–

185. Oxford University Press.
Wright, C. (2020). Is There Basic *A Priori* Knowledge of Necessary Truth? Elementary Arithmetic as a Case Study.