Expansionist abstraction

Jonathan Payne

Thesis submitted for the degree of Doctor of Philosophy

Department of Philosophy
University of Sheffield

February 2013

Abstract

The subject of this thesis is a position in the philosophy of mathematics—defended by Bob Hale and Crispin Wright—known variously as *neo-Fregeanism*, *neo-Logicism* or *abstractionism*, and which claims that knowledge of mathematical objects can be based on principles—known as *abstraction principles*—which are in important respects like definitions of mathematical language.

In the thesis, I make a distinction between two ways in which the abstractionist programme might be carried out. These are the standardly defended *static* view, according to which abstraction principles can used to discover previously unrecognised objects lying within some fixed domain of quantification. The second is an *expansionist* view, according to which abstraction principles allow one to introduce new quantificational vocabulary, and thus expand one's domain of quantification to one containing referents of mathematical terms.

There are then two main aims. The first is to examine the static position, so as to identify the components of that view which make it committed to a standard domain, and to argue against the view. My main argument against the view concerns what has become known as the *bad company problem*. I argue that there is an epistemological component to the bad company problem which can not be avoided by the static abstractionist.

The second aim of the thesis is to argue for and defend the expansionist view. In particular, I will claim that the expansionist view avoids the bad company problem, and that the expansionist view allows for an abstractionist foundation for set theory—an aim which (or so I will argue) has so far eluded the static view.

Acknowledgements

I have been fortunate to have benefited greatly from the help of a large number of people during the preparation of this thesis. To them all I am most grateful.

First of all, I would like to thank my primary supervisor Bob Hale and secondary supervisor Rosanna Keefe, who have been a constant source of incredibly useful criticism, discussion and support. I would also like to thank Øystein Linnebo, who supervised me when I visited Birkbeck College and his *Plurals, Predicates and Paradox* project there for a semester in 2011. My experience during that time has been invaluable.

Thanks to the postgraduate community at the University of Sheffield. The opportunity to present my work several times at the weekly graduate seminar has been of immense help. I have benefited greatly from the ensuing questions, comments and discussions, both in and out of the seminar room. Likewise, the community at Birkbeck, and especially all those on the PPP project, have been an excellent source of stimulating discussion.

I have presented various aspects of this thesis at conferences and workshops in London, Cambridge, Bristol, Barcelona, Ghent and Nancy. Thanks to all those who provided me with very useful question, comments and discussions at these events.

Thanks also to Aldo Antonelli and Marco Panza, both of whom very kindly sent me detailed feedback on early drafts of what has become chapter 7.

The Arts and Humanities Research Council supplied funding for me whilst I completed the thesis. Without this funding, I would not have been able to undertake the project.

Finally, special thanks to Katie Potter for her unerring love and support over the past years, as well as for her proofreading, and for putting up with me.

Contents

Part I

Static abstraction

Chapter 1

Logicism and neo-logicism

The aim of this thesis is to examine the prospects of a philosophical foundation of mathematics based on a certain class of axioms known as *abstraction principles*. These are, roughly, criteria of identity of the form:

$$\text{the abstract of } F = \text{the abstract of } G \text{ iff } R(F, G)$$

where R is an equivalence relation on the domain of the (usually second-order) variables F and G. I will say in more detail what abstraction principles are later on.

The claim that abstraction principles can play an important role in the metaphysics and epistemology of mathematics has recently been defended by, in particular, Bob Hale and Crispin Wright (e.g. Hale, 1987; Hale and Wright, 2001a; Wright, 1983) (but see also Cook (2009b) for another defender of the position). Since it is usually claimed that abstraction principles are in certain respects like definitions, and that much of mathematics follows from these together with logic, such a position is claimed to be a form of logicism. This position is known variously as *neo-logicism*, *neo-Fregeanism* or *abstractionism*.

The purpose of this chapter will be to set the scene for what I intend to argue later on. Sections 1.1–1.2 will give an overview of the relevant background, section 1.1 dealing with Frege's logicism—as espoused in his *Grundlagen der Arithmetik* (Frege, 1884) and *Grundgesetze der Arithmetik* (Frege, 1893)—and section 1.2 with the neo-logicist programme of Hale and Wright.

In section 1.3, I will argue that a distinction should be made in how one may think of the role that abstraction principles play, or are intended to play. This will be between a *static* approach—according to which the role of abstraction principles is to expose previously unrecognised existential commitments in some fixed domain of quantification—and an *expansionist* approach—according to which the role of abstraction principles is to allow one to expand one's domain of quantification. This distinction will play an increasingly important role in the latter part of this thesis, in which I will argue for and defend the expansionist approach.

1.1 Frege's logicism

On the topic of Frege's logicism, I shall only be brief; my aim is not to give a full historical study. Rather, I intend to give enough background to put abstractionism in context. As such, I shall concern myself mainly with the aspects of Frege's programme which have been made use of by the neo-logicists. In particular, these are his use of abstraction principles (although he did not call them that) and his justification of implicit definitions via the *context principle*.

1.1.1 Aims and motivations

There are two aims that Frege explicitly states in the first few sections of *Grundlagen*—one mathematical, and the other philosophical, though both are related.

The first aim, which is stated in §§1–2, is to complete the task started in the earlier part of the 19th century by the rigorous treatment of analysis by Bolzano, Riemann, Weierstrass and others. That task (according to Frege) was to return mathematics, and in particular arithmetic (constituted broadly, so as to include the theory of real numbers, complex numbers and so on) to 'the old Euclidean standard of rigour' (p.1). So, as analysis shows that 'the concepts of function, of continuity, of limit and of infinity have been shown to stand in need of sharper definition' (p.1), so too should the same attention be paid to the concept of *number*, and other basic arithmetical (considered narrowly) concepts. Similarly, in analysis 'proof is now demanded of many things that formerly passed as self-evident'. Frege does not give an example, but one might be the seemingly self-evident claim that, given a continuous function of reals which takes a negative value for some argument a, and a positive value for another argument b, there must be some point in between a and b where the function takes the value of 0 (this is the *intermediate value theorem*).[1] So then, Frege's aim was to give proofs for similarly seemingly self evident arithmetical propositions, such as '5 + 7 = 12' and the associativity of addition.

Now, this is not clearly a peculiarly *logicist* aim; there is nothing particularly logicist in character about the rigourisation of analysis, so why should there be about the rigourisation of arithmetic? However, if the analogy is to be taken to its conclusion, it can be seen that a logicist gloss is rather appropriate. The achievements of analysis might be seen to arise from a process of firstly sharply defining concepts, and then proving propositions just using those definitions. So, for example, the intermediate value theorem might be seen to follow fairly immediately from the rigorous ϵ-δ definition of continuity. But, of course, this is not quite the case. The eligibility of the definitions involved, and the application of them will depend, in this case, on more basic arithmetical propositions. So, in some sense, this amounts to a reduction of analysis to arithmetic.[2]

So, if Frege's aim is to continue this process, the aim will be to produce definitions for arithmetical concepts, and to prove seemingly self-evident propositions concerning them. Then, if *these* depend on more basic assumptions (in the same way as the proof of the intermediate value theorem depends on arithmetic), then these assumptions should

[1] The first rigorous proof of this theorem was by Bolzano (1817). Prior to this, most 'proofs' involved an appeal to geometric intuition.

[2] Of course, we now know that to move from arithmetic to analysis, at least a modicum of set theory is required. But Frege considered class theory to be a part of logic. So, for him, the work of the analysts may have seemed to be a (perhaps partial) reduction of analysis to arithmetic.

be based on rigorous definitions, and so on. If carried to its conclusion (and, if indeed it is possible to carry it on to its conclusion), a desirable outcome would be for the final grounding to be on pure logic together with definitions.

Frege's second aim, which he put forward in §3, is more squarely aimed at philosophers. It is to give an answer to Kant's questions of whether mathematics is *a priori* or *a posteriori*, and whether it is analytic or synthetic. Frege's aim ultimately is to argue that arithmetic is analytic (and hence also *a priori*). It is in this aim that Frege's project is more explicitly avowed to be a form of logicism, if logicism is taken to be the thesis that mathematics can be reduced to logic plus definitions.

Frege's view that mathematics is analytic depends on a particular notion of analyticity that he puts forward, but which goes beyond Kant's notion of analyticity (albeit claimed by Frege to merely amount to 'stat[ing] more accurately what … Kant … [has] meant by them' (p.3 fn. 1)).[3] For Frege, the key to whether a sentence is analytic or synthetic, *a priori* or *a posteriori* is how it can be justified, and, in particular, what a *proof* of it will depend on—the 'ultimate ground' of the proposition. So, concerning the analytic/synthetic distinction, he writes:

> The problem becomes, in fact, that of finding the proof of a proposition, and following it up right back to the primitive truths. If, in carrying out this process, we come only on general logical laws and definitions, then the truth is an analytic one, bearing in mind that we must take account also of all propositions upon which the admissibility of any of the definitions depends. If, however, it is impossible to give the proof without making use of truths which are not of a general logical nature, but belong to the sphere of some special science, then the proposition is a synthetic one. (p.4)

and concerning the *a priori*/*a posteriori* distinction:

> For a truth to be a posteriori, it must be impossible to construct a proof of it without including an appeal to facts, i.e., truths which cannot be proved and are not general, since they contain assertions about particular objects. But if, on the contrary, its proof can be derived exclusively from general laws, which themselves neither need nor admit proof, then the truth is a priori. (p.4)

Thus, the claim that mathematics is analytic is precisely that the truths of mathematics can be proved using only logic and definitions. It can also be seen how this aim matches up with the first, mathematical aim. For the project of carrying the rigourisation of mathematics to its full conclusion must entail finding the 'ultimate ground' of mathematics. If this ultimate ground is just logic and definitions, then it will be shown that mathematics is analytic.[4]

[3] There is much that could be said about the relationship between Frege's notion of analyticity and Kant's (e.g. Dummett, 1991, ch. 3; MacFarlane, 2002; Benacerraf, 1981; Blanchette, 1994), but I shall not be concerning myself with this to any great extent.

[4] It is important to note that, for Frege (and indeed, for neo-logicists), 'logic' means at least second-order logic. That is, as well as permitting quantification into nominal position—with quantifiers ranging over objects—quantification into predicate position is also permitted. These quantifiers range over *concepts* (which stand to predicates as objects do to singular terms). Indeed, as we shall see, concepts play a crucial role in both Frege's programme and in the neo-logicist reconstruction of it.

Since the notions of analyticity and *a priority* which Frege is working with are broadly epistemological (concerning, as they do, justification), it follows that his aims are also broadly epistemological (though this has been disputed, by, e.g. Benacerraf (1981)). Frege is not particularly explicit on this point, however, and it is not clear how much of a role it plays in his views. It is worth however flagging it up now, since similar epistemological aims *are* explicit in neo-Fregeanism.

1.1.2 Definition, the context principle and abstraction principles

Since the notion of *definition* plays such a prominent role in Frege's stated aims, it is important for him to be clear about two questions. The first is: given that part of the aim is to provide definitions for arithmetical vocabulary, for what parts of arithmetical vocabulary in particular must definitions be furnished? The second is a demand for a fairly general account of how definitions may be provided, and, more specifically, how to provide the definitions for arithmetical vocabulary.

I shall not say much about the first question. After some consideration of various competing options, Frege decides that we must supply definitions for terms of the form 'the number of *F*s', where *F* denotes a concept. There are two features to this: the first is that numbers correspond to *concepts*, rather than, for example, physical aggregates or 'sets of units'. This is to avoid various problems with other approaches that Frege sets out in §§21–44. The second feature is that the terms to be defined are singular terms, rather than, for example, adjectives. By doing so, both adjectival and nominal uses of number words can be accounted for. So, for example, an adjectival use of 'seven' in 'there are seven cities in Yorkshire' can be paraphrased using 'seven' nominally: 'the number of cities in Yorkshire is seven' (where 'is' signifies identity, rather than predication). By contrast, some nominal uses of number words can not easily be expressed in terms of adjectival uses. Consider, for example 'the number of cities in Yorkshire is prime'.[5]

The details of the second question are of greater significance to the neo-logicist programme, since two aspects of the answer—the use of the context principle to (perhaps) justify implicit definitions, and the use (to some extent) of abstraction principles—play a central role in the neo-logicist treatment.

What about the first part of this question: how is it, in general, that we are to define a word? That is, how do we confer a word with meaning? Frege answers this requirement—at least partly and at least as expressed in *Grundlagen*—with his *context principle*.[6]

Frege states the context principle right at the beginning of *Grundlagen* (p.x), along with two other 'fundamental principles'. It is

> never to ask for the meaning of a word in isolation, but only in the context of a proposition.

[5]Though see Dummett (1991, ch. 9) for criticism of Frege's hastiness in arguing for a purely nominal analysis of number.

[6]The precise role that the context principle plays in *Grundlagen*, the extent to which Frege retained it (albeit non-explicitly) in later writings, and the relationship between the context principle and his later distinction between sense and reference are controversial (See, e.g. Dummett (1991, ch. 16–17), Milne (1986) and the works cited therein). I do not wish to wade into these debates. Instead, since it is the neo-Fregean programme which is my primary interest, for present purposes I will accept the interpretation of Wright (1983, ch. 1) and Hale (1987, ch. 7) that the primary role of the context principle was to justify implicit definitions.

So, if we want to ask after the meaning of a word, we must ask not just of the word itself, but ask what contribution the word makes to the meanings of whole sentences.

Part of the purpose of adopting the principle, which Frege explicitly states, is to avoid a psychologistic view of meaning, whereby the meaning of a word is something like a mental picture that it conjures up. Such a view of meaning, Frege believed, would go against the seeming objectivity of mathematics.

But the context principle also has consequences for definitions. Just as what is required to ask of the meaning of a word is to ask of its contribution to whole sentences, what is required to *give* meaning to a novel word is to stipulate the behaviour of whole sentences in which it appears. And indeed Frege repeats his commitment to the context principle when discussing definitions. So, in §62 he writes:

> How, then, are numbers to be given to us, if we cannot have any ideas or intuitions of them? Since it is only in the context of a proposition that words have any meaning, our problem becomes this: To define the sense of a proposition in which a number word occurs. (p. 73)

Thus, Frege's task is to define arithmetical vocabulary by stipulating the meanings of an appropriate range of sentences in which arithmetical vocabulary occurs. That is, the aim is to give *implicit definitions* of arithmetical vocabulary.

The matter then moves on to the specifics for number terms. What contexts involving number words need to be given truth conditions, and how are these truth conditions to be given? Since it has already been decided by Frege that it is numerical singular terms—which refer to objects—that we are interested in, we have a natural candidate for the kinds of sentences for which we must give truth conditions:

> we have already settled that number words are to be understood as standing for self-subsistent objects. And that is enough to give us a class of propositions which must have a sense, namely those which express our recognition of a number as the same again. If we are to use the symbol a to signify an object, we must have a criterion for deciding in all cases whether b is the same as a. (p. 73)

So, if we write 'NF' for the number of Fs, we must supply the truth conditions for '$a = NF$', where a is another singular term. An important particular case of this is where a is also a number term. That is, we want to supply truth conditions of statements of the form '$NF = NG$'. Moreover, Frege suggests such truth conditions (which he attributes to Hume, although they are perhaps more accurately attributable to Cantor). $NF = NG$ just in case F and G can be put into one-to-one correspondence. If we write the relation of one-one correspondence between the Fs and the Gs as $F \approx G$, we obtain as a principle:

$$\text{(HP)} \qquad \forall F \forall G [NF = NG \leftrightarrow F \approx G]$$

which has become known as *Hume's Principle*.

$F \approx G$ can then be given a definition in pure second-order logic. Frege does not supply one in *Grundlagen*, but does do so in *Grundgesetze*. Various equivalent definitions can be given, but one would be:

$$F \approx G \stackrel{\text{def}}{=} \exists R \left(\begin{array}{l} \forall x(Fx \rightarrow \exists y \forall z((Gz \wedge Rxz) \leftrightarrow z = y)) \wedge \\ \quad \forall x(Gx \rightarrow \exists y \forall z((Fz \wedge Rzx) \leftrightarrow z = y)) \end{array} \right) \tag{1.1}$$

Hume's principle is an *abstraction principle*. These are sentences of the form:

(AP) $$\forall\alpha\forall\beta[\S\alpha = \S\beta \leftrightarrow \alpha \sim \beta]$$

where '§' is an *abstraction operator*, which attaches to terms of the type of α and β to form a singular term, and '$\sim$' is an equivalence relation on the range of the variables α and β. In the case of HP, α and β are second-order variables ranging over concepts, and $\sim$ is the relation of equinumerosity between concepts, as defined in (1.1).

Frege goes on to justify the possibility of using HP as a definition of number in §§63–64, by comparing it to another abstraction principle—that which gives identity conditions for *directions of lines*. It is:

(DE) $$\forall\ell_1\forall\ell_2[D\ell_1 = D\ell_2 \leftrightarrow \ell_1 \parallel \ell_2]$$

Here, ℓ_1 and ℓ_2 range over *lines*, $D\ell$ denotes the direction of ℓ, and $\parallel$ is the relation of parallelism. This abstraction principle is often called the *direction equivalence*.

There are a couple of notable differences between DE and HP. One is that DE is a *first-order* abstraction principle; the initial quantifiers range over *objects* rather than concepts. As a consequence, the abstracts (i.e. directions) are of the same type as the things which are being abstracted from (lines). Secondly, the equivalence relation on the right hand side is non-logical. Unlike equinumerosity, there is no obvious paraphrase of parallelism in purely (second-order) logical language. Although these differences will be important later on, for the moment they do not matter much.

The purpose of introducing DE is to justify the possible use of abstraction principles as definitions. Frege introduces the metaphor of *content recarving* to explain how such implicit definitions may work. The idea is that the left hand side of an abstraction principles serves to *recarve* the content expressed on the right hand side in a different way:

> The judgement "line a is parallel to line b" …can be taken as an identity. If we do this, we obtain the concept of direction, and say: "the direction of line a is identical with the direction of line b. Thus we replace the symbol $\parallel$ by the more generic symbol $=$, through removing what is specific in the content of the former and dividing it between a and b. We carve up the content in a way different from the original way, and this yields us a new concept. (pp. 74–5)

Presumably, the context principle is at work here; by recarving the content in this way, we succeed in providing the content of sentences involving new vocabulary, and thus confer this new vocabulary with meaning.[7]

However, Frege is ultimately not satisfied with DE, nor, for the same reasons, with HP. Although they both give the truth conditions for identity contexts of the form $D\ell_1 = D\ell_2$ (respectively $NF = NG$), they do not give the truth conditions of identity contexts in general. So, he writes:

> [O]ur definition [DE] affords us a means of recognising this object [the direction of a] as the same again, in case it should happen to crop up in

[7] *Grundlagen* was written before Frege made his famous distinction between sense and reference. As such, it is unclear which of these (if either) is meant by 'content'.

> some other guise, say as the direction of b. But this does not provide for all cases. It will not, for instance, decide for us whether England is the same as the direction of the Earth's axis—if I may be forgiven an example which looks nonsensical. Naturally no one is going to confuse England with the direction of the Earth's axis; but that is no thanks to our definition of direction. (pp. 77–8)

This has become known as the *Julius Caesar problem*, after a similar remark that Frege makes earlier concerning an adjectival analysis of numerical vocabulary:

> [W]e can never—to take a crude example—decide by means of our definitions whether any concept has the number JULIUS CAESAR belonging to it, or whether that familiar conqueror of Gaul is a number or is not. (p. 68)

Consequently, Frege rejects the method of defining arithmetical vocabulary by means of abstraction principles. Instead, he chooses to give an *explicit* definition in terms of *classes*, or *extensions of concepts*.

1.1.3 Frege's final definition of '*NF*' and Basic Law V

In *Grundlagen*, Frege settles on defining the defining the number of *F*s as being a class of concepts, or the extension of a second-level concept. He writes, as the proposed definition:

> the Number which belongs to the concept F is the extension of the concept "equinumerous[8] to the concept F". (pp. 79–80)

In set-theoretic notation, we might write this as:

$$NF = \{G : G \approx F\} \tag{1.2}$$

In *Grundgesetze*, his definition changes somewhat. There, he does not countenance classes of concepts, but only classes of objects. A very similar definition is still however available: NF is to be the class of all the extensions of all concepts which are equinumerous with F:

$$NF = \{x : \exists G(x = \{y : Gy\} \wedge G \approx F)\} \tag{1.3}$$

For such a definition, Frege obviously requires a theory of extensions or classes. In *Grundlagen*, he does not provide such a theory, but merely asserts that 'I assume that it is known what the extension of a concept is' (p. 80 n. 1). He does, however, provide such a theory in *Grundgesetze*. Extensions are a specific case of *value-ranges* of functions, where the value range of a function f can be thought of more-or-less as the *graph* of f. Concepts are then a specific kind of function, which map objects to truth values.

Governing value ranges is the following principle, which Frege calls Basic Law V. Where εf denotes the value range of f:

$$\forall f \forall g[\varepsilon f = \varepsilon g \leftrightarrow \forall x_1 \ldots x_n(f(x_1, \ldots x_n) = g(x_1 \ldots, x_n))]$$

[8] I have replaced Austin's translation of *Gleichzablig* as 'equal' with 'equinumerous', to fit in with my use so far.

In the specific case of concepts (which will be the only case which from now on I shall consider), we have:

(BLV) $$\forall F \forall G \big[\varepsilon F = \varepsilon G \leftrightarrow \forall x (Fx \leftrightarrow Gx) \big]$$

Now, it is clear that again we have an abstraction principle, giving us our third example. Here, the abstraction operator is 'ε', and the equivalence relation is co-extensionality.

Frege does not, however, seek to justify BLV by the same means as HP and DE, at least not explicitly. Indeed there is no mention of the context principle or recarving of content in *Grundgesetze*.

In any case, with BLV and definition 1.3, Frege can prove HP, and from that prove basic laws of arithmetic which are equivalent to the Dedekind–Peano axioms. I shall not go in to any detail about how this is done; to do so would take some time, and would not add much to what I intend to argue for. Sketches of the derivation can be found in Wright (1983, ch. 4), Boolos (1990), Heck (1993) and Zalta (2010).

1.1.4 Russell's paradox and the collapse of Frege's system

Alas, Frege's theory of extensions, as embodied in BLV (together with a principle of substitution which is captured in a modern setting by the comprehension principle for second-order logic), is inconsistent. One consequence of Frege's system in *Grundgesetze* is that every open formula corresponds to a concept, and every concept has an extension. This commitment—as is well known—leads quickly to contradiction.

First, it can be noted that it is possible to define a usual set membership relation '$\in$' using the abstraction operator 'ε' as a primitive:

(1.4) $$x \in y \stackrel{\text{df}}{=} \exists F (y = \varepsilon F \wedge Fx).$$

Then, there will be a concept R corresponding to the formula $x \notin x$, and this in turn will have an extension r. But then it is relatively simple to prove, making use of the equivalence between $x \in \varepsilon F$ and Fx, that $r \in r \leftrightarrow r \notin r$, which is a contradiction.

As a consequence of this discovery—which was communicated to Frege by Russell in a letter (Russell, 1901)—Frege ultimately abandoned his project.

1.2 Neo-Logicism

Recently, Bob Hale and Crispin Wright have defended a revised version of Frege's programme which aims to avoid the problem of contradiction. The aims of neo-logicism are similar to Frege's aims, but with different emphasis. In particular, there is an emphasis on reconciling a realist conception of mathematics with a reasonable epistemology of mathematics.

One aspect of Frege's philosophy of mathematics, which I did not make much mention of in the last section, is that it is *realist* or *platonist*. That is, it takes number terms to be genuinely referential singular terms which refer to abstract objects (and, moreover, that sentences including these terms are true). Hence, the view is committed to there being abstract objects.

Platonism has come under attack in the latter half of the 20th century for epistemological reasons (e.g. Benacerraf, 1973; Field, 1989). The charge is that the platonist view

can not account for knowledge of mathematics without appealing to some mysterious faculty of intuition. Benacerraf argues that knowledge of acausal, non-spatio-temporal objects (as the platonist claims numbers are) conflicts with a causal theory of knowledge, which he claims to be the best theory available. Field makes a similar argument, but without relying upon a causal theory of knowledge (at least, not *explicitly*). His argument is that the platonist will not be able to explain the correlation between mathematicians' beliefs and mathematical truths (which will be about an abstract mathematical realm, according to the platonist).

One of the main aims of neo-logicism is to respond to such arguments. If something like Frege's project were successful, then an answer can be given to the epistemological challenges: It is not the case that, in order to have knowledge of some sentence which involves mathematical singular terms that we must *first* have some access to their referents (be it through causal mechanisms or through some faculty of intuition). Rather, it is the other way around; mathematical terms acquire reference through their use in whole sentences. Moreover, since some of these sentences may serve as something like definitions, they will be analytic, and hence knowable *a priori*.

Hale and Wright have sought to resuscitate Frege's programme by making use of HP directly, without an intermediate appeal to the inconsistent BLV. In this section, I shall sketch the details of their neo-logicist programme. In 1.2.1–1.2.3 I shall sketch the formal development of the programme—how HP fits into a consistent second-order theory which suffices for the derivation of much of arithmetic. In 1.2.4, I will discuss the epistemological claims that Hale and Wright make.

1.2.1 Frege arithmetic

It turns out that it is possible to interpret all of arithmetic in the system that results from adding HP to second-order logic—a system known as *Frege Arithmetic* (Boolos, 1987). Moreover, the resulting system is—in contrast to BLV—consistent. Before going into details of how this is done, it will be useful first to get a little clearer on the notation which I am using (and which I will be using for the remainder of the thesis).

Recall that I have treated the 'number of' operator (and indeed, all abstraction operators) as denoting a function from concepts to objects. I.e. it attaches to a concept term, such as a second-order variable or predicate, to produce a singular term. Such a function is sometimes called a *type-lowering* function, since it maps second-order entities to first order entities.

An alternative way to express HP and other abstraction principles would be by means of a *variable-binding term-forming operator* (VBTO). That is, a symbol which binds a free variable in an open formula—as a quantifier does—and results in a singular term. In the case of the number operator, this is often denoted '$Nx:\phi(x)$'.

The advantage of VBTOs is that they allow one to refer to the abstracts of concepts defined by specific formulas. For example, the number of things which are F-or-G can be denoted by '$Nx:(Fx \vee Gx)$'. Such a notation is not *directly* allowable with the functional approach; something like $N(F \vee G)$ is simply not well formed.

However, if we stick to the functional approach, a VBTO can be emulated in a couple of different ways. The first is to treat $Nx:\phi(x)$ as a definite description. So, we define:

$$Nx:\phi(x) \stackrel{\text{df}}{=} (\imath x)\big[\exists F(\forall y(Fy \leftrightarrow \phi(y)) \wedge x = NF\big]$$

Then, by using the usual Russellian contextual definition of definite descriptions, VBTOs can always be eliminated in favour of terms formed with the function.

Alternatively, a device for 'naming' concepts could be introduced into the language, such as λ-abstraction: Where ϕ is a formula with x free, $\lambda x \phi$ will be a concept term (i.e. a predicate) for the concept defined by ϕ. More generally, where $x_1, \ldots, x_n$ are free variables in ϕ, $\lambda x_1 \ldots x_n \phi$ will be an n-ary relation term. With the λ notation, $Nx{:}\phi(x)$ can simply be an abbreviation of $N(\lambda x \phi)$.

Then, either λ-abstraction can become an official part of the language, by adding appropriate introduction and elimination rules:

$$(\lambda\text{-I}) \frac{\phi(t_1/x_1, \ldots, t_n/x_n)}{(\lambda x_1, \ldots, x_n \phi) t_1 \ldots t_n} \qquad (\lambda\text{-E}) \frac{(\lambda x_1, \ldots, x_n \phi) t_1 \ldots t_n}{\phi(t_1/x_1, \ldots, t_n/x_n)}$$

Or, they can be eliminated in a similar Russellian way.

The functional approach has an advantage over the VBTO approach when it comes to considering models of abstraction principles (which I shall be doing at various points). For then, the interpretation of an abstraction operator will simply be a function from the second-order domain (usually the powerset of the first-order domain) to the first-order domain.[9] As such, I will officially be treating abstraction operators as type-lowering functions. But the notation for VBTOs will still feature, as unofficial abbreviations.

So, Frege Arithmetic (FA) can be defined formally. The language is that of second-order logic with a single non-logical constant: 'N', whose type is such that, where F is a second-order variable, NF is a well formed term of the same type as the first-order variables. FA is the theory in this language which consists of second-order logic, including the full comprehension scheme:

$$\exists F \forall x (Fx \leftrightarrow \phi)$$

and HP as its only axiom.

1.2.2 Frege's Theorem

As mentioned, HP together with second-order logic suffices to interpret second-order arithmetic. I.e. definitions of arithmetical vocabulary (zero, successor, natural number) can be given within the language of HP, and the second-order Dedekind–Peano axioms—as expressed using these definitions—can then be proved using HP. That this is the case is known as Frege's Theorem, since the proof is essentially the one which is sketched by Frege in *Grundlagen* (as with Frege Arithmetic, this term was coined by Boolos).[10]

The first thing to do is to define the usual language of arithmetic within the language of HP. That is, we need a term for zero, a successor relation symbol, and the predicate 'is a natural number'. These definitions can be given as follows:

Definition 1.1. $0 \stackrel{\text{df}}{=} Nx{:}x \neq x$

[9] I will later be considering modal logic, where things get a little more complicated than that. But in that case, the simplicity of the functional approach is even more useful.

[10] Whether the formal proof in *Grundgesetze* could constitute a proof of the theorem is less clear, since it makes use of BLV throughout. Heck (1993) has argued that all appeals to BLV in the proof in *Grundgesetze* (after having derived HP) are inessential.

Definition 1.2. $Sab \stackrel{\text{df}}{=} \exists F \exists x(b = NF \wedge Fx \wedge b = Ny{:}(Fy \wedge y \neq x))$

Definition 1.3. $\mathbb{N}a \stackrel{\text{df}}{=} \forall F\big[F0 \wedge \forall x \forall y(Fx \wedge Sxy \rightarrow Fy) \rightarrow Fa\big]$[11]

Definitions 1.1 and 1.2 are natural and fairly self-explanatory. Definition 1.3 is a specific instance of Frege's definition of the ancestral of a relation (Frege, 1879); an object is a natural number if it is related to 0 by the ancestral of the successor relation.

Then, arithmetic can be interpreted in this system in the form of second-order Peano arithmetic, the system which results from adding the following axioms to second-order logic (with full comprehension):

DP1) $\mathbb{N}0$

DP2) $\forall x(\mathbb{N}x \rightarrow \exists y(\mathbb{N}y \wedge Sxy))$

DP3) $\forall x \forall y \forall z(Sxz \wedge Syz \rightarrow x = y)$

DP4) $\forall x \forall y \forall z(Szx \wedge Szy \rightarrow x = y)$

DP5) $\forall x(\neg Sx0)$

DP6) $\forall F\big[F0 \wedge \forall x \forall y(Fx \wedge Sxy \rightarrow Fy) \rightarrow \forall x(\mathbb{N}x \rightarrow Fx)\big]$

I shall not go through the proof here. For details see Boolos (1990).

1.2.3 Model theory and the consistency of HP

It will be useful for various purposes to consider the model theory of abstraction principles. This is simply a special case of the semantics for second-order logic in general. A model of an abstraction operator will be a pair $\mathcal{M} = \langle D, \S^{\mathcal{M}} \rangle$. D is the domain of objects over which the first-order quantifiers range. The range of the second-order quantifiers can then be modelled by the power set of D, $\mathcal{P}(D)$. Finally, $\S^{\mathcal{M}}$ is the interpretation of the abstraction operator. It is a function $\S^{\mathcal{M}} : \mathcal{P}(D) \rightarrow D$.

Clauses for truth in a model can be given in the standard way. In particular, if F is a concept term which is assigned $X \subseteq D$, 'εF' will be assigned the object $\S^{\mathcal{M}}(X)$.

Given this, where the equivalence relation in the abstraction principle is $\Phi(F, G)$, a structure $\mathcal{M}$ will be a model of the abstraction principle just in case: For all $X, Y \subseteq D$,

$$\S^{\mathcal{M}}(X) = \S^{\mathcal{M}}(Y) \text{ iff } \mathcal{M} \vDash \Phi(X, Y)$$

We can apply this model theory to HP, and in doing so prove that it is consistent. A natural model is achieved by letting $D = \mathbb{N} \cup \{\aleph_0\}$, and $N^{\mathcal{M}}(X) = |X|$ (i.e. the cardinality of the set X). It is easy to then check that this is a model of HP, and hence that HP is consistent (with the same level of certainty as any other mathematical claim). But an improvement can be made; there is a model of HP whose domain is the natural numbers, by letting $N^{\mathcal{M}}$ be:

$$N^{\mathcal{M}}(X) = \begin{cases} |X| + 1, & X \text{ finite} \\ 0, & X \text{ infinite} \end{cases}$$

[11] A note on operator precedence: I will always assume that implication $\rightarrow$ takes lower precedence than the other connectives. That is, '$p \rightarrow q \wedge r$' is equivalent to '$p \rightarrow (q \wedge r)$' and so on. In situations like this I will omit the parentheses.

Again, it can be checked that this is indeed a model of HP (see Boolos, 1987).

Since $N^{\mathcal{M}}$ can be defined in standard second-order arithmetic, we have an interpretation of HP in second-order arithmetic. Since there is an interpretation both ways, we can thus conclude that HP is consistent if and only if second-order arithmetic is consistent. So, we can be sure (at least, as sure as could be reasonably demanded) that HP will not face any similar problems to those suffered by BLV.

1.2.4 Epistemology and Hero

So, HP is sufficient as an axiom for arithmetic. But neo-logicists make a further claim—as already mentioned—that HP can underwrite our *knowledge* of arithmetic in a way in which the usual axioms do not.

The claim is that HP is, if not a *definition* of the number operator, then at least an *explanation* of the number concept; anybody who has an understanding of the number operator will thereby have the means to acquire knowledge of the existence of numbers, and of their properties.

A thought experiment which is useful for the purposes of visualising this claim is suggested by Wright (2001b). He suggests that we picture a character—'Hero'—who makes use of HP to gain knowledge of arithmetic. At the start, Hero neither has knowledge of natural numbers, nor any vocabulary or concepts with which to express any arithmetical claims. Hero is, however, proficient with the language of second-order logic. Wright claims that, by laying down HP as an implicit definition of the number operator, Hero can simultaneously gain an understanding of arithmetical vocabulary and the wherewithal to gain knowledge of arithmetical truths.[12]

So, Hero will start off with an understanding of the various logical connectives, the first- and second-order quantifiers and so on. Just what this understanding consists in—especially in the case of the quantifiers—will play an important role in the distinction that I make later in section 1.3. Hero may also have a smattering of various non-logical predicates and names, perhaps standing for, e.g. physical properties and objects. But, importantly, these will not include any mathematical vocabulary.

Then, Hero may extend his[13] language by the addition of the abstraction operator 'N', which is to be explained by means of HP. Since Hero is provided with an explanation of the new symbol in the form of the abstraction principle, he will thus have an understanding of the new vocabulary. Moreover, by means of HP, Hero will be able to prove various propositions which use the new symbols.

For example, he may deduce that there are numbers thus: Take a concept, say that denoted by $\lambda x(x \neq x)$. It will be a matter of pure second-order logic that $\lambda x(x \neq x) \approx \lambda x(x \neq x)$, and since Hero has a full understanding of second-order logic, he will be able to prove as much. By this stage, nothing concerning numbers has been introduced.

[12] The principal reason for which Wright introduces the story of Hero is actually to claim that this understanding or knowledge can be acquired in a *predicative* manner. I shall discuss this later 5, but for now, the main purpose of introducing Hero is to provide an example concerning the more general claims about epistemology made by the neo-Fregeans.

[13] As a fictional character, there are many irrelevant properties of Hero which are undetermined; Hero has no definitive height, hair colour, nationality, and so on. It is an unfortunate feature of English that another irrelevant feature—Hero's gender—may not be left indeterminate. I have decided to vary which pronouns I use from chapter to chapter.

But now, Hero may reason from right to left across HP, to deduce that $Nx{:}x \neq x = Nx{:}x \neq x$. Thus (by the use of the usual rules for the existential quantifier), Hero can deduce that there is such an object as $Nx{:}x \neq x$, which may be called 0. And the same method can be used to show that, for any concept F, NF exists.

Hero can then go on to gain an understanding of the other numerals. So, 1 is $Nx{:}x = 0$, 2 is $Nx{:}(x = 0 \vee x = 1)$ and so on. Furthermore, Hero will be in a position to make use of other definitions, perhaps those given in (1.1)–(1.3), from which he can prove the Dedekind–Peano axioms. He can then go on to prove any number of arithmetical theorems.

This story—of Hero gaining knowledge of mathematics through the use of abstraction principles—need not be specific to HP and arithmetic. Part of the neo-logicist claim is that this will generalise to some other abstraction principles; in certain circumstances, Hero may lay down an abstraction principle and thereby both gain an understanding of the abstraction operator involved, and knowledge of the resulting abstracts. So, for example, suppose Hero starts off with an understanding of some limited geometrical vocabulary involving lines and parallelism. The abstractionist claim is that he will then be able to gain an understanding and knowledge of *directions* by means of DE. Or, it might be hoped that there is an abstraction principle (or some abstraction principles) will would allow Hero to gain a knowledge of mathematical objects beyond the natural numbers, such as the real numbers or sets.[14]

1.3 Two kinds of abstraction

There are two ways in which one may think of the process of abstraction, the distinction between which will play an important role throughout the thesis. These are between a *static* view, whereby the domain of quantification is fixed throughout the process, and an *expansionist* view, whereby an abstraction principle also may serve to *expand* the domain of the first-order quantifiers.[15]

This is obviously in need of further explanation and detail. But before giving such further details, it will be worth contrasting this distinction (as stated rather vaguely) with a similar one made by Fine (2002, pp.56–7) between *standard* and *creative* definitions. According to Fine, standard definitions 'are made from a standpoint in which the existence of the objects or items that are to be assigned to the defined terms is presupposed.' The purpose of such definitions is to 'make an appropriate assignment of the objects already in the domain to the terms that are to be defined.' By contrast, in a creative definition, 'the existence of the objects that are to be assigned to the terms is not presupposed.' Since this distinction concerns domains of quantification, it seems that it may serve my purposes.

But, as it stands, I do not think that Fine's distinction is particularly useful. The reason is that the distinction mixes up (or at least, fails to sharply distinguish between) an epistemic reading of a definition 'presupposing' objects, and a more semantic or

[14] It should be noted that—at least according to the view of abstraction put forward by Hale and Wright—BLV will not be such an abstraction principle. Being inconsistent, it can not be a source of knowledge.

[15] There may of course be yet further ways in which one may think about abstraction (e.g. Leitgeb (Forthcoming) and Antonelli (2010a,b) have views about abstraction principles which do not fit within either of these paradigms). I will not be concerning myself with such views.

metaphysical reading. So, the distinction sometimes concerns distinctively epistemic notions, such as us being 'sure that the required objects . . . exist', or us 'knowing prior to the definition being made that the objects of the required sort exist.' At other times, however, it merely concerns the domain of quantification. But it is perfectly consistent (and indeed, very natural) to suppose that we may succeed in quantifying over objects which one has no knowledge of.

So, a definition may presuppose some objects in one of two ways—either it presupposes *knowledge* of those objects, or it merely presupposes the ability to quantify over those objects (without the requirement that a speaker knows that they are quantifying over those objects).

Given this distinction, the terminology of 'creative' versus 'standard' definitions is perhaps unfortunate. For a view which does not claim presupposition of *knowledge* of objects need have nothing to do with creation of those objects. Even in the semantic case, literal creation of objects may not be required, since if a definition allows a speaker to quantify over objects that he did not previously quantify over, there is nothing to say that the act of definition brought those objects into existence in some creationist manner[16] It is perhaps less misleading to refer to *presuppositional* and *non-presuppositional* definitions.[17]

It should be clear that, taken in the epistemic sense, it is of no use to regard HP as a presuppositional definition. For then it would presuppose knowledge of infinitely many objects. But this is precisely what the use of abstraction principles by neo-logicists is intended to avoid. As such, the distinction that I intend to make concerns only the semantic presupposition of objects, in a sense that I hope to make clear. Epistemically speaking, both sides of the distinction will claim that abstraction principles are non-presuppositional definitions (in either the epistemic or semantic sense).

1.3.1 The static view

It will be useful to discuss the two views, and the differences between them, within the context of Hero; they will be two different views concerning Hero's understanding of the first-order quantifiers. According to the static view, the domain of quantification over which Hero's quantifiers range remains fixed. So, the same objects are quantified over both before and after laying down an abstraction principle like HP. In particular, Hero's quantifiers will range over the natural numbers before laying down HP, and one of the roles of HP is simply to allow Hero to *discover* this fact.

To make this, and the distinction between the views, clearer, it will be useful to distinguish between two kinds of 'grasping' of a domain of quantification. One will lie firmly on the epistemic side, while the other will be more semantic in character.

Say that a speaker *understands* a particular domain of quantification, D, if there is something (be it their use of quantified language, some particular mental state, or whatever) that fixes the meaning of their quantifiers as being such so as to range over D.[18]

[16] I argue for this claim more fully in chapter 6. Compare also Fine's own (non-abstractionist) view, in Fine (2005) and Fine (2007), which is clearly a creative view in Fine's sense, but which he denies requires the literal *creation* of objects.

[17] See Hale (2006) for a similar point.

[18] Now is as good a time as any to remark on my use of 'D' as a singular term to denote a domain of quantification. There will be points later on in this thesis in which it would be begging the question on my part

But an understanding of a domain in this sense falls well short of knowing substantial facts about the domain, such as what objects fall within it, or its cardinality. I may, for example, be completely proficient in quantifying over all stars, and it be completely determinate over what the quantifier phrase 'all stars' ranges when uttered by me. But, nonetheless, I am still utterly clueless about, say, the cardinality of this domain, even approximately.[19]

This distinction, between knowing about a domain and merely understanding the domain, will lead on to a similar distinction between a couple of domains of quantification that may be discussed in the context of Hero. On the one hand, one may talk about Hero's *semantic domain*—the domain over which Hero's quantifiers actually range, which is (at least partially) fixed by Hero's understanding of the first-order quantifiers. On the other hand, there is what might be called Hero's *epistemic domain* (or, perhaps, range of domains). These will be the domains which are compatible with what Hero *knows* about the range of his quantifiers. Since there will be many domains of quantification which are compatible with what Hero knows at any one point, I will use 'epistemic domain' sometimes to mean the *minimal* epistemic domain, that is, the smallest domain of quantification which is compatible with what Hero knows about the range of his quantifiers.

It is the former of these which is kept fixed under the static view. Before laying down HP, for all Hero knows, there may be just a small finite number of objects over which his quantifiers range (depending on some extent to the non-logical vocabulary that Hero has to start with). Thus Hero's possible epistemic domains include small finite domains. But the *semantic* domain is nonetheless infinite, since it contains all of the natural numbers.

Then, after laying down HP, Hero's semantic domain will remain the same (since that is the hallmark of the static view). But his *epistemic* domain will expand, for now what he knows about his domain of quantification includes the fact that it is infinite.

1.3.2 A toy model for the static view

It will be helpful to explain the distinction between the two views in terms of toy models. These will be, roughly, informal model-theoretic characterisations of the process which Hero goes through in each case.

In the static case, the picture is more or less as described in 1.2.3, when describing what a model of an abstraction principle is. There is one fixed domain, D, and this is what Hero's first-order quantifiers range over. There will also be a domain, perhaps related to D, over which Hero's second-order quantifiers range. Call this **D**. (In the case of the standard semantics of second-order logic, this will be $\mathcal{P}(D)$). We might also consider Hero's epistemic domain as a subset of D, say D_E.

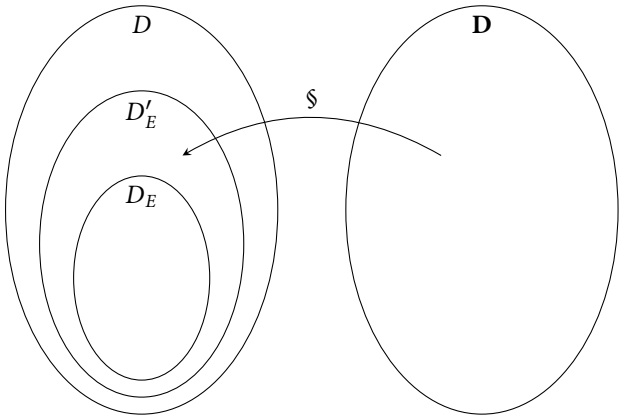
to assume that a domain of quantification is a single object, such as a set (cf. Cartwright, 1994). For this reason, it is important that my use of this singular terminology is not read as being committed to this 'all-in-one' view. Where there is a possibility that this would make a real difference, 'D' can be read as a disguised plural term. Instead of the quantifiers ranging over a single object, D, they may range over *some* objects, the Ds.

[19]This is not to deny that understanding a domain is not itself at least partially an epistemic notion. Presumably an understanding of a domain involves at least some knowledge about the domain, perhaps, for example, what kinds of thing lie in the domain and so on. As such the distinction is not quite one between something epistemic and something non-epistemic, but rather between one kind of knowledge about a domain and another kind of knowledge about that domain.

The abstraction principle will then describe a function § : $\mathbf{D} \to D$ which will serve to provide names for various objects in D, namely the abstracts. Moreover, since it is assumed that Hero started off with no knowledge of the abstracts, the image of § will lie outside of D_E.

Information about this function and the corresponding abstracts (as given by the abstraction principle) can then be used to deduce information about the domain D. This new information will result in a wider epistemic domain, D'_E, which will include the abstracts. The picture which emerges will be something like that illustrated in figure 1.1.

Figure 1.1: The static view



Now, the information about D that results from the abstraction principle could be deduced from the point of view of the object language, as in 1.2.2 and 1.2.4. But it can also potentially be illuminating to view the situation 'from the outside' as it were, in the context of the toy model itself. We can do this by considering a property of abstraction principles which Fine (2002) calls *inflation*.

Consider HP, and a domain D which has cardinality κ. Suppose also that the second-order domain is just that of the standard semantics, so that $\mathbf{D} = \mathcal{P}(D)$. The equivalence relation of equinumerosity will partition $\mathbf{D}$ into $\kappa + 1$ equivalence classes.[20] But since the function denoted by N must map equivalent subsets of D onto the same element of D, we must then have that $\kappa + 1 \leq \kappa$.

Since, for finite κ, $\kappa < \kappa + 1$, HP would require in such a circumstance that there be more numbers than there are objects; it can be said to *inflate* on finite domains. But since a core part of the Fregean view is that numbers *are* objects, this can not be the case. By contrast, there is no conflict in the infinite case, since for infinite cardinalities κ, $\kappa + 1 = \kappa$. Hence, if HP is true, then the original domain must be infinite.

A similar picture can be given to explain the problems with BLV. In this case, the equivalence relation will partition the domain into 2^κ partitions (i.e. one for every subset of D). Since, by Cantor's Theorem, $\kappa < 2^\kappa$ for any κ, BLV will inflate on *any*

[20] Why? There will be subsets of D with every cardinality between 0 (e.g. the empty set), and κ (e.g. D itself). There will be $\kappa + 1$ such cardinalities.

domain, and must therefore be declared unacceptable. We thus have something like a model-theoretic explanation of Russell's paradox.[21]

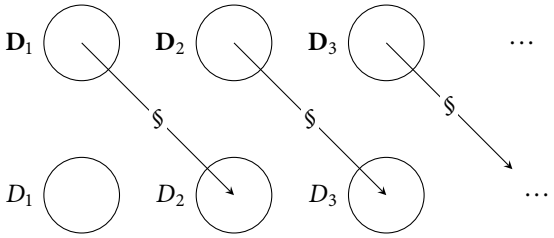
1.3.3 The expansionist view

Whereas on the static view it is only Hero's epistemic domain which expands with the laying down of an abstraction principle, on the expansionist view the semantic domain is also permitted to expand. So, before laying down an abstraction principle, Hero's quantifiers range over some domain, and after laying down the abstraction principle, they may range over a *wider* domain.

Again, a kind of toy model can be described. We will have a starting domain, D_1, with associated second-order domain $\mathbf{D}_1$. Now, instead of the abstraction operator being a mapping $\mathbf{D}_1 \to D_1$, as on the static approach, it will be a mapping $\mathbf{D}_1 \to D_2$, where $D_2 \supseteq D_1$ is the expanded domain. Thus, on this picture, Hero uses the abstraction principle, not to gain knowledge about the original domain D_1, but to gain an understanding of the wider domain of quantification D_2.

Typically, the relation on the right hand side of an abstraction principle will involve quantification, and hence the application of the abstraction principle will depend upon the range of the quantifiers, i.e. on D_1. So, the abstraction principle could be applied *again*, this time with the quantifiers ranging over D_2 (and perhaps with the second-order quantifiers ranging over an associated second-order domain, $\mathbf{D}_2$). Then the abstraction operator will be a mapping $\mathbf{D}_2 \to D_3$, and so on. The emerging picture will be something like that illustrated in figure 1.2.

Figure 1.2: The expansionist picture



This picture, it should be noted, is still compatible with the domain being fixed, as in the static case. For there is nothing to say that the various D_is are not in fact all identical, so that $D_1 = D_2 = \ldots$.

[21] There are reasons, however, to suspect that this may not be as explanatory as it seems. Russell's paradox arises whenever we have full impredicative comprehension. But impredicative comprehension is perfectly compatible with a second-order domain which falls well short of the full power set of the first-order domain, say, the (countable) set of all and only *definable* subsets of D. In this case, Russell's paradox is still present, but inflation will do nothing to explain its presence.

Another way in which inflation may fail to be a satisfactory as an explanation of Russell's paradox concerns features of the toy model outlook, and, in particular, the fact that this treats domains as *sets*, and thus as having a determinate cardinality. If it could be claimed that domains only have a size in some indeterminate manner (say, by having lower bounds to their cardinality, but not an upper bound), then the notion of inflation may not apply.

But, as I shall argue for in more detail in chapter 5, this does not mean that there is little or no formal difference between the two views.[22] The reason is that, although the view is compatible with $D_1 = D_2 = \ldots$, it does not permit us to *assume* that this is the case. This will restrict certain methods of proof, which are essential to the standard proof of the Dedekind–Peano axioms from HP. But, as I will show in chapter 7, this will also allow the blocking of contradictions.

The two views are also different when we consider inflation. Consider again HP over a domain D_1 with cardinality κ. Again, $\approx$ will partition $\mathbf{D}_1$ into $\kappa + 1$ equivalence classes. But this time, N will map the equivalence classes to a possibly expanded D_2, rather than to D_1. Hence, we no longer have the requirement that $\kappa + 1 \leq \kappa$, but merely that $\kappa + 1 \leq |D_2|$. This is not a restriction on the size of D_1, but rather a restriction on how it can expand; if D_1 is finite, then it must be the case that $D_1 \neq D_2$. I.e. the expansion is a genuine (one might say, 'proper') expansion.

Similarly, in the case of BLV, inflation is not such a stumbling block. In this case, we will have the restriction that $2^\kappa \leq |D_2|$. This is not a contradiction; it is simply a proclamation that *whatever* the cardinality of a domain, an application of BLV will require it to expand.

Nothing I have said here should be decisive concerning the differences between these two approaches. For one thing, these are just toy models to give a rough idea of the differences between the two views. For another, far more detail needs to be given, especially about the expansionist view: What does the view look like in the object language? Does it really avoid Russell's Paradox, and if so, how? What is the appropriate metaphysical picture to go along rather unclear concept of 'domain expansion'? And so on. I shall postpone giving such further details until later in the thesis (chapters 5–8), in which I shall defend the expansionist position.

Before then, in the next few chapters, I will set out in more detail how I see the static view, and will push what I see as some of the most major problems with it. In the next chapter, I will argue that the background logic of static abstraction has a certain feature—that it is a *negative free logic*—and connect this feature with the staticness of the domain. This will also have relevance to certain kinds of *restricted* abstraction principles. Then, in chapters 3–4, I will be concerned with the *bad company* problem for static abstraction.

[22] This is in contrast with the attitude that Fine appears to take in his distinction between standard and creative definitions. For in his discussion of the distinction, he does not mention at all any difference that may occur in the technical part of the abstractionist programme as a consequence of conceiving of abstraction principles in one way rather than another. Rather, his main concern seems to be with the philosophical consequences of each view (and, in particular, the relation to the context principle).

Chapter 2

Abstraction and free logic

This chapter concerns a feature of the background logic of abstraction (or, at least, a feature that I claim that it should have) which has received some, but not much, attention in the literature. That feature is that the background logic should be a *free* logic. In section 2.1, I shall say a little about what a free logic is and the difference between a couple of ways of conceiving of it (in particular, between *negative* and *positive* free logic). In section 2.2, I shall argue that neo-Fregeans require a free logic, and, in particular, a negative free logic. Moreover, I shall argue that the use of such a logic is justified, given both the context principle and the assumption of a fixed domain. Finally, in section 2.3 I shall discuss the applications of free logic to a particular kind of *restricted* abstraction principle. The result will be a way of restricting abstraction principles which has—as far as I know—not yet been considered.

2.1 Free logic

2.1.1 Existential presuppositions and quantifier rules

A *free logic* is one in which it is not assumed that every term of the language refers. So, one may have a language which features apparent empty names, such as 'Vulcan' or 'Pegasus', or that admits possibly empty definite descriptions as terms, without eliminating them in the usual Russellian way (although in free logics with definite descriptions, the Russellian contextual definition of definite descriptions can usually be derived as *theorems*). This goes too for other complex singular terms, and—especially importantly for the present purposes—functional expressions of the form $f(a)$ (with a special case being abstract terms).

That every term refers in a classical, non-free logic is a simple consequence of the $\forall$ elimination rule and $\exists$ introduction rule (which are interderivable given an equivalence between $\forall x\phi$ and $\neg\exists x\neg\phi$). These rules are:

$$(\forall\text{-E})\ \frac{\forall x\phi(x)}{\phi(t)} \qquad (\exists\text{-I})\ \frac{\phi(t)}{\exists x\phi(x)}$$

So, for example, given the platitude 'everything exists' ($\forall x \exists y(x = y)$), we can infer $\ulcorner t$ exists$\urcorner$ for any term t, regardless of whether t succeeds in referring or not.

This requirement can also be seen in the usual model theoretic semantics; an interpretation function is required that assigns a referent in the domain to any constant in the language. In addition, every functional term in the language is assigned a *total* function on the domain, so that every functional term of the form '$f(a)$' receives a referent.

Free logics avoid the presupposition that every term refers by modifying the quantifier rules. The ($\forall$-E) and ($\exists$-I) rules are weakened by requiring an additional premise that the terms involved refer, that is, that for a term t, $\exists x(x = t)$. This existence statement I will abbreviate as $E!t$. The resulting rules are:

$$(\forall\text{-E})\ \frac{\forall x \phi(x) \qquad E!t}{\phi(t)} \qquad\qquad (\exists\text{-I})\ \frac{\phi(t) \qquad E!t}{\exists x \phi(x)}$$

The ($\forall$-I) and ($\exists$-E) rules are correspondingly *strengthened*—both allow the discharge of an additional assumption:

$$(\forall\text{-I})\ \frac{\begin{array}{c}\overline{E!t}\\ \vdots \\ \phi(t)\end{array}}{\forall x \phi(x)} \qquad\qquad (\exists\text{-E})\ \frac{\exists x \phi(x) \qquad \begin{array}{c}\overline{\phi(t) \quad E!t}\\ \vdots \\ \psi\end{array}}{\psi}$$

2.1.2 Atomic formulas

This modification of the quantifier rules is common to all approaches to free logic. But there are different attitudes that one can take towards the truth of unquantified sentences involving non-referential terms and, in particular, the truth-values of *atomic* sentences involving non-referential terms. A particularly important special case of this will be statements of identity.

Negative free logics hold that all atomic sentences which feature a non-referential singular term are *false*. In particular, in a negative free logic, even $\ulcorner t = t\urcorner$ will be false if t fails to refer. *Positive* free logics hold that at least some atomic sentences featuring non-referential singular terms—perhaps, e.g. 'Pegasus flies'—are true. In particular, they hold that $\ulcorner t = t\urcorner$ holds for any term whatsoever, regardless of whether t refers or not.[1]

This difference will manifest itself in differences in the rules for identity, and in the existence of a supplementary rule for negative free logic.

In a positive free logic, since $\ulcorner t = t\urcorner$ will hold for any term, we have the following, which could be taken as an introduction rule for identity:

$$(=\text{-I}^{+})\ \frac{}{t = t}$$

[1] There are also *neutral* free logics, according to which sentences involving non-referential terms are truth-valueless. I shall not be considering these. For the present purposes, they will come out much the same as negative free logic, since they will allow an inference from an atomic formula featuring a term t to the existence of a referent of t.

The corresponding elimination rule will be Leibniz's Law:

$$(=\text{-E})\ \frac{s = t \qquad \phi(s)}{\phi(t)}$$

There is a further question as to when $\ulcorner s = t \urcorner$ is true if s and t are distinct non-referring terms. The simplest approach is to take all such sentences as true. But an alternative might be to judge this as true if s and t have the same sense, say, and false otherwise. But, for my purposes, not much hangs on how this choice is made; the only important feature of positive free logic for the present purposes is that identity statements may in some cases be true when the ingredient terms do not refer.

In a negative free logic, $\ulcorner t = t \urcorner$ will only be true when t refers. As such (=-I) must be restricted in a similar way to the quantifier rules:

$$(=\text{-I}^{-})\ \frac{\exists x(x = t)}{t = t}$$

But, since an atomic formula can be true only if all of the component terms refer, we will have also have the following rule:

$$(E!\text{-I})\ \frac{A(t) \quad (A \text{ atomic})}{\exists x(x = t)}$$

This raises the question of just what should count as an atomic formula. In the case of a first-order language, this is a simple question; for each non-logical n-ary relation R, $Rx_1 \ldots x_n$ will be an atomic formula, and $x = y$ will be an atomic formula. In the case of the second-order language that is under consideration, however, there are a couple of issues that may arise. Firstly, should Fx, where F is second-order *variable*, count as an atomic formula? Secondly, if we have lambda abstraction in the language, should $(\lambda x \phi(x))y$ count as an atomic formula? These two questions intersect, with the answer to one of them putting constraints on the answer to the other. It will be fine, however, to postpone this question. For the present purposes, the only atomic formula which will be under consideration will be identity.

It should be noted that, in either positive or negative free logic, a formula which behaves in the same way as the identity of the other logic can be defined. So, from a positive free logic, one can define:

$$s =_{-} t \text{ iff } \exists x(x = s \wedge x = t) \tag{2.1}$$

and in a negative free logic:

$$s =_{+} t \text{ iff } \forall x(x = s \leftrightarrow x = t), \tag{2.2}$$

provided that one is happy with kind of positive free logic in which $\ulcorner s = t \urcorner$ is true when neither s nor t refers.

2.2 Free logic and abstractionism

The appropriate background logic for abstraction principles, I claim, is a free logic. Moreover, it is a negative free logic (at least, in the case of static abstraction). There are two main reasons for this. Firstly, a non-free logic risks begging the question of the existence of abstracts. Secondly, there are abstraction operators for which it is natural to suppose that not every abstract term formed from them refers.

2.2.1 Abstraction and existence assumptions

First, it should be noted that one of the aims of abstractionism is to argue for the existence of numbers and other abstract objects. But, if the background logic is non-free, this would just be an *assumption*. Hence, the position would simply beg the question. Moreover, such an assumption does not appear to be required; it seems that we can argue directly to the existence of numbers from HP. Recall the reasoning by which somebody like Hero is supposed to deduce the existence of numbers. It goes as follows:

(i)	$F \approx F$	(logic)
(ii)	$NF = NF$	(HP)
(iii)	$\exists x(x = NF)$	($\exists$ introduction)
(iv)	$\forall F \exists x(x = NF)$	(second-order $\forall$ introduction)

This proof is supposed to make a substantive use of HP in step (ii). But, in a non-free logic, the conclusion is a trivial logical truth, since 'NF' is bound to refer. For the proof to be of interest, there must be the possibility of the abstract terms failing to refer. And for this to be the case, the logic has to be free.

Moreover, for the proof to be valid, the logic in question must be a *negative* free logic. For in a positive free logic, (iii) does not follow from (ii); in order to apply existential introduction to a formula containing 'NF', one must already have (iii) as a supplementary premise. This is what prevents a positive free logic from collapsing to a non-free logic, since (ii) can be derived in positive free logic without the use of HP (indeed, it is a logical axiom). Without the claim that number terms refer being derivable from HP, so too would the Dedekind–Peano axioms be underivable. HP would need to be supplemented by what would be, in effect, explicit existence assumptions.

In a negative free logic, however, these issues do not occur. The step from (ii) to (iii) does not proceed by means of ($\exists$-I) which would, as in the case of the positive free logic, require (iii) already as a premise. Instead, it can make use of the ($E!$-I) rule, since (ii) is an atomic sentence.

So then, for the purpose of abstractionism, a negative free logic is required. A non-free logic opens the position up to accusations of question begging, and in a positive free logic, it is impossible to prove that the number terms refer, and thus the central technical plank of the abstractionist programme—Frege's Theorem—can not be derived.[2]

Indeed, when the matter arises, neo-logicists appear to claim that the background logic for abstraction principles will be a negative free logic. For example, Hale and Wright (2009a, p.464) claim, in the context of a situation where we 'require, for universal instantiation, a supplementary premise asserting the existence of a referent for the instantial term, with a similar restriction on existential generalisation' (i.e., in a free logic), that 'identity-contexts are so understood that they cannot be true unless their terms have reference' (i.e. the logic is a negative free logic).

That it is a negative free logic that is both required and desired by neo-logicists is thus nigh on indisputable. The question of whether the use of a negative rather than positive free logic is *justified* is, however, another matter. Several commentators have

[2]It might be thought that there could be some argument other than that above which allows one to infer from HP that number terms refer, and that such an argument *is* valid in a positive free logic. A simple model theoretic argument can show that this is not the case; in a positive free logic, HP will have models with an empty domain, and in which no abstract terms is supplied with a denotation.

written things which suggest that they see the appropriate free logic as being a positive free logic, or at least, something similar to a positive free logic. So, for example Potter and Smiley (2001) write:

> Allowing for [empty terms] brings with it the realisation that there are two variant readings of identity. There is a strong reading under which $a = b$ is false if either a or b or both fail to refer. But there is also a weak reading under which $a = b$ is false when one term refers and the other doesn't, but is true—one might say 'vacuously true'—when both terms fail to refer. ... The weak reading comes into its own whenever non-existence is a serious issue. ... Hale [in Hale (1999)] has taken for granted the strong reading of identity. (p.336)

Here, the strong and weak readings of identity clearly coincide with a negative and a positive free logic, respectively. Similarly, Shapiro and Weir (2000) claim that a negative free logic begs the question in much the same as a non-free logic does. They write:

> [I]f [identity] does have existential import, then Frege's Theorem holds but the interpretation of the required abstraction principles ... will beg the question. (p. 188).

Finally, a similar claim is made by Rumfitt (2003, pp. 208–9). On HP he writes:

> Within a free logic, in such a context it bifurcates in an interesting way. On the one side, we have that part of Hume's principle which gives the criterion of identity for numbers when a number belongs to a concept, viz.:
>
> > (HP minus) For any concepts X and Y, if there is such an object as the number of Xs, or the number of Ys, then the number of Xs is identical with the number of Ys iff there are just as many Xs as Ys.
>
> On the other side, we have the part of Hume's principle that tells us when there is such a thing as the number of Xs:
>
> > (Universal countability) For any concept X, there is such an object as the number of Xs.

It is not entirely clear whether this is a claim that a positive free logic is required; after all, in a positive free logic, there would be no need to restrict (HP minus) to strip it of existential import, since it would already lack existential import. But Rumfitt's claim is similar to the claim that a positive free logic is required, in that he requires that there be an additional existence assumption.

So, what can be said in defence of negative free logic as a background logic for abstraction? Hale (2001a) writes:

> It would be ... wrong to suppose that the truth of instances of the left hand side [of an abstraction principle] is simply a matter of stipulation, if by that it is meant that their truth (and hence the existence of referents for their ingredient terms) is stipulated. All that is stipulated is the truth of a (universally quantified) biconditional. In general, this will leave entirely open the question whether terms of the type provided for by the left hand

> side have reference or not—and it will do so, regardless of whether the identity predicate is understood as signifying a strong or rather weak identity relation in Potter and Smiley's sense. There is therefore no good ground, for all we have seen so far, to insist that if an abstraction principle is to be the subject of legitimate stipulation, it must be existentially bowdlerised by deploying the weak identity relation in the way Potter and Smiley suggest. (p. 347)

And, considering the proposal that the identity on the left hand side of HP should be taken as in positive free logic, Hale and Wright (2003) write:

> The neo-Fregean will reply that it is not of the truth-conditions of identity contexts of that etiolated kind that he is proposing Hume's Principle as a stipulative explanation, but of contexts whose being true precisely will license straight-forward existential generalisation.(p. 260 n. 9).

Whereas in Hale and Wright (2009a), they simply write that

> it is sometimes proposed that '$t = t$' is to be understood so as to be true even if its ingredient term lacks reference. But we are under no pressure to accept such a view. (p. 464.n 15)

The thought seems to be that, with a choice between two readings of identity, the abstractionist is free to choose which one they will be stipulating the truth conditions of (by means of an abstraction principle). This choice is permissible since in neither case will the identity statement *itself* be stipulated—to do that in the case of the identity of negative free logic would clearly be question begging—but rather it is only the truth-*conditions* which are stipulated. Whether these obtain or not is not a matter of stipulation.

But this can not be quite the whole story. What reason is there that we can choose freely which of the identity statements we stipulate the truth conditions of? It seems that there are two options. One is that there is something special about *identity* contexts which means that they are suitable to be the contexts for which truth conditions may be stipulated. Or, there is not, and we are free to explain the meaning of a new singular term t (or range of singular terms, via an abstraction operator) by stipulating the truth conditions of whichever contexts involving t that we choose. In the first of these cases, an argument would then be required as to why it is the identity of negative, rather than positive, free logic which has a special status. But in the second case, since there is no longer anything special about identity, it seems that the abstractionist may be burdened with having to explain a host of *generalised* abstraction principles. These would be principles of the form:

$$\forall\alpha\forall\beta[\psi(\S\alpha, \S\beta) \leftrightarrow \phi(\alpha, \beta)] \tag{2.3}$$

where ϕ is a formula which does not make use of the § abstraction operator. Normal abstraction principles such as HP are clearly instances of this, where $\psi(\S\alpha, \S\beta)$ is just the identity '$\S\alpha = \S\beta$'.

If the second of these options is taken, then a response may be possible by accepting that certain generalised abstraction principles may be able to serve the same purpose as HP in implicitly defining new concepts. Obviously it would be undesirable if a consequence were that it would be possible to make use of any such generalised abstraction

principle, with any formula featuring on the left hand side. But this may be resisted; there must be some restriction on what abstraction principles proper are viable as implicit definitions, so as to rule out, for example, BLV (see chapters 3 and 4). These restrictions will no doubt be applicable to generalised abstraction principles as well, and would be expected to rule out those which may be most problematic.

However, such a response—which still leaves the abstractionist with any difficulties arising from generalised abstraction principles—may not be necessary. Instead, a response can be given more in line with the first option, by arguing directly in favour of negative free logic. That is, by arguing that genuine identity contexts permit existential generalisation. Then it will be the case that there is not really a choice between a strong and a weak reading of identity. There is instead simply a choice between identity (which will permit existential generalisation), and the complex formula as given in (2.2). Although this latter formula may behave in some ways like identity, it is not in fact identity, so may simply be disregarded.

Such a response requires two components. Firstly, something should be said in favour of the view that there is something special about identity which makes it suitable for the left hand side of abstraction principles. Then, an argument is needed to the effect that genuine identity contexts are as negative free logic says that they are. This latter argument can be given, I believe, by appealing to two components of the static abstractionist view, namely the context principle, and the assumption that the domain of quantification remains fixed.

I will only be brief on the first component of this argument, concerning why we should only take identity contexts as being suitable to appear on the left hand side of an abstraction principle. Strictly speaking, all that is required is that identity contexts *are* suitable, not the stronger claim that *only* such contexts are suitable. And, if abstractionism is to go anywhere at all, the first of these can not be denied. To deny that identity contexts are suitable for use in this way is simply to reject abstraction principles altogether, well before it comes to determining the correct background logic for them.

But something could perhaps also be said for the stronger claim. There are a number of reasons that one may think that it is identity contexts in particular which may play such a role. So, for example, we may be following Frege in the *Grundlagen* in demanding that 'if we are to use the symbol a to signify an object, we must have a criterion for deciding in all cases whether b is the same as a' (p. 73). Or, one may think, like Quine (1950), that criteria of identity are essential for fixing reference.

So, given the assumption that an abstraction principle may be used to give the truth conditions of an identity statement, what can be said about the second part of the argument? How might a negative free logic be defended directly? One way could be as follows (c.f. Burge, 1975): an atomic formula $\ulcorner Pt \urcorner$ serves to ascribe a property (that signified by the predicate P) to an object (that denoted by the singular term t). So, if t fails to refer, the atomic formula must be false; there is no object to which it can be ascribing a property. The same goes for relation symbols more generally. An atomic sentence $\ulcorner Pt_1 \dots t_n \urcorner$ asserts that a particular relation (that signified by P) holds between certain objects (those denoted by $t_1, \dots, t_n$). Hence the sentence can not be true if the terms involved do not refer. To put it another way: an atomic formula is true partly *because* the singular terms in it refer. Hence the characteristic rule of a negative free logic ($E!$-I) must be valid.

However, by making use of the context principle, an alternative route to the ($E!$-I) rule is available which turns the aforementioned explanation of the truth of atomic sentences on its head. Roughly: it is not because the terms involved in an atomic sentence refer that that sentence is true, but rather the other way around. Singular terms refer because they appear in true atomic sentences; to appear in a true atomic sentence just is what it is for a singular term to refer.

But how does this follow from the context principle? What the context principle tells us is that claims about the semantics (e.g. reference) of singular terms are to be explained in terms of claims about the semantics (e.g. truth-value) of whole sentences. Similarly, questions about the semantics of singular terms are to be answered in terms of the semantics of whole sentences. Thus, in asking if and why a singular term t refers, the correct answer will be of the form 'because this sentence/these sentences containing t is/are true' (or something similar). This is essentially the part of the context principle which Wright (1983) calls the syntactic priority thesis. It is that 'possession of reference is imposed on a singular term by its occurrence in true statements of an appropriate type' (p.53).

There remains then the question of what kind of sentence is required (for clearly, it can not be the truth of *any* sentence involving t which suffices). A natural answer to this question, and one which Hale and Wright appear to be committed to,[3] is that it is the most basic kind of sentences—atomic sentences—which play such a role.[4]

But there is another element which is required in order to defend the ($E!$-I) rule, which is supplied by the assumption of a fixed domain. The need for this assumption can be revealed by examining more closely the ($E!$-I) rule.

What the conclusion of this rule—$\exists x(x = t)$—says is that the term t refers to an object *in the range of the quantifier* $\exists$. But there is *prima facie* space for there to be gap between a term referring, and it referring to an object in a particular domain of quantification. Of course, this gap may be closed if, for example, it is claimed that there is an absolutely unrestricted domain of quantification, as the proponent of static abstraction proposes. But, since this assumption will be questioned in the latter part of this thesis, it is important to note that the assumption is needed.

Or, to put things another way, suppose that an abstraction principle allows us to prove that $\S\alpha = \S\alpha$, and so that '$\S\alpha$' refers. Now, what object this term refers to will depend partly on the first-order domain of quantification which was in play in the context in which the abstraction principle was laid down. This is because an abstraction principle will (typically) involve first-order quantification on its right hand side. So, the

[3] See, for example, Hale and Wright (2009b, p.197).

[4] There is of course room to reject this. Another plausible alternative might be that sentences of the form $\ulcorner \exists x(x = t) \urcorner$ play that role. Thus the issue of whether a singular term t refers should be answered in terms of whether the sentence $\ulcorner \exists x(x = t) \urcorner$ is true.

This would put abstractionists in a difficult position were it the case. For then, rather than having to explain the truth of atomic sentences involving number terms, and then the truth of quantified sentences involving those terms by means of the (E!-I) rule, the task would become to explain the truth of quantified sentences involving number terms directly. But, since my present aim is to show how a negative free logic might be motivated using a form of the context principle which Hale and Wright are already committed to, I will not take this issue any further.

It is worth noting, however, that the use I make of the context principle in chapter 6 is not affected by a similar response. For, in that chapter, I aim to justify the referentiality of abstract terms by means of directly explaining the truth-values of *quantified* sentences involving them (rather than just explaining the truth value of identity statements involving them).

inference licensed by ($E!$-I) is that if an abstract term appears in an atomic sentence, then it will refer to an object within the very same domain of quantification by which its reference is (partly) determined. And this is, more or less, just what the assumption of a fixed domain amounts to. Again, this might be justified by appeal to an absolutely unrestricted domain.

These two components then provide a recipe for the justification of ($E!$-I): Given an atomic formula $A(t)$, t refers by the context principle. Then, by the assumption that the domain is fixed, t will refer to an object which lies in the present domain of quantification. Thus, we have the conclusion of the rule, that $\exists x(x = t)$.

2.3 Free logic and restricted abstraction principles

In the context of an abstraction principle such as HP, the adoption of a free logic—as long as it is a negative free logic—ultimately makes little or no difference (as Hale and Wright are eager to point out in, e.g. Hale and Wright (2009a, p. 464)). Since, for any term t, it can be proved from HP that $E!t$, the logic will behave exactly as a non-free logic would. But there is a certain class of abstraction principles for which a negative free logic can be put to a good use. These are *restricted* abstraction principles. This useful application provides a further reason for adopting a free logic as the background logic for abstraction.

For an abstraction principle such as HP, it is desirable that for any concept, there is a resulting well defined abstract.[5] This would be a consequence of taking arithmetic, as Frege did, as being essentially universally applicable; no matter what we choose to concern ourselves with, we can count it. For other abstraction operators that we may wish to introduce, however, this is not the case. For example, we may wish to consider the direction operator as applying not just to lines, but to any object whatsoever. In that case, objects which are not lines would simply not have a direction associated with them. Or, given what we know about the inconsistency of naïve set theory, one may wish to introduce a 'set of' operator in such a way such that not every concept has a corresponding set. Another use might be to take care of certain 'degenerate' mathematical terms—if they were to be introduced by abstraction—such as fractions whose denominator is zero.

Since in these cases the desired outcome is that the abstraction operator denotes what is, in effect, a (merely) *partial* function, free logic seems to be an appropriate background in which to pursue such restricted abstraction principles. And indeed, I shall argue that these kinds of restrictions can most naturally be accommodated in a free logical setting. But before doing so, it will be useful to review a few ways in which restricted abstraction principles have been pursued in a setting of classical non-free logic. My concern is with restricted abstraction principles in general, but, for the sake of concreteness, and because they are by far the most discussed in the literature, I will only discuss restricted versions of Basic Law V. But everything I say should apply also to

[5] Actually, this is not quite the case even for HP. There are a number of concepts for various reasons that we may not wish to attach a number to, such as vague concepts (e.g. 'tall man'), non-sortal concepts (e.g. 'red'), or perhaps *indefinitely extensible* concepts (e.g. 'self-identical') (see, e.g. Wright, 2001c, p. 207–8). The method of restriction which I will discuss here will also apply to such as these, but I have in mind in particular restrictions which even rule out definite, sortal concepts.

restricted abstraction principles more generally.

2.3.1 New V

The principal aim of restricted versions of Basic Law V has been to rule out problematic sets such as the Russell class, or a universal set. The first proposal in this direction was by Boolos (1989) who proposed that BLV be restricted to concepts which are 'small'. A small concept is one which is not in a one-to-one correspondence with everything that there is. Or, equivalently, it is a concept F such that there is not an injection from everything that there is into F. This may be given a formal definition along similar lines to the definition of equinumerosity:

$$\text{Small}(F) \stackrel{\text{df}}{=} \neg\exists R[\forall x\exists y\forall z((Fz \wedge Rxz) \leftrightarrow z = y)] \tag{2.4}$$

Perhaps more notable than the specific restriction, however, is the way in which Boolos suggests the restriction be made. He proposes the following abstraction principle, known as *New V*:

$$\forall F\forall G[\varepsilon F = \varepsilon G \leftrightarrow ((\text{Small}(F) \vee \text{Small}(G)) \rightarrow \forall x(Fx \leftrightarrow Gx))] \tag{NV}$$

But, of course, there is nothing specific about the restriction to smallness that makes this possible. We may replace 'Small' with an arbitrary formula $\phi(F)$, where the concepts which are not ϕ are ruled to be bad somehow (although it would be reasonable to demand that ϕ does not contain any occurrence of ε). So, more generally, we have:

$$\forall F\forall G[\varepsilon F = \varepsilon G \leftrightarrow ((\phi(F) \vee \phi(G)) \rightarrow \forall x(Fx \leftrightarrow Gx))] \tag{NV$_\phi$}$$

Some alternatives to smallness which have been suggested are *double*-smallness (roughly, smaller than some concept which is smaller than the universe) (e.g. Hale, 2000a, 2005) and *definiteness* (e.g. Shapiro, 2003; Shapiro and Wright, 2006).

What is the effect of abstraction principles such as NV$_\phi$? First, it should be noted that it is simple to prove that the formula on the right hand side will be an equivalence relation. As a consequence, it will be possible to prove $\forall F\exists y(y = \varepsilon F)$ in exactly the same way that it is possible to prove $\forall F\exists x(x = NF)$ from HP, and so NV$_\phi$ will be consistent with non-free logic. So, every concept—including problematic concepts such as a universal concept, or one which defines the Russell class—will have a corresponding 'set'. Instead, however, all of these problematic concepts will map to the *same* abstract. That is, we will be able to prove $\forall F\forall G(\phi(F) \wedge \phi(G) \rightarrow \varepsilon F = \varepsilon G)$. This results in a system which bares some resemblance to one of the precursors of free logic—the 'chosen object' theory which is sometimes attributed to Frege (1948, pp.70–71) and Carnap (1947, pp.35–6). This approach to empty singular terms first identifies a particular *null* object. Then terms which would otherwise be taken to be non-denoting would be stipulated to have the null object as their referent.

Now, NV$_\phi$ has what seem to be some undesirable (but by no means insurmountable) consequences. So, for example, (assuming that a restriction rules out the same kind of sets that, say, ZF set theory does), it identifies the set of all sets, the set of all objects whatsoever, the set of cardinals, the set of ordinals and so on. Moreover, if we are to define $\in$ in the natural way (as in (1.4)), we will encounter further strange consequences.

For example, writing U for $\{x : x = x\}$ and O for $\{x : x \text{ is an ordinal}\}$, we have such consequences as $U \in O$, $\forall x(x \in O)$ and so on. Even if we restrict attention to objects which are not the null object, we get strange results, since NV permits us to form new sets out of the null object.

As mentioned, these consequences are by no means insurmountable. Boolos deals with them by restricting quantifiers to only those abstracts which are (a) not the null object, and (b) are pure sets, in the sense that they have no non-sets (including the null object) in their transitive closure. But the presence of these artifacts of the approach and the need to avoid them in various ways at least suggests that we might look for a more natural way of restricting abstraction principles.

2.3.2 Direct restriction

Another way of restricting abstraction principles is simply to explicitly restrict the outermost quantifiers. So, if we continue to just consider restrictions on BLV by a condition ϕ, we have:

$$(\mathrm{RV}_\phi) \qquad \forall F \forall G[\phi(F) \wedge \phi(G) \rightarrow (\varepsilon F = \varepsilon G \leftrightarrow \forall x(Fx \leftrightarrow Gx))]$$

A restriction of this form is considered by Hale (2000a).

I will not say much about this form of restriction yet. It is similar in many ways to the method of restriction that I will propose shortly. There are a number of notable differences between RV_ϕ and NV_ϕ which are worth pointing out at this moment. Rather than mapping bad concepts onto the same null object, RV_ϕ does not say *anything* about them; it does not say what objects they refer to, nor does it say that they fail to refer. It will also transpire that it is not as powerful as NV_ϕ, at least in the case where ϕ is smallness.

One objection that may be levelled against RV_ϕ is that, under a strict criterion of what counts as an abstraction principle, it will not be counted as such, since it does not have the requisite form of being a universally quantified biconditional. Of course, such a criterion could readily be amended to permit restricted abstraction principles. But, at the very least, refusing to count principles of the form of RV_ϕ as abstraction principles proper will allow for a certain amount of simplification later on (especially in chapter 3), since then we only need consider abstraction principles of a single form.

2.3.3 Using free logic to restrict abstraction principles

As I pointed out in my discussion of NV_ϕ, the right hand side of such principles is an equivalence relation, and this is what makes it compatible with a non-free logic. In fact, it is very often required by various writers that an abstraction principle must have an equivalence relation on its right hand side. The reasoning for this is simple; the right hand side is being stipulated to be equivalent to an identity, and must therefore inherit the properties of reflexivity, symmetry and transitivity from those of identity; to do otherwise would simply cause a contradiction.

However, in the context of a negative free logic, it is *not* required that the right hand side be reflexive. An argument for its reflexivity might be given much as the following: from the reflexivity of identity, $\S F = \S F$. So, reasoning left to right across the abstraction principle, we have $F \sim F$. Thus, by universal generalisation, $\forall F(F \sim F)$. But the premise

of this argument is not a general truth in a negative free logic. $\S F = \S F$ will be true only if an additional assumption is granted that $\S F$ exists, which is just what is missing in a free logic. Identity in a negative free logic is reflexive in one sense, namely that $\forall x(x = x)$ holds in it. But the stronger claim that, for all terms t, $\ulcorner t = t \urcorner$ is true, does not hold.

And in many of the natural cases in which we might expect abstraction operators to behave as merely partial functions, we would also expect the abstraction relation to be non-reflexive. This is often noted in regards to the direction equivalence. So, for example, Wright (2001a) notes that

> if the failure of parallelism between my hat and my shoe is down to the unsuitability of of either object to be parallel to anything, then by the same token they are not *self*-parellel, and DE provides no incentive to regard either as having a direction at all. (p. 314)

And this thought transfers very naturally to the situation in which we wish to place an explicit restriction on an abstraction principle, as we do in the case of BLV. We want an abstraction relation $\sim$ such that, for 'bad' concepts F, $F \not\sim F$. When the aim is to restrict an already given abstraction principle—say BLV—to concepts which satisfy some condition ϕ, a way of doing this is easily available. We can have, as such an abstraction principle:

$$(\mathrm{FLV}_\phi) \qquad \forall F \forall G[\varepsilon F = \varepsilon G \leftrightarrow \phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)]$$

It is then simple to check that, although the relation on the right hand side is transitive and symmetric, it is not reflexive. In particular, in the case where $\neg\phi(F)$, the relation will not hold between F and itself.

As a consequence, there are a couple of things that can be immediately noted. Like RV_ϕ and unlike NV_ϕ, this does not a assign an object to non-ϕ concepts. But unlike RV_ϕ, it does say something about the non-ϕ concepts, namely that they do not define sets (the converse, that ϕ concepts do define sets, is also readily provable). That is, we can prove the following proposition:

Proposition 2.1. *$\mathrm{FLV}_\phi \vdash \forall F(\phi(F) \leftrightarrow E!\varepsilon F)$*

Proof. Suppose that $\phi(F)$. Then clearly also $\phi(F) \wedge \phi(F) \wedge \forall x(Fx \leftrightarrow Fx)$. Hence, by reasoning right to left across FLV_ϕ, $\varepsilon F = \varepsilon F$. Since this is an atomic formula, we may derive $E!\varepsilon F$ by ($E!$-I).

For the converse, suppose $E!\varepsilon F$. Then, by (=-I) for negative free logic, $\varepsilon F = \varepsilon F$. So, reasoning left to right across FLV_ϕ, $\phi(F)$ as required. □

In addition, in contrast to RV_ϕ, FLV_ϕ will count as an abstraction principle according to the strict condition that it be of some specific form.

2.3.4 The relationship between the restrictions

It is possible to do better than these brief remarks concerning the differences between the three approaches, and to map out precisely what the relationships between them are. The comparison between RV_ϕ and FLV_ϕ is by far the simplest, so I will cover that first.

As already mentioned, one thing that FLV_ϕ does which RV_ϕ does not do is prove that non-ϕ concepts do not have corresponding abstracts. And, in fact, it turns out that that is the only difference between them. In particular, we have the following theorem:

Proposition 2.2. *$\text{FLV}_\phi \equiv [\text{RV}_\phi \wedge \forall F(\neg\phi(F) \rightarrow \neg E!\varepsilon F)]$ (where '$\equiv$' denotes logical equivalence).*

Proof. For the left to right direction: Assume FLV_ϕ. We have already shown that $\forall F(\neg\phi(F) \rightarrow \neg E!\varepsilon F)$, so it remains to show RV_ϕ. Consider concepts F and G. If $\neg\phi(F)$ or $\neg\phi(G)$, then there is nothing to prove (RV_ϕ would be vacuous). So suppose that $\phi(F) \wedge \phi(G)$. The right to left direction of BLV embedded within RV_ϕ can then easily be proved using the right to left direction of FLV_ϕ together with the assumption that $\phi(F) \wedge \phi(G)$. The left to right direction of BLV embedded within RV_ϕ follows directly (eliminating unneeded conjunctions) from the left to right direction of FLV_ϕ.

For the right to left direction: Assume RV_ϕ and $\forall F(\neg\phi(F) \rightarrow \neg E!\varepsilon F)$. For the left to right direction of FLV_ϕ, suppose F and G are concepts such that $\varepsilon F = \varepsilon G$. By (E!-I), we have that $E!\varepsilon F$ and $E!\varepsilon G$, and hence by our second assumption, that $\phi(F)$ and $\phi(G)$. We can thus apply RV_ϕ and so prove that $\forall x(Fx \leftrightarrow Gx)$, which gives us the right hand side of FLV_ϕ as required. For the right to left direction of FLV_ϕ, suppose that F and G are concepts such that $\phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)$. Since $\phi(F)$ and $\phi(G)$, we can thus apply RV_ϕ, and deduce that $\varepsilon F = \varepsilon G$, as required. □

The relationship between FLV_ϕ and NV_ϕ is not so simple. In contrast to the relationship between FLV_ϕ and RV_ϕ, there can be no simple relationship of entailment or equivalence between any theory which extends FLV_ϕ and any theory which extends NV_ϕ. The reason is that the two theories are formally inconsistent. FLV_ϕ entails $\exists F\forall x(x \neq \varepsilon F)$, whereas NV_ϕ entails its negation, $\forall F\exists x(x = \varepsilon F)$ (whether in the context of a non-free logic or a negative free logic).

But this is perhaps to be expected since, in each case, the abstraction operator seems intuitively to denote something different. In the case of FLV_ϕ, it is straightforwardly a 'set of' operator, whereas in the case of NV_ϕ, the abstract terms denote a wider class of entities—Boolos calls them *subtensions*—which include sets proper, the null object and any sets formed out of the null object. To make this difference clearer, it will be helpful to make use of a different symbol in each case. I will use ε_1 for the 'set of' operator of FLV_ϕ, and ε_2 for the 'subtension of' operator of NV_ϕ.

Since the symbols involved in each abstraction principle intuitively have different meanings, we might hope instead to find a relationship of *interpretability* between them, by giving a definition of ε_1 is terms of ε_2 and vice-versa. One half of this is to some extent fulfilled by Boolos (1989), since it is obviously an aim of his to interpret some set theory within the theory of subtensions given by New V.

But again, things will not be quite so simple. The reason is that NV_ϕ is strictly stronger than FLV_ϕ. Consider just the case where ϕ is smallness. Boolos shows that in this case, NV_ϕ interprets second-order arithmetic, and so only has infinite models. By contrast, FLV_ϕ by itself will have finite models. We can, for example, give a model of FLV_ϕ with a single element in its domain as follows: Let $\mathcal{M} = \langle D, \varepsilon^{\mathcal{M}}\rangle$, where $D = \{a\}$ and $\varepsilon^{\mathcal{M}} : \mathcal{P}(D) \rightarrow D$ is given by letting $\varepsilon^{\mathcal{M}}(\varnothing) = a$ and $\varepsilon^{\mathcal{M}}(\{a\})$ be undefined. Then it is simple to see that this satisfies FLV_ϕ; it assigns a 'set' to the only small concept, and

assigns none to the only large concept. This shows that FLV_ϕ cannot interpret NV_ϕ, for in that case it too would only have infinite models.

But a suitable strengthening of FLV_ϕ can be achieved—so as to achieve a mutual interpretability result—by noting how it is that NV_ϕ gets its strength. A crucial step in Boolos' derivation of some set-theoretic axioms from (NV) is a kind of bootstrapping process which makes essential use of the null object. Firstly, since there must be at least one abstract (since every abstract term refers), the empty concept $\lambda x(x \neq x)$ is guaranteed to be small. Moreover, the universal concept $\lambda x(x = x)$ is guaranteed to be large and hence their respective subtensions must be distinct. So, there must be at least two objects, and thus singleton concepts (e.g. $\lambda x(x = \varepsilon\lambda x(x \neq x))$) will be small. Then we can infer the existence of at least three objects and so on.

Without a null object, however, this kind of reasoning will not get off the ground in the case FLV_ϕ. But, all that is required instead is merely that there be at least one *urelement*, that is, an object which is not itself a set.[6] So, we can add to FLV_ϕ a sentence which states just that. It is also convenient to give such an urelement a name. So, consider an expansion of the language of FLV_ϕ by an individual constant c, and let ψ be the sentence $\exists x(x = c \wedge \forall F(x \neq \varepsilon_1 F))$. Then, we have the following relationship:

Theorem 2.3. *NV_ϕ and $\mathrm{FLV}_\phi + \psi$ are mutually interpretable.*

In order to show this result, it is necessary to provide translations between the two languages, by defining the language from one abstraction principle in the language of the other. Then, it will need to be proved that each of NV_ϕ and FLV_ϕ proves the translation of the other.

First, consider interpreting NV_ϕ in FLV_ϕ. In order to do this, we will need to give a definition of ε_2 in terms of ε_1 and c. This definition will be a formula $\theta(F, x)$ such that

$$\mathrm{FLV}_\phi + \psi \vdash \forall F \exists! x \theta(F, x)$$

If we write $\mathcal{L}_{NV}$ for the language of NV_ϕ (i.e. with sole non-logical symbol ε_2), and $\mathcal{L}_{FLV}$ for the language of FLV_ϕ (i.e. with non-logical symbols ε_2 and c), this definition will then induce a translation $\tau : \mathcal{L}_{NV} \to \mathcal{L}_{FLV}$, by treating ε_2 as a definite description $(\imath x)(\theta(F, x))$.[7]

A suitable definition is the following:

$$\theta(F, x) \stackrel{\mathrm{df}}{=} (\phi(F) \wedge x = \varepsilon F) \vee (\neg\phi(F) \wedge x = c) \tag{2.5}$$

We can then prove the required uniqueness and existence claim:

Lemma 2.4. *$\mathrm{FLV}_\phi + \psi \vdash \forall F \exists! x \theta(F, x)$*

Proof. For existence: if $\phi(F)$, then $E!\varepsilon F$ (by Proposition 2.1) and $\theta(x, \varepsilon F)$ (by the definition of θ). If $\neg\phi(F)$, then $\theta(F, c)$ (by the definition of θ), and $E!c$ (by ψ). In either case, we thus have that $\exists x \theta(F, x)$.

For uniqueness, suppose $\theta(F, x)$ and $\theta(F, y)$. If $\phi(F)$, then $x = \varepsilon F = y$, so $x = y$. If $\neg\phi(F)$, then $x = c = y$, so $x = y$. □

[6] It would also do to assume that there are at least two objects of any kind.

[7] Usually, an interpretation will also feature a domain restriction. But a domain restriction will not be required to prove the result.

Then we need to show that the formula resulting from translating NV_ϕ can be proved using FLV_ϕ and ψ:

Lemma 2.5. *$FLV_\phi + \psi \vdash \tau(\mathrm{NV}_\phi)$*

Proof. First note that:

$$\begin{aligned}\tau(NV_\phi) &= \tau(\forall F\forall G(\varepsilon_2 F = \varepsilon_2 G \leftrightarrow (\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)))\\ &= \forall F\forall G(\tau(\varepsilon_2 F = \varepsilon_2 G) \leftrightarrow (\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)))\\ &= \forall F\forall G[\forall x\forall y(\theta(F,x) \wedge \theta(G,y) \rightarrow x = y) \leftrightarrow\\ &\qquad (\phi(F) \vee \phi(G)) \rightarrow \forall x(Fx \leftrightarrow Gx)]\end{aligned}$$

Now, we can prove the left to right direction of $\tau(\mathrm{NV}_\phi)$ as follows: Assume $\forall x\forall y(\theta(F,x) \wedge \theta(G,y) \rightarrow x = y)$ with the aim of showing that $\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)$. So, suppose that $\phi(F) \vee \phi(G)$. Without loss of generality, suppose that $\phi(F)$. Then, by the definition of θ, $\theta(F, \varepsilon_1 F)$. But then $\phi(G)$ and hence $\theta(G, \varepsilon_1 G)$; for if not, then $\theta(G, c)$, and so $\varepsilon_1 F = c$, contradicting ψ. Hence $\varepsilon_1 F = \varepsilon_1 G$. So, by the left to right direction of FLV_ϕ, we have $\forall x(Fx \leftrightarrow Gx)$, and so $\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)$ as required.

For the right to left direction, assume that $\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)$, with the aim of showing that $\forall x\forall y(\theta(F,x) \wedge \theta(G,y) \rightarrow x = y)$. So, consider arbitrary x and y such that $\theta(F,x)$ and $\theta(G,y)$. If either $\phi(F)$ or $\phi(G)$, then, by our assumption, $\forall x(Fx \leftrightarrow Gx)$, and thus also $\phi(G)$ (assuming that ϕ is stated in a purely extensional language, so that it is a congruence with respect to coextensiveness). Thus, by the definition of θ, $x = \varepsilon_1 F$ and $y = \varepsilon_1 G$. But from the right to left direction of FLV_ϕ, we have $\varepsilon_1 F = \varepsilon_1 G$, and hence $x = y$ as required.

Suppose instead that both $\neg\phi(F)$ and $\neg\phi(G)$. Then, by the definition of θ, $x = c$ and $y = c$, and thus $x = y$ as required. □

We can also prove corresponding lemmas for the converse direction. What will be required is a formula $\theta(F,x)$ which defines ε_1. But, since ε_1 is merely a partial function, it will not be required that it satisfies both existence and uniqueness conditions, but merely uniqueness conditions. That is, it is required that:

$$\mathrm{NV}_\phi \vdash \forall F\forall x\forall y(\theta(F,x) \wedge \theta(F,y) \rightarrow x = y)$$

In addition, it will be required to provide a definition for the constant c. This will be a formula $\gamma(x)$ for which

$$\mathrm{NV}_\phi \vdash \exists! x\gamma(x)$$

As in the previous case, these definitions will induce a translation $\tau : \mathcal{L}_{FLV} \rightarrow \mathcal{L}_{NV}$ by treating $\varepsilon_1 F$ and c as definite descriptions $(\imath x)(\theta(F,x))$ and $(\imath x)(\gamma(x))$.

Suitable definitions are then the following:

$$\theta(F,x) \stackrel{\mathrm{df}}{=} \phi(F) \wedge x = \varepsilon_2 F \tag{2.6}$$

$$\gamma(x) \stackrel{\mathrm{df}}{=} \exists F(\neg\phi(F) \wedge x = \varepsilon_2 F) \tag{2.7}$$

We can then prove the appropriate properties for θ and γ:

Lemma 2.6.

$$NV_\phi \vdash \forall F \forall x \forall y (\theta(F,x) \wedge \theta(F,y) \rightarrow x = y)$$
$$NV_\phi \vdash \exists x (\gamma(x) \wedge \forall y (\gamma(y) \rightarrow y = x))$$

Proof. For the first, suppose that $\theta(F,x)$ and $\theta(F,y)$. Then $\phi(F)$, $x = \varepsilon_2 F$ and $y = \varepsilon_2 F$. Thus $x = y$ as required.

For the second: To show existence, let F be the Russell concept . Then, on pain of contradiction, $\neg\phi(F)$. Since we have $\exists x(x = \varepsilon_2 F)$, we thus have $\exists x(\gamma(x))$ as required. To show uniqueness, suppose that $\gamma(x)$ and $\gamma(y)$, so that there are F and G such that $\neg\phi(F)$, $\neg\phi(G)$, $x = \varepsilon_2 F$ and $y = \varepsilon_2 G$. But, by NV_ϕ, we can deduce $\varepsilon_2 F = \varepsilon_2 G$, and thus $x = y$ as required. □

It remains again to show that the resulting translation results in an interpretation. That is, we need to show the following:

Lemma 2.7.

$$NV_\phi \vdash \tau(FLV_\phi) \wedge \tau(\psi)$$

Proof. First, note that:

$$\begin{aligned}
\tau(\mathrm{FLV}_\phi) &= \tau(\forall F \forall G[\varepsilon F = \varepsilon G \leftrightarrow \phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)]) \\
&= \forall F \forall G[\tau(\varepsilon F = \varepsilon G) \leftrightarrow \phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)] \\
&= \forall F \forall G[\forall x \forall y(\phi(F) \wedge x = \varepsilon_2 F \wedge \phi(G) \wedge y = \varepsilon_2 G \rightarrow x = y) \leftrightarrow \\
&\qquad \phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)] \\
&\equiv \forall F \forall G[(\phi(F) \wedge \phi(G) \wedge \varepsilon_2 F = \varepsilon_2 G) \wedge y = \varepsilon_1 G \rightarrow x = y) \leftrightarrow \\
&\qquad \phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)]
\end{aligned}$$

and that:

$$\begin{aligned}
\tau(\psi) &= \tau(\exists x(x = c \wedge \forall F(x \neq \varepsilon_1 F))) \\
&= \exists x(\gamma(x) \wedge \forall F(\neg\theta(F,x))) \\
&= \exists x(\exists F(\neg\phi(F) \wedge x = \varepsilon_2 F) \wedge \forall F(\neg\phi(F) \rightarrow x = \varepsilon_2 F))
\end{aligned}$$

Then we can prove the lemma. First, consider proving $\tau(\psi)$ from NV_ϕ. Since from NV_ϕ it follows that all non-ϕ concepts have the same subtension, the two conjuncts of $\tau(\psi)$ are equivalent, so we just need to show that $\exists x \exists F(\neg\phi(F) \wedge x = \varepsilon_2 F)$. But we have already shown just that, in lemma 2.6.

Now consider $\tau(\mathrm{FLV}_\phi)$. First, we will show the left to right direction. Consider arbitrary concepts F and G and assume that $\phi(F) \wedge \phi(G) \wedge \varepsilon_2 F = \varepsilon_2 G$, with the aim of showing $\phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)$. The first two conjuncts obviously follow immediately and trivially. Since $\varepsilon_2 F = \varepsilon_2 G$, we may reason left to right across NV_ϕ to obtain $\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)$. But since we have $\phi(F)$ and $\phi(G)$, a simple application of *modus ponens* gives us the required result that $\forall x(Fx \leftrightarrow Gx)$.

For the right to left direction: consider arbitrary concepts F and G and assume that $\phi(F) \wedge \phi(G) \wedge \forall x(Fx \leftrightarrow Gx)$, with the aim of showing $\phi(G) \wedge \phi(G) \wedge \varepsilon_2 F = \varepsilon_2 G$. Again, the first two conjuncts are trivial. Since $\forall x(Fx \leftrightarrow Gx)$, we have $\phi(F) \vee \phi(G) \rightarrow \forall x(Fx \leftrightarrow Gx)$, so we may reason right to left across NV_ϕ to obtain $\varepsilon_2 F = \varepsilon_2 G$ as required. □

It simply remains to bring lemmas 2.4–2.7 together in a proof of the theorem:

Proof (of theorem 2.3). What is needed is that for any ϕ in $\mathcal{L}_{NV}$, $\mathrm{FLV}_\phi + \psi \vdash \tau(\phi)$, and similarly for the converse direction. This will follow from lemmas 2.5 and 2.7, together with the following general claim about translations of the appropriate sort:

$$\text{If } \phi \vdash \psi \text{ then } \tau(\phi) \vdash \tau(\psi)$$

But this is well know to hold in the case of definitions of the sort given by τ, which make use of formulas with appropriate existence and uniqueness properties (the proof is a fairly simple but tedious one by induction on formula complexity). And this is just what lemmas 2.4 and 2.6 show. □

2.4 The implications of these restrictions

What conclusions can we draw from the foregoing results? For one thing, they show that this way of restricting abstraction principles will never get us more in the way of mathematics derivable from abstraction principles, since in every case, the New V-style restriction will be stronger. What then, is the point of even considering such abstraction principles? I do not wish to claim that there are any particularly deep conclusions to draw, but I think that there are a couple of more tentative conclusions which can be drawn.

Firstly, I claim that free logic provides the most natural approach to restricted abstraction principles, or to abstraction principles where it would be expected that not every abstract term has a reference. In contrast to what RV_ϕ allows, we frequently do want to say something about the concepts or objects which do not define an abstract, namely that they do not define an abstract. Rather than keeping silent about, say, the direction of shoes and hats, as the equivalent of RV_ϕ for directions would have us, we want to say that hats and shoes *do not have directions*. And I hope that it should be clear that abstraction principles along the lines of FLV_ϕ are far more natural in their consequences than those along the lines of NV_ϕ. Abstraction principles along the lines of NV_ϕ require that we introduce a seemingly superfluous object to be, say, the direction of hats and shoes. Moreover, in certain cases, such as set theory, it will not only require the introduction of just one strange object, but a whole hierarchy of objects which involve this object in some way. Worse, it seems that we may even have to have distinct null objects for every abstraction principle, so that there is a null-set, null-number, null-direction and so on.

The free logical approach to restricted abstraction principles, by contrast, suffers from none of these problems. It permits us to say of hats and shoes that they do not have directions, for example. Even better, in certain cases, it does not even require us to make modifications to abstraction principles. Rather, it allows us to consider the relations on the right hand side as something less than equivalence relations—which a non-free logic would prevent—so that non-lines are not self-parallel, or so that non-sortal concepts are not self-equinumerous and so on.

Secondly, and with regards to abstraction principles for set theory in particular, I wish to claim that the weakness of FLV_ϕ compared with NV_ϕ reveals something of a sleight of hand involved in the use of the latter. New V can be seen as a way of formalising

the limitation of size conception of set within an abstractionist system, and of showing that the limitation of size conception is thus enough to recover a reasonable portion of set theory. But, if I am right that FLV_ϕ is really the most natural way of restricting BLV to ϕ concepts, then the result of formalising the limitation of size conception in an abstraction setting is really the instance of FLV_ϕ in which ϕ is smallness.

But, in that case, it is not true that the limitation of size conception is itself sufficient for deriving a substantial amount of set theory. Rather, it is the limitation of size conception *together* with an assumption to the effect that there is at least one non-set (or that there are at least two objects). New V sneaks this additional assumption in by having the null object provide such an object.

Chapter 3

The bad company problem, and how to think about it

3.1 The bad company problem

One of the main problems facing abstractionism, particularly the standard static form of abstractionism propounded by Hale and Wright is the *bad company problem*.[1] In its simplest form, the objection is that the inconsistency in BLV infects the whole abstractionist programme, rendering abstraction principles in general unsuitable as implicit definitions of abstraction operators.

So, for example, Dummett (1991) writes:

> If the context principle, as expounded by Wright, is enough to validate the 'contextual' method of introducing the cardinality operator, it must also be enough to validate a similar means of introducing the abstraction operator [for classes]. ...Frege's method of introducing the abstraction operator [for classes]—that is, of introducing value-ranges—was, notoriously, not in order. (p. 188)

The reasoning is then something like the following: The suitability of an abstraction principle to serve as a definition is in virtue of its form. But then, BLV would be suitable as an implicit definition, in virtue of being an abstraction principle. But BLV is inconsistent, and thus not suitable as an implicit definition (since it is not capable of being true). Hence abstraction principles in general are not suitable as implicit definitions.

This line of argument is, however, too quick, as Wright (1997) points out. It is not unreasonable to think that a method of definition may require restrictions on which specific attempts to give a definition succeed. Wright gives as an example the (presumably perfectly acceptable) method of defining a predicate by specifying application conditions. But this will inevitably have instances which lead to inconsistency. So, for example, suppose we define the predicate '(is) heterological' by specifying that it applies to predicates which do not apply to themselves. As is well known, this leads

[1] I shall argue later that the expansionist approach to abstraction does not suffer from the bad company problem.

immediately to a contradiction (since 'heterological' will be heterological just in case it is not heterological), yet this does not mean that we should abandon the specification of application conditions as a way of defining predicates.

Another (perhaps less uncontroversial) example might concern logical connectives. It might be thought that logical connectives have the meanings that they do by virtue of their introduction and elimination rules. But Prior's (1960) example of the 'Tonk' connective[2] shows that not any old introduction and elimination rules will do. So, proponents of this inferentialist view have suggested criteria that introduction and elimination rules must satisfy so as to be in good order.

So then, it is reasonable to allow that some additional constraints may be placed on abstraction principles before they may be considered acceptable, and that these constraints will rule out abstraction principles such as BLV. The bad company problem then becomes, not an outright *objection* to abstractionism, but rather a *challenge*; the challenge being to find appropriate restrictions on which abstraction principles are to count as acceptable.

It will not do, however, simply to propose that an abstraction principle shall be considered acceptable if and only if it is consistent. For while this would rule out BLV and other inconsistent abstraction principles, another problem remains. That problem is that there are abstraction principles which, although individually consistent, are mutually inconsistent.[3]

The standard example given of such a clash is between HP—which entails that the universe is infinite—and some abstraction principle which entails that the universe is finite. Such a principle is provided by the following 'Nuisance Principle' (Wright, 1997, p.290):

(NP) $\quad \nu(F) = \nu(G) \leftrightarrow$ there are finitely many x s.t. $(Fx \wedge \neg Gx) \vee (Gx \wedge \neg Fx)$

which has models of any finite cardinality, but no infinite models. Thus HP and NP are jointly unsatisfiable.[4]

A word of caution is perhaps appropriate here. Although it is commonly claimed that this shows that HP and NP are *inconsistent*, not quite as much has been shown. Whilst

[2]Tonk has the same introduction rules as 'or', and the same elimination rules as 'and'. I.e.:

$$\text{(Tonk-I}_1\text{)}\ \frac{A}{A\ \text{Tonk}\ B} \qquad \text{(Tonk-I}_2\text{)}\ \frac{A}{B\ \text{Tonk}\ A}$$

and

$$\text{(Tonk-E}_1\text{)}\ \frac{A\ \text{Tonk}\ B}{A} \qquad \text{(Tonk-E}_2\text{)}\ \frac{A\ \text{Tonk}\ B}{B}$$

Then it is simple to see that we can derive any conclusion B from any premise A: First derive 'A Tonk B' from (Tonk-I), then use (Tonk-E) to derive B.

[3]It is worth pointing out that this is also the case with predicates being defined by means of application conditions. For a Curry-style paradox can be constructed as follows. Say that a predicate P is ϕ-heterological iff P does not apply to itself, or ϕ. Then ϕ will be a consequence of this definition. So, if ϕ and ψ are such that neither is inconsistent, but their conjunction is, the definitions of ϕ-heterologicality and ψ-heterologicality will both be consistent, but will be mutually inconsistent.

[4]NP was actually the second of such restrictions to appear. Boolos (1999) introduced the *parity principle*, which differs from NP only in that 'there are finitely many' is replaced by 'there are an even (and finite) number of'. It nonetheless has the same effect.

satisfiability and consistency are one and the same in first-order logic (as a consequence of the completeness theorem) the same can not be said of second-order logic; it is possible for a set of sentences to have no (full) models, yet still be consistent. In order to show that NP and HP are together inconsistent, it would be required to show that NP and HP have no *Henkin* models (i.e. models where the second-order domain may comprise less than the entire powerset of the first-order domain). The reason is that for Henkin semantics, there is a completeness theorem, and so satisfiability and consistency coincide. Whether NP and HP share a Henkin model is less clear. In particular, since NP has models of every finite cardinality, a standard compactness argument can show that it also has infinite models. This is, however, not enough to show that HP and NP are consistent; although they both have infinite models, it would need to be shown that they *share* an infinite model, and this is no easy feat. I will not, however, say more on this here.

The issue in this particular case is, anyhow, to a large extent moot. For it is possible to construct abstraction principles which *are* proof-theoretically inconsistent with one another. An entire class of abstraction principles along such lines is provided by the various principles NV_ϕ from the previous chapter.[5] Recall that these are:

$$(\mathrm{NV}_\phi) \qquad \forall F \forall G[\varepsilon F = \varepsilon G \leftrightarrow (\phi(F) \vee \phi(G)) \rightarrow \forall x(Fx \leftrightarrow Gx)]$$

It follows from NV_ϕ that $\exists F \neg \phi(F)$, since, if not, the principle collapses into BLV and a contradiction ensues.

It is then simple to select ϕ and ψ so that NV_ϕ and NV_ψ are inconsistent. The simplest way to show this would be to take ϕ not to depend on F, for example, a sentence which says that F is coextensive with itself, and that the universe contains exactly 3 objects. Then ψ could be taken to be the negation of ϕ.

The challenge which the bad company problem raises is then the following: how do we place restrictions on abstraction principles so as not just to rule out inconsistent abstraction principles such as BLV, but also to decide between inconsistent abstraction principles? One way to do this would be to proceed in a somewhat piecemeal manner, by proposing individual restrictions, and evaluating them individually. If some restriction is unsatisfactory for some reason, then it may be discarded and new restrictions proposed in its stead.

This has been, more or less, how the debate concerning how to restrict abstraction principles has proceeded. Instead, however, I wish to claim that a more systematic approach should be taken, and this in turn requires that more needs to be done to make the challenge precise: what is to count as a restriction, for example, and what is to count as a restriction being successful?

The aim of this chapter will be to do just that. Before suggesting a framework in which we can ask questions relating to proposed solutions to the bad company problem, I will in section 3.2 review a number of restrictions which have been suggested and the relationships between them. In section 3.3, I will argue that a clear framework is required in which to consider questions about proposed solutions. In particular, there is a *prima facie* conflict between the need to restrict abstraction principles and the desire to develop stronger and stronger mathematics in an abstractionist way. A clear framework

[5] Weir (2003) makes heavy use of these to provide a stock of mutually inconsistent abstraction principles. The general idea of using modifications of BLV for such an effect is due to Heck (1992).

would provide a base from which to resolve such issues. Finally, in section 3.4, I will endeavour to set up such a framework.

3.2 Survey of restrictions

Before considering the bad company problem in more detail, it will be useful to give an overview of various restrictions that have been proposed, and the relationships between them. My aim is not to be exhaustive in such an overview, but merely to provide a good enough stock of examples for my arguments in this chaper and the next. In particular, although I include within the examples all restrictions which have been taken seriously as possible solutions to the bad company problem, I have left open certain questions about the precise formal relationships between some restrictions which I mention. Nonetheless, the restrictions which I do mention, and the relationships which I state, will be adequate for my purposes.

Solutions which have been proposed so far can be seen to fall roughly into four categories: restrictions based on the *consequences* of an abstraction principle; restrictions based on the class of models of an abstraction principle; restrictions which partly involve the *form* of an abstraction principle; and restrictions which make use of the fixed domain of quantification over which—on the static view—the quantifiers of an abstraction principle range (typically the absolutely universal domain).

3.2.1 Consequences

In the first category lie restrictions based on the logical consequences of an abstraction principle. Since the background logic is second-order, a distinction needs to be made between semantic consequence and proof-theoretic consequence. So, each definition I give here will actually describe two (or more) restrictions. I shall denote consequence generally by '$\mathrel{\sim}$', where this may either be understood as semantic consequence ($\models$)—giving rise to one restriction—or deductive consequence ($\vdash$)—giving rise to another.

Consistency

The most obvious restriction to consider is that an abstraction principle be *consistent*. That is:

Definition 3.1. An abstraction principle A is *consistent* iff $A \not\mathrel{\sim} \bot$.

This will have two variants, depending on whether consequence is taken to be deductive consequence or semantic consequence. It is well known that semantic consistency entails deductive consistency (as an immediate consequence of the soundness theorem of second-order logic), but not vice-versa (which would amount to a completeness theorem).

As I have already mentioned, consistency will be too weak as a restriction, since it says nothing about individually consistent, yet mutually inconsistent abstraction principles.

Strong conservativeness

A much stronger requirement which might be considered, and which might be motivated by an analogy with explicit definitions is *conservativeness*, or *strong conservativeness* to distinguish it from a restriction which is more often discussed in the abstractionist literature under the name of 'conservativeness'. An abstraction principle is conservative, roughly, if it does not have any consequences which do not involve the abstraction operator. More formally:

Definition 3.2. Let T be a theory stated in a language $\mathcal{L}$ (which does not feature the abstraction operator). Then an abstraction principle A is *strongly conservative* over T iff for any sentence ϕ of $\mathcal{L}$

$$\text{if } T, A \vdash \phi, \text{ then } T \vdash \phi.$$

A is strongly conservative simpliciter if it is strongly conservative over all such T.

The thought behind strong conservativeness is that, given some theory (which may, for example, be a complete theory about all physical objects), an implicit definition should not allow one to derive new consequences about the objects it quantifies over (e.g., physical objects). That is, it should not be possible to draw consequences stateable in the original language which were not already consequences. Explicit definitions are strongly conservative, and thus it might be thought desirable for implicit definitions as well.

Again, there will be both semantic and deductive versions of strong conservativeness. It might also be thought that a *mixed* version of conservativeness might be desirable, whereby it is permitted for an abstraction principle to deductively entail consequences which were not already deductive consequences, but which were semantic consequences of T.

A moment's thought, however, shows that strong conservativeness is too *strong* as a restriction, since it will rule out any abstraction principle that says anything of interest. For example, any abstraction principle which entails the existence of infinitely many objects (as HP does) will fail to be strongly conservative over a theory which states that there are exactly three objects, say.

Weak conservativeness

As such, most attention has been paid to restrictions which involve a weaker notion of conservativeness, which I shall call *weak conservativeness*, or just conservativeness simpliciter if it is obvious from the context that it is the weak variant at issue. Roughly, an abstraction principle will be weakly conservative over a theory T if it has no consequences *for the non-abstracts* that are not already consequences of T. More precisely:

Definition 3.3. Let T be a theory stated in a language $\mathcal{L}$ as before. For any sentence ϕ, let $\phi^{\neg\S}$ result from ϕ by restricting all of its quantifiers by the formula $\neg\exists F(x = \S F)$. I.e. $\forall x \ldots x \ldots$ becomes $\forall x(\neg\exists F(x = \S F) \rightarrow \ldots x \ldots)$ and $\forall F \ldots F \ldots$ becomes $\forall F(\forall x(Fx \rightarrow \neg\exists G(x = \S G)) \rightarrow \ldots F \ldots)$.

Then A is *weakly conservative* over T iff for any sentence ϕ of the language $\mathcal{L}$,

$$\text{if } A, T^{\neg\S} \vdash \phi^{\neg\S}, \text{ then } T \vdash \phi.$$

A is weakly conservative simpliciter iff it is weakly conservative over all such T.

As for strong conservativeness, there are three ways in which this may be taken, depending on how the consequence relation is taken.

This captures more accurately the intuition behind the requirement of strong conservativeness. Recall that this was to forbid abstraction principles which have consequences which are not about the abstracts which they define. But this made use of the principle that a sentence ϕ is about those abstracts only if it features the abstraction operator. But in the presence of absolutely unrestricted quantification—which is a crucial feature of the static approach to abstraction—this principle is flawed. For then even purely logical sentences can, in some sense, be about the abstracts. For example, the sentence that states that there are infinitely many objects may be true just because there are infinitely many abstracts of a certain kind. Thus, its following from an abstraction principle will not violate the requirement that an abstraction principle only have additional consequences which are about the abstracts which it seeks to define.

The modification made in weak conservativeness eliminates this. By explicitly restricting the quantifiers in ϕ, it automatically only considers consequences which are not about the abstracts in question. Similarly for T—by restricting the quantifiers in T, we ensure that we are not capturing consequences which are partly due to something which T says about the abstracts through its use of absolutely unrestricted quantification.

Unfortunately, weak conservativeness is too weak as a restriction. Weir (2003, § 4) shows that we can construct distraction principles which are conservative but mutually unsatisfiable.

Irenicity

Weir suggests as an alternative a stronger notion, which he calls *irenicity*. An irenic abstraction principle is one which is weakly conservative and compatible with all other weakly conservative abstraction principles. As with the other examples, this may be taken in either a semantic or deductive sense. In contrast to the previously considered restrictions, any set of irenic abstraction principles is consistent (Weir, 2003).[6]

3.2.2 Classes of models

A much discussed kind of restriction on abstraction principles concerns the class of models that an abstraction principle has. These restrictions will classify an abstraction as acceptable just in case it has models of an appropriate kind (usually related to cardinality). These have mostly been discussed with respect to *full* second-order models, i.e. those whose second-order domain is the full power set of the first-order domain. It will be useful, however, also to discuss analogues which consider Henkin models, where the second-order domain may comprise less than the full power set of the domain.

Since most of these restrictions involve an abstraction principle being satisfiable by a model of a particular cardinality, it is convenient to introduce a bit of notation. Where κ is a cardinality, say $\kappa \vDash A$ iff there is a model $\mathcal{M} \vDash A$ with $|\mathcal{M}| = \kappa$.

[6] It has been disputed by Linnebo and Uzquiano (2009) whether this means that irenicity is successful as a restriction. In particular, they claim that there are proper classes of irenic abstraction principles which are not consistent taken as a whole. I will argue in section 3.4, however, that such a situation can not in fact arise.

Satisfiability

An obvious model-theoretic correlate to consistency is *satisfiability*. A abstraction principle is satisfiable iff it has a model, or, to use the cardinality language, $\exists\kappa(\kappa \vDash A)$. If we require a full model, then satisfiability will obviously correspond to semantic consistency. If we only require a Henkin model, then, as a consequence of the completeness of Henkin semantics[7], this will correspond to deductive consistency.

Boundedness

Discussions concerning restrictions of this kind have, however, involved more demanding restrictions. Weir (2003) considers *unboundedness*, which he introduces to correspond to conservativeness. It is defined as follows:

Definition 3.4. An abstraction principle A is *unbounded* iff $\forall\kappa\exists\lambda > \kappa$ s.t. $\lambda \vDash A$.

Weir then proves that conservativeness entails unboundedness. He claims also to have proved the converse. Linnebo (2010b) shows that Weir's proof has a gap. In particular, the implication may fail for abstraction principles which feature non-logical vocabulary on their right hand sides. What we need is the following definition:

Definition 3.5. An abstraction principle is *purely logical* if the relation on the right hand side only involves (second-order) logical vocabulary.

Then, Weir's proof in fact shows that if an abstraction principle is purely logical and unbounded, then it is conservative.

Linnebo suggests an improvement on unboundedness so as to do without the additional assumption. He gives the following definition:

Definition 3.6. An abstraction principle A (stated in a language $\mathcal{L}$) is *uniformly unbounded* iff for any $\mathcal{L}$-structure $\mathcal{M} = \langle D, I\rangle$, there is a model $\mathcal{N} = \langle D', I'\rangle$ such that:

- I' interprets all non-logical vocabulary of $\mathcal{L}$ in the same way as I,
- The non-abstracts in $\mathcal{N}$ are precisely the objects of the original domain. That is, $\{a \in D' : \mathcal{N} \vDash \neg\exists F(\S F = a)\} = D$
- $\mathcal{N} \vDash A$.

He then proves that this is equivalent to semantic conservativeness.

These kinds of boundedness all involve full models, and, as such, correspond to the restrictions based on *semantic* consequence. But I suspect that it would be possible to produce similar restrictions by considering Henkin models instead.

Stability

A more stringent restriction—first considered by Fine (2002), and suggested by Weir as a correlate to irenicity—is *stability*. Roughly speaking, an abstraction principle is stable if its class of models is not only unbounded above, but contains no 'gaps', in the sense that it is satisfiable at every cardinality above a certain threshold. This actually leads to two related restrictions:

[7] See, e.g. Shapiro (1991, §4.3)

Definition 3.7. Let A be an abstraction principle. Then:

- A is *weakly stable* iff $\exists\kappa\forall\lambda(\lambda \geq \kappa \Rightarrow \lambda \vDash A)$
- A is *strongly stable* iff $\exists\kappa\forall\lambda(\lambda \geq \kappa \Leftrightarrow \lambda \vDash A)$.

Weir does not distinguish between these, and claims an equivalence between stability and irenicity. The actual situation (proved by Linnebo) is that, if an abstraction principle is semantically irenic, then it is weakly stable, and that if a purely logical abstraction principle is weakly stable then it is irenic.

3.2.3 Inflation and satisfiability

I have already discussed the notion of inflation as a possible diagnosis of what goes wrong with BLV. Recall that an abstraction principle is said to inflate on a domain D if the abstraction relation partitions $\mathcal{P}(D)$ into more equivalence classes than there are objects in D. This notion of inflation has been suggested as forming the basis of a restriction on abstraction principles, notably by Fine (2002).

Now, as it stands, inflation is a *relative* notion; an abstraction principle is inflationary only relative to a domain. But this can be rectified by considering the notion relative to a *particular* domain, namely the fixed domain in which abstraction principles are considered on the static view. That is, inflation can be specified in the very language in which abstraction principles are stated in, with the quantifiers ranging over the same—supposedly absolutely unrestricted—domain.

We thus have the following definition (due to Fine):

Definition 3.8. Let A be an abstraction principle with abstraction relation $\Phi(F, G)$. Then A is *non-inflationary* iff

$$\exists\mathbf{R}[\forall F\exists x\forall y(\mathbf{R}(F, y) \leftrightarrow y = x) \wedge \forall x\exists F\forall G(\mathbf{R}(G, x) \leftrightarrow \Phi(F, G))]$$

Some explanation of this definition is in order. Firstly, unlike the restrictions considered so far, this is a definition in the object language of abstraction itself, rather than the metalanguage. That is, all of its quantifiers are taken to range over exactly what the corresponding quantifiers in abstraction principles themselves range over. The statement, as given this way, is similar to the (object language) statement that says that a concept is small, or that concepts are in one-to-one correspondence. Secondly, in contrast to abstraction principles (at least, the kind of abstraction principles which are currently under consideration), it is *third*-order, since it features quantification over relations between concepts and objects.

But, if we are to consider restrictions which are stated in the object language, it seems that there are some (more or less trivialising) restrictions on abstractions which are also worth considering. For example, why not just restrict to abstraction principles which are *true* (on the fixed domain)?

This is not quite possible, however. For the truth of an abstraction principle will depend on the interpretation of the abstraction operator, which is just what the abstraction principle aims to fix; a restriction stated in the object language, such as inflation, should be stateable without using the abstraction operator. But something very close to truth can be stated—namely, *satisfiability* on the absolutely unrestricted domain—by Ramsifying the abstraction principle. So, we have the following:

Definition 3.9. Let A be an abstraction principle with abstraction relation $\mathbf{R}$. Then A is *absolutely satisfiable* iff

$$\exists \mathbf{f}\big[\forall F \forall G(\mathbf{f}(F) = \mathbf{f}(G) \leftrightarrow \mathbf{R}(F, G))\big]$$

It turns out that these are, in fact, equivalent. Fine (2002) proves that the metalinguistic counterparts to non-inflation and satisfiability are equivalent, and indeed states the object-language version of non-inflation in a way which is very close to how I have defined absolute satisfiability.

There is not, however, much that can be said about the relationship between these restrictions and those canvassed in the previous two sections. The reason is that what relationship holds will be very sensitive to features of the fixed domain itself. So, for example, if the fixed domain is a *set*, then we would be able to say that absolute satisfiability entails satisfiability. But this much may be denied. Likewise, it is tempting to say that if an abstraction principle is stable then it would be absolutely satisfiable, by appealing to a principle that states that the fixed domain is so large that it will be of a cardinality greater than the stabilising point of any stable abstraction principle. But it might be doubted that the notion of cardinality could apply to the fixed domain, being, as it is, supposedly absolutely unrestricted and therefore presumably too large to be a set.

3.2.4 Arrogance

All of the restrictions considered in the previous three sections have be closed under logical equivalence. That is, if an abstraction principle A satisfies a restriction, and B is logically equivalent to A, then B also satisfies that restriction. But one proposed restriction—which has played a prominent role in neo-Logicists' recent discussions of the matter[8]—does not satisfy such a condition.

That restriction is one of *avoidance of arrogance*, which Hale and Wright (2000) describe thus (in the context of implicit definitions in general):

> Let us call *arrogant* any stipulation of a sentence, '#*f*', whose truth, such as is the antecedent meaning of '#_' and the syntactic type of '*f*', cannot justifiably be affirmed without collateral (a posteriori) epistemic work. (p. 128)

They give as an example of an arrogant stipulation 'Jack the Ripper is the perpetrator of these killings', since this presupposes that there is a unique individual who perpetrated the series of killings. They give as another example (in Hale and Wright, 2009a) a stipulation of the Dedekind–Peano axioms outright.

The most notable feature of this restriction is that there is no need to regard it as closed under logical consequence. And indeed, Hale and Wright all but claim that it is not in their appendix to Hale and Wright (2009a); although they are obviously committed to HP satisfying any restrictions on abstraction principle, including non-arrogance, they claim of a particular logically equivalent abstraction principle that it *is* arrogant.[9]

[8] See, for example, Hale and Wright (2009a).

[9] This other abstraction principle was constructed by Ebert and Shapiro (2009) in order to support a claim that abstractionism leads to the view that knowledge of advanced arithmetical theorems can be known without

Again, it is hard to map out just what the relation between non-arrogance and other restrictions might be. The notion is not precise enough as it stands to say much on the matter. But it seems that Hale and Wright see non-arrogance as being an issue orthogonal to matters such as conservativity and the like, since they propose it as a restriction in addition to conservativity.

3.3 The bad company problem and higher mathematics

The challenge of finding an acceptable solution to the bad company problem is liable to intersect with another challenge for abstractionism—that of finding abstraction principles which suffice for the derivation of more powerful mathematics than arithmetic and, in particular, set theory. One aspect of finding a restriction on abstraction principles is that it must be strong enough. That is, it must rule out enough potential abstraction principles so that the ones that are left do not jointly entail a contradiction. But there is also a risk that, in finding a strong enough restriction, it will be *too* strong. It may be that a restriction rules out enough abstraction principles that the ones left do not entail a substantial amount of set theory.

Indeed, there are specific examples of this conflict between finding a restriction that is strong enough, and finding an abstraction principle (or abstraction principles) which can serve as a foundation of at least some set theory. One promising example of an abstraction principle for set theory which has already been discussed is Boolos' New V. This principle is both consistent and recovers some set theory (albeit still far less than might be hoped for).

But it turns out that New V is not conservative (see Shapiro and Weir, 1999). The reason is as follows: By the reasoning of the Buralli-Forti paradox, the concept 'is an ordinal' must not define a set. Thus, by New V, it must be in a one-to-one correspondence with the universe. The presence of this one-to-one mapping then shows that the universe is well-orderable and, in particular, that even the non-abstracts are well-orderable. But this fact will not, in general, follow without the use of New V.[10]

Another instance of how a solution for the bad company may conflict with attempts to found set theory on abstraction principles concerns not any particular abstractionist set theory, but set theory in general. The solution in question is stability (in either its strong or weak form). Note first that stability is a notion which applies, not just to abstraction principles, but any axiomatic theory in general. Uzquiano (2009) argues that *any* theory in which a sufficient amount of set theory can be interpreted is not stable. A consequence of this, then, is that any abstraction principle which recovers a large amount of set theory will not be stable.

Both of these clashes between proposed solutions to bad company and proposed

proof (what they call the *problem of easy mathematical knowledge*). I shall be discussing this claim in the next chapter 4.

It seems that perhaps when they first proposed non-arrogance in Hale and Wright (2000), Hale and Wright saw it as being closed under logical consequence. So, they claim that a stipulation may be non-arrogant by virtue of being equivalent to a non-arrogant stipulation (p.130 n.25). But the claim that non-arrogance allows the avoidance of the problem of easy mathematical knowledge seems to suggest that their more developed view is that non-arrogance is not closed under logical consequence.

[10]The same reasoning will go through even with the other forms of restriction discussed in the previous chapter.

abstractionist foundations of set theory can be resisted. So, for example, it could be claimed that the well-orderability of the universe is a consequence of pure second-order logic, if that is taken to include a principle of global choice (as it often is). Likewise, there are parts of Uzquiano's argument that may be resisted, such as its reliance on the *urelement axiom*—the principle that there is a set which contains all non-sets.

But, regardless of the details of these individual cases, it would be very much desirable to foresee such (potential) clashes. That is, we should not be asking two separate questions: 'is there an abstraction principle which recovers set theory?' and 'is there an appropriate solution to the bad company problem?'. Rather, we should be asking the combined question: 'is there an abstraction principle A, and a restriction on abstraction principles S such that A recovers set theory, S is an acceptable solution to bad company, and A is acceptable by S's lights?'.

In order to be able to answer such a question, we need to set up a more precise, less piecemeal framework concerning solutions to the bad company problem. That is, we would like to be able to survey the space of, not only restrictions which have actually been proposed, but all *possible* restrictions. We may then ask whether any of these solutions is both acceptable as a solution to the bad company problem (whatever that requires), and permits an abstraction principle (or abstraction principles) sufficient for set theory.

The aim of the next section of this chapter will be to provide such a framework, by considering potential solutions to the bad company to be subsets of all abstraction principles.

3.4 What are abstraction principles?

The title of this section may seem strange. After all, have I not already answered such a question in chapter 1? But there are issues that must be resolved if we want to consider the collection of *all* abstraction principles, as we need to if we want to consider restrictions as being sub-collections.

Since all abstraction principles are of the same form,[11] and differ only in what

[11]There are three classes of potential counterexamples to this claim. I believe, however, that these can be accommodated.

The first of these is that of restricted abstraction principles. I showed in the last chapter how these could be accommodated by allowing the abstraction relation to be non-reflexive.

The second concerns where the outermost variables in an abstraction are of a type other than that of concepts variables. We have already seen abstraction principles where the variables range over objects (such as the direction principle). But there have also been examples of abstraction principles where the variables range over *relations* (e.g. Hazen, 1985; Hodes, 1984a), over higher-order concepts (e.g. Cook, 2009a), or perhaps over more exotic types from further up the type-theoretic hierarchy. I will not deal with these cases directly. Instead, it may simply be noted that the method I am about to give (p. 50) could easily be adapted to encompass these. Since there are only countably many types, when it comes to enumerating formulas with two free concept variables, we could easily instead enumerate all formulas with two free variables of any type (as long as they are of the same type).

The final possible counterexample concerns, not abstraction operators of different *types*, but abstraction operators of different *arity*. So, for example, Hale (2000b) makes use of an abstraction operator which is a two-place function on objects, so that abstract terms have the form $\S(t_1, t_2)$ and the relation on the right-hand side is a four-place relation. There are two ways in which such abstraction principles could be accommodated. The first would be to do the same as above, and consider more formulas in our enumeration (in particular, all formulas with an even number of free variables). But alternatively, abstraction principles with a higher arity can be reduced to monodic abstraction principles of a higher type. So, for example, consider a first-oder dyadic abstraction principle (i.e. the abstraction operator is a two-place function from objects to objects):

$$\forall x \forall y \forall z \forall w (\S(x, y) = \S(z, w) \leftrightarrow \Phi(x, y, z, w)) \tag{3.1}$$

abstraction relation they feature on the right hand side, the question reduces to that of what abstraction relations there are. Such an answer will depend on what abstraction relations *themselves* are.

This again may seem like a strange question to ask, but there are indeed two ways in which one could go. The main choice to make is whether abstraction relations are some kind of linguistic entity, such as formulas of a language, or whether they are non-linguistic entities (i.e. higher-order relations) which go beyond just what can be expressed by formulas. These could then be generalised over by means of third-order quantification. There are some immediate attractions to the linguistic approach. For if abstraction principles are thought to be definitions, they must be laid down in some language or other, and there simply is no need to consider inexpressible relations. Furthermore, by considering abstraction relations as formulas, it is possible to make distinctions which cannot easily be made with the non-linguistic approach. For example, we may wish to distinguish between coextensive but differently expressed relations on concepts. But the standard semantics for third-order quantification will individuate relations on concepts extensionally. Even individuating relations on concepts *intensionally* (so, perhaps treating them as something like functions from possible worlds to sets of pairs of concepts) may fail to be fine-grained enough. This is because there may be logically equivalent—and so necessarily coextensive—formulas which we may wish to distinguish between for the purposes of, for example, distinguishing between arrogant and non-arrogant abstraction principles.

There are some problems with the linguistic approach, however. Firstly, it may be claimed that there are indeed ways of laying down individually inexpressible abstraction principles. There are two ways this could be done. The first is to consider formulas which have not only F and G as free variables, but also one or more *parameters* (similar methods are considered by Fine (2002, p.6 n.2) and Linnebo and Uzquiano (2009)). For example, the formula '$Fx \leftrightarrow Gx$' expresses the equivalence relation which partitions concepts into those which apply to x and those that do not (where x is a parameter, referring to some unspecified object). Then infinitely many abstraction principles can be laid down—one for each object x—by laying down the abstraction principle with the free variable x and then quantifying out:

$$\forall x\big[\forall F\forall G\big(\S_x F = \S_x G \leftrightarrow Fx \leftrightarrow Gx\big)\big]$$

Another way would be to use third-order quantification directly. So, for example, a multitude of abstraction principles could be laid down as follows:

$$\forall \mathbf{R}\big[\ldots \mathbf{R} \ldots \rightarrow \forall F\forall G\big(\S_{\mathbf{R}} F = \S_{\mathbf{R}} G \leftrightarrow \mathbf{R}FG\big)\big]$$

This can be replaced by a monadic abstraction principle which ranges over first-order relations. Say that a relation R *encodes* the pair of objects x, y if it holds between those two objects, and no other objects. Symbolically: $\mathrm{Enc}(R, x, y) \stackrel{\mathrm{df}}{=} \forall z\forall w(Rzw \leftrightarrow z = x \wedge w = y)$. Then we can replace (3.1) by:

(3.1′) $\quad \forall R\forall S\big(\S R = \S S \leftrightarrow \exists x\exists y\exists z\exists w(\mathrm{Enc}(R, x, y) \wedge \mathrm{Enc}(S, z, w) \wedge \Phi(x, y, z, w))\big)$

The relation on the right hand side here will not be reflexive. In particular, it will not be reflexive for relations R which do not encode a pair of objects. As such, the effect is that the abstraction principle is restricted to those relations which encode a pair. It can also be seen that this will be equivalent to the original, dyadic, abstraction principle. The same could also be done for polyadic abstraction principles of different types (so, an n-adic abstraction principle whose variables range over entities of type τ can be replaced by a monadic abstraction principle whose variable range over n-ary relations of entities of type τ).

This is in fact the main way in which all abstraction principles in Fine's general theory of abstraction are stipulated (Fine, 2002, pp.165–192).

But are these really ways of laying down infinitely many abstraction principles, including ones which may not be individually expressible? I would say not. For it is not that multiple definitions are being laid down, but rather that just one sentence is being laid down, which appears similar in many ways to abstraction principles. Perhaps such a process—call it *extended abstraction*—could share many of the features and supposed advantages of abstraction. However, if it leads to additional problems regarding bad company, then this is a problem for extended abstraction, not for abstraction simpliciter. In any case, an extended abstraction principle can only play the epistemological role of abstraction principles—that is, providing a means to learn of a new kind of abstract object—by means of a specified instance of it, and this will already be covered by expressible abstraction principles, and hence by the linguistic account.

A second problem for the linguistic account is the question of what language it is that we are to consider formulas of. We could just consider pure second-order logic, and thus restrict attention to the *purely logical* abstraction principles. This would then include a large number of the abstraction principles which have been made use of, such as HP. But it might be desired to have non-logical abstraction principles, and in particular abstraction principles which make reference in their relation to abstracts which have already been obtained by means of abstraction principles. For example, abstractionist approaches to the real numbers by Hale (2000b) and Shapiro (2000) make use of such abstraction principles. We then face the problem of identifying a language which will feature all such abstraction principles.

The problem with identifying such a language is that it seems that it will always be possible to expand such a language by adding more abstraction operators. And so the relations on concepts which are expressible will continually expand as well. It is here in particular that treating abstraction relations as non-linguistic—as whatever third-order quantifiers range over—has an advantage. For third-order quantifiers will not just range over all relations expressible in a language, but will automatically range over all relations expressible in any possible language (and possibly over more than just that). As a consequence, they will already take into account any possible expansion of a language.

But this problem can be avoided if it could be shown that there is a language 'big enough' in some sense. By that, I mean a language which features enough abstraction operators in its collection of non-logical terms and enough potential abstraction principles as sentences, that the addition of any more abstraction operators and principles will not yield any new relations. And indeed, such a language can be constructed as follows.

3.4.1 A language of abstraction

The construction follows what might be taken to be a natural process in building up non-logical abstraction principles in stages. First, logical abstraction principles are considered, then abstraction principles which make reference to the logical abstraction principles and so on.[12]

[12]The resulting hierarchy bears some resemblance to the hierarchy of abstract objects discussed in Hale (1987, chap. 3).

More precisely, we build up languages recursively and then we build up sets of abstraction relations based on this hierarchy of languages as follows: Let $\mathcal{L}_0$ be a second-order language which does not feature any abstraction operators. It might contain other non-logical vocabulary, such as various terms for physical properties and so on. I will assume that such a language is countable.[13] Now, for each $n > 0$, let $\mathcal{L}_{n+1}$ result from $\mathcal{L}_n$ by adding a set of abstraction operators $\{\S_{n,i} : i \in \mathbb{N}\}$. So $\mathcal{L}_1$ is $\mathcal{L}_0 \cup \{\S_{0,0}, \S_{0,1}, \S_{0,2}, \ldots\}$, $\mathcal{L}_2$ is $\mathcal{L}_1 \cup \{S_{1,0}, \S_{1,1}, \S_{1,2}, \ldots\}$ and so on. Finally, let $\mathcal{L} = \mathcal{L}_\omega = \bigcup_{i\in\mathbb{N}} \mathcal{L}_i$. It would also be possible to continue this process into the transfinite in a similar way (but, I shall argue in a bit that this is not necessary).

Now at each stage we can define the set of abstraction principles which can be formed at that stage. At the base level, let REL_0 be the set of formulas of $\mathcal{L}_0$ with the only free variables being F and G.[14] Since the language is countable, this set is countable. Let $\{\phi_i : i \in \mathbb{N}\}$ be an enumeration of it. Now, for each $i \in \mathbb{N}$, let $A_{0,i}$ be the sentence $\ulcorner \forall F \forall G [\S_{0,i} F = \S_{0,i} G \leftrightarrow \phi_i(F, G)] \urcorner$. Note that this is a sentence of $\mathcal{L}_1$. Let $\mathrm{AP}_0 = \{A_{0,i} : i \in \mathbb{N}\}$. So, what we have at this stage is an enumeration of the set of all purely logical abstraction principles.

Similarly, for each $n \in \mathbb{N}$, let REL_n be the set of formulas of $\mathcal{L}_n$ with the only free variables being F and G. Let $\{\phi_i : i \in \mathbb{N}\}$ be an enumeration of REL_n. Now, for each $i \in \mathbb{N}$, let $A_{n,i}$ be the sentence $\ulcorner \forall F \forall G [\S_{n,i} F = \S_{n,i} G \leftrightarrow \phi_i(F, G)] \urcorner$. Note that this is a sentence of $\mathcal{L}_{n+1}$. Let $\mathrm{AP}_n = \{A_{n,i} : i \in \mathbb{N}\}$. So, AP_1 is the set of all abstraction principles which refer to the kinds of objects defined by purely logical abstraction principles and so on.

Finally, let $\mathrm{AP} = \mathrm{AP}_\omega = \bigcup_{i\in\mathbb{N}} \mathrm{AP}_i$. Note that AP is a set of sentences in $\mathcal{L}$, unlike for the finite case, where AP_n is a set of sentences in $\mathcal{L}_{n+1}$.

Now, in what sense is this language and this set of formulas 'big enough'? Firstly, it contains enough abstraction principles for anybody who is only capable of performing finite tasks. That is, any language of abstraction which could actually spoken by a finite being will be contained within $\mathcal{L}$. For every finite process of laying down more and more abstraction principles, no matter how large, will only succeed in laying down abstraction principles which are in AP.

Even if some kind of supertask effort of laying down abstraction principles were permitted, the language and set of abstraction principles considered here will be sufficient. For suppose that some creature manages to expand their language to actually include all of $\mathcal{L}_\omega$ and has considered every one of AP_ω. Then any further expansion of the language and addition of abstraction principles will yield nothing new. This is because any relation on concepts expressible in the language will already be expressible in one of the languages $\mathcal{L}_n$, and so the corresponding abstraction principle will have been formed at stage n. For consider some $\phi \in \mathrm{REL}_\omega$. Since the sentence is finite and so only contains finitely many abstraction operators, there will be some greatest $i \in \mathbb{N}$ such that $\S_{i,j}$ appears in ϕ, for some j. Now, $\S_{i,j}$ is in $\mathcal{L}_{i+1}$. Since each abstraction operator in ϕ has first index $\leq i$, every abstraction operator in ϕ is also in $\mathcal{L}_{i+1}$. Hence, ϕ is in $\mathcal{L}_{i+1}$, and so the corresponding abstraction principle will be in AP_i and hence in AP.

[13] The account can be modified to apply to uncountable languages, by indexing members of each REL_i by some uncountable ordinal.

[14] We can safely ignore the requirement that the relations in abstraction principles be transitive and symmetric. Since any abstraction principle whose relation is non-transitive or non-symmetric will be inconsistent, any solution to the bad company problem will rule them out along with other inconsistent abstraction principles.

Now, we are in a position to make precise any requirements on what should be allowed as a solution to the bad company problem.

3.5 Solutions and sets of solutions

Having decided that abstraction principles should be thought of as sentences, which can be treated as sentences in the language $\mathcal{L}$, it is then possible to treat potential solutions to the bad company problem as subsets of AP, the set of all abstraction principles in $\mathcal{L}$. So, each proposed solution will simply be some set $S \subseteq \mathrm{AP}$. The relationships between them, as in section 3.2, will simply be one of subsethood, or, in the case of equivalence, simply identity.

Treating restrictions as sets of abstraction principles has a couple of advantages over treating them as some kind of description of what is to count as an acceptable abstraction principle. Firstly, sets of abstraction principles will themselves be theories, which are amenable to simple investigation. So, for example, the question of whether a restriction S allows for abstraction principles which can recover set theory just becomes the question of how much set theory can be interpreted in S itself. Or, the question of whether a restriction S lets through mutually inconsistent abstraction principles simply becomes the question of whether S itself is consistent.

Secondly, given a particular restriction, it allows us to distinguish between the question of which abstraction principles are permitted by that restriction, and the details of how that restriction is—or could be—described. This then allows us to separate questions purely concerning the *consequences* of a restriction (such as whether it permits mutually inconsistent abstraction principles, or how much set theory it permits)—which will just depend on which abstraction principles satisfy the restriction—and questions concerning, for example, whether and how that restriction might be described in some language or other. This latter type of question will become particularly important for the purposes of chapter 4.

As well as considering sets of abstraction principles, we may also wish to consider sets of *solutions*, that is, sets of sets of abstraction principles.

A set of potential solutions can be thought of as representing a criterion of adequacy for solutions. And there will be various candidates for such a criterion, depending on what one thinks of as the aim of potential solutions. It might be thought that there is an objective fact of the matter concerning which abstraction principles are acceptable, and thus a set S_0 which is the set of all and only those abstraction principles which are objectively capable of serving as implicit definitions. Then one potential aim—and really the ultimate aim—of proposing restrictions is to find S_0.

But, a more modest aim might be targeted instead. Whether some proposed restriction is in fact S_0 will be a somewhat intractable problem. Instead, we might aim for a simpler, more formal criterion of what is required for a restriction to be successful. Such an aim might be, for example, to avoid inconsistency. Any such aim will then determine a *success set*, $\mathsf{SUC} \subseteq \mathcal{P}(\mathrm{AP})$, consisting of all and only those restrictions on abstraction principles which succeed in the aim.

There there may well be various potential success sets, corresponding to various potential aims. Obvious potential aims are to rule out inconsistency or to rule out unsatisfiability. These would correspond to the sets $\{S : S \nvdash \bot\}$ and $\{S : S \nvDash \bot\}$

respectively (note that these are sets of *solutions*, not sets of abstraction principles, and thus do not correspond directly to consistency and satisfiability as restrictions on abstraction principles). But other options are available; some of the other proposals in section 3.2 could also be taken as proposals concerning when a solution is successful or not.

3.5.1 Restrictions and definability

By treating restrictions extensionally, as sets of abstraction principles, the manner in which restrictions are described has been abstracted from completely. So, for example, conservativity and uniform unboundedness are, according to the foregoing account, one and the same restriction. But it will be useful to relate restrictions, conceived of in such a way, back to possible descriptions of them.

Now, a restriction, if it can be specified at all, must be specified in some language or other, making use of various linguistic resources. In some cases, the resources needed are relatively moderate—as perhaps is the case for some restrictions concerning consequences, which require only the notion of provability or logical consequence. And in other cases, the resources needed are more sophisticated—as is the case in various model-theoretic restrictions, which are stated using quite a bit of set-theoretical vocabulary.

The notion which will do the job of relating sets of abstraction principles to particular descriptions of these sets is that of *definability*. This will be a relative notion—a set of abstraction principles will be definable relative to some language $\mathcal{L}_H$.[15]

Now, any language capable of defining a set of sentences of $\mathcal{L}$ (which is what restrictions are on the present view) will of course have to have the expressive power to talk about, to some extent, the syntax of $\mathcal{L}$. So, for example, it must have terms that refer to sentences and formulas of $\mathcal{L}$, and must have predicates which correspond to certain properties of such sentences.

It would no doubt be possible to construct some minimal language to do this, with terms for each sentence of $\mathcal{L}$ and so on. However, languages capable of expressing syntax are sufficiently similar to the language of (first-order) arithmetic that we may as well use the language of arithmetic as such a language, and represent the sentences and formulas of $\mathcal{L}$ by means of a Gödel numbering.[16]

So let $\mathcal{L}_H$ be the usual language of first-order arithmetic, let $\ulcorner A \urcorner$ denote the Gödel number of A, and let $\overline{\ulcorner A \urcorner}$ be numeral of the Gödel number of A in the formal language $\mathcal{L}_A$. Standard definitions concerning the definability of some set S can then be made:

- $S \subseteq \mathrm{AP}$ is *definable* in $\mathcal{L}_H$ iff there is some formula $\phi(x)$ of $\mathcal{L}_H$ such that for any $A \in \mathrm{AP}$, $A \in S$ if and only if $\phi(\overline{\ulcorner A \urcorner})$ is true (under the standard interpretation). In this case we say that S *is defined by* ϕ.

It is also possible to generalise this to refer to definability by formulas of specific complexity (by, for example considering the place of certain formulas in the arithmetic

[15] So labelled firstly to distinguish it from $\mathcal{L}$, the language in which abstraction principles may be stated (from section 3.4), and secondly, because this will later represent the language which somebody like Hero may speak, before laying down any abstraction principles.

[16] But this will not have to be paired with any particularly strong *theory* of arithmetic. This is especially important considering the use that $\mathcal{L}_H$ will play in the next chapter—as the language that Hero possesses.

hierarchy). Three instances of this will be particularly important for the purposes of the next chapter. These are: definability by a Σ_1 formula, definability by a Π_1 formula, and definability by both a Π_1 and Σ_1 formula. (A formula ϕ is Σ_1 (resp. Π_1) iff it is logically equivalent to one of the form $\exists x_1 \ldots \exists x_n \sigma$ (resp. $\forall x_1 \ldots \forall x_n \sigma$), where σ is a formula which features only bounded quantifiers.)

We thus have the following definitions:

- $S \subseteq \mathrm{AP}$ is Σ_1-*definable* (respectively Π_1-*definable*) iff S is defined by a Σ_1 (Π_1) formula.
- $S \subseteq \mathrm{AP}$ is Δ_1-*definable* iff it is both Σ_1-definable and Π_1 definable.

These types of definition will be important since they relate to *computability*. In particular, the following relationships hold:

- S is Σ_1-definable iff it is *positively semidecidable*: There is an algorithm which, given any $A \in S$, halts and confirms that $A \in S$.
- S is Π_1-definable iff it is *negatively semidecidable*: There is an algorithm which, given any $A \notin S$, halts and confirms that $A \notin S$.
- S is Δ_1-definable iff it is *decidable*: There is an algorithm which, for any A, outputs 1 if $A \in S$ and 0 if $A \notin S$.

So, if a restriction is Δ_1-definable, it will be possible to mechanically compute of a given abstraction principle whether it is acceptable according to that restriction or not. If it is Σ_1-definable, then there will be a procedure which will (eventually) reveal which abstraction principles are acceptable, but may never give an answer for unacceptable abstraction principles (and similarly for Π_1-definability).

I shall also be interested in what can be *proved* about the membership of S given some theory. Let T be some theory in the language $\mathcal{L}_H$. Then we have the following definitions:

- $S \subseteq \mathrm{AP}$ is *positively representable* in T if there is some sentence ϕ of $\mathcal{L}_H$ such that for any $A \in \mathrm{AP}$, $A \in S$ if and only if $T \vdash \phi(\overline{\ulcorner A \urcorner})$. In this case we say that S is (positively) represented by ϕ.
- $S \subseteq \mathrm{AP}$ is *negatively representable* in T if there is some sentence ϕ of $\mathcal{L}_H$ such that for any $A \in \mathrm{AP}$, $A \notin S$ if and only if $T \vdash \phi(\overline{\ulcorner A \urcorner})$.
- $S \subseteq \mathrm{AP}$ is *representable* simpliciter in T if it is both negatively and positively representable in T.

Success sets

As well as the definability of restrictions $S \subset \mathrm{AP}$, we can also consider the definability of success conditions $\mathsf{SUC} \subseteq \mathcal{P}(\mathrm{AP})$. A few difficulties arise here which are not present in the case of the definability of restrictions. In particular, there will not be terms for *sets* of abstraction principles in the language $\mathcal{L}_H$, as there are terms for abstraction principles, so we can not treat definability in the same way. But we can, for the purposes of definability, treat SUC not as a set of sets of abstraction principles, but rather as a

set of formulas of $\mathcal{L}_H$, which themselves define sets of abstraction principles. This will require us to consider a Gödel numbering on the formulas of $\mathcal{L}_H$ (which will thus be required to talk of its own syntax) as well as of $\mathcal{L}$

For a formula ϕ of $\mathcal{L}_A$, let $S_\phi \subset \text{AP}$ be the set of abstraction principles which is defined by ϕ. That is, $S_\phi = \{A \in \text{AP} : \phi(\overline{\ulcorner A \urcorner}) \text{ is true}\}$. Then, we have the following definitions:

- $\mathsf{SUC} \subseteq \mathcal{P}(\text{AP})$ is *definable* iff there is some formula Φ of $\mathcal{L}_A$ such that for each formula ϕ of $\mathcal{L}_A$, $S_\phi \in \mathsf{SUC}$ if and only if $\Phi(\overline{\ulcorner \phi \urcorner})$ is true.
- SUC is *positively representable* in T iff there is some formula Φ of $\mathcal{L}_A$ such that for each formula ϕ of $\mathcal{L}_A$, $S_\phi \in \mathsf{SUC}$ if and only if $T \vdash \Phi(\overline{\ulcorner \phi \urcorner})$.
- SUC is *negatively representable* in T iff there is some formula Φ of $\mathcal{L}_A$ such that for each formula ϕ of $\mathcal{L}_A$, $S_\phi \notin \mathsf{SUC}$ if and only if $T \vdash \Phi(\overline{\ulcorner \phi \urcorner})$.

3.6 Conclusion

The aim of this chapter has not been to say much which directly concerns the *substance* of the bad company problem; I have not sought to defend any particular solutions to the problem, nor to criticise any proposed solutions. Nor have I said anything concerning the consequences of the bad company problem in general for abstractionism.

Rather, my concern has principally been with the *methodology* around the bad company problem. I have claimed that, instead of proceeding in a piecemeal manner, by considering potential restrictions in isolation from one another, we should instead survey the space of possible restrictions as a whole, for which we require a framework to consider such questions.

Providing such a framework for considering restrictions on abstraction principles as a whole has been my main aim. Restrictions on abstraction principles should be considered primarily as sets of abstraction principles. But a restriction may also be treated as some description of a property of abstraction principles, by considering the definability of that set in some language or other.

It is also important to consider what would be required for a proposed solution to be successful in solving the bad company problem. Such requirements can also be considered as a set—this time a set of possible restrictions—and for this as well we can consider particular definability characteristics.

I have not sought in this chapter to put this framework to use. In the next chapter, however, I will do so, by arguing that any restriction, if it is to be capable of playing the appropriate epistemic role in Hero's development of arithmetic, must have certain definability characteristics.

Chapter 4

The epistemological bad company problem

4.1 Introduction

In this chapter, I shall argue that there is an epistemological element to the bad company problem, which a solution must overcome by satisfying certain requirements. I shall furthermore argue that none of the proposed solutions satisfy these requirements and that, although we can construct restrictions with these requirements in mind, the resulting restrictions fail to be successful.

Recall that neo-Fregeanism has as a key part to it an epistemological aim—to show that knowledge of mathematics can be gained through the use of abstraction principles. Then a question arises concerning what epistemological role a restriction on abstraction principles must play. We can again bring the character of Hero into play, and frame the question in terms of her.

Suppose that S is the set of acceptable abstraction principles. We can then ask: What role must S play in Hero's development of mathematics? What must Hero know about S in order to gain knowledge from laying down an acceptable abstraction principle? Now, with this question in place, a potential problem emerges. It may turn out that some proposed S may not be capable of playing the role required of it.

A problem along these lines is raised by Ebert and Shapiro (2009). They consider various options concerning the role that restrictions might play and the requirement on what Hero must know. They find all of these options unsatisfactory for various reasons. I will be considering some options which are very similar to those that Ebert and Shapiro consider.

My aim in this chapter will be to develop an argument similar to that given by Ebert and Shapiro. Along the way, I will make use of an argument given by Weir (2003) which, although not an epistemological argument in its original form can—I believe—be used to press an epistemological point.

My approach will be similar to that of Ebert and Shapiro in that I shall consider various options concerning the role that S may have in Hero's development of mathematics. My argument will differ as follows: Firstly, Shapiro and Ebert ask what is required

for *knowledge*, whereas I shall ask what is required for *justification*. I shall argue in the next section why I believe that it is better to ask the question in terms of justification. Secondly, for each option considered for the role that *S* may play in the epistemology, I shall identify the properties that *S* must have in order to play that role. Finally, I shall consider an option not considered by Ebert and Shapiro, which does not suffer from any of the problems that they raise, nor from the problem raised by Weir. I will however argue that this final option will rule out any of the restrictions which have been proposed.

4.2 The role of restrictions in neo-Fregean epistemology

So, the questions that I aim to answer are the following: Firstly, given some restriction *S* and abstraction principle *A*, what must Hero know (or be able to know) in order for her to gain knowledge that *A* is true simply by stipulating it? Secondly, given an answer to the first question, what must a solution be like in order for the right kind of conditions to be in place for Hero to be able to gain knowledge by stipulating abstraction principles?

The first question can also be asked, not in terms of *knowledge*, but simply in terms of justification. The question then becomes: What does Hero have to know (or have justified belief in) in order to be *justified* in believing the consequences of an abstraction principle?

Indeed, I think that there are reasons as to why it is preferable to ask the question in terms of justification. One reason is principally pragmatic. By only considering justification, we do not need to worry ourselves with a number of tricky issues which may arise due to sceptical worries or the possibility of Gettier cases. So, for example, someone may perhaps claim that in some circumstances, although Hero is justified in believing an abstraction principle *A*, she nonetheless fails to *know* this through the world failing to cooperate in some way or other. For example, *A* may fail to be true, or the true justified belief may fail to be knowledge (as in a Gettier case), each through no fault of Hero's. In considering only the justification question, such problems will not arise. Or, at least, they can then be dealt with separately.

But the justification question will also be adequate to my purposes, since an epistemological challenge which calls into doubt the possibility of justification of a certain kind will be stronger than one that calls into doubt the possibility of knowledge. It will also not be susceptible to a response that it is simply a sceptical worry that may perhaps be brushed off. There are two epistemological aims that abstractionism could be though to have, one weak and one strong. The weak aim is to show how mathematical beliefs—platonistically construed—can be justified. The stronger claim is to show how such beliefs can amount to knowledge. By considering questions about what is required for Hero to be justified in believing an abstraction principle, my target in this chapter is the weaker of these two aims. And if there is a problem for this weaker aim, then clearly there will be too for the stronger aim.[1]

[1] Although *prima facie* it is the stronger of these aims which is required for answering the epistemological challenges to platonism, I believe that a case could be made for the weaker aim sufficing. For if Hero is justified in believing mathematics, platonistically construed, then we will successfully have shown that belief in platonistic mathematics can be justified, and hence that *platonism* can be justified. This seems to me to be adequate to the task at hand. If it could be shown that mathematical beliefs—platonistically construed—are in

Whilst on the topic of justification, it will be useful to briefly consider an issue concerning the nature of justification which will play a role in my discussion. That issue is the debate between *internalists* and *externalists* about justification. Roughly speaking, internalists claim that whether a belief is justified depends only on factors internal to the person who has that belief, whereas externalists do not. More precisely, there are two ways in which one can be an internalist. The first, which is often called *access internalism* holds that, when A is justified in believing that p, A will have access to her source of justification. By contrast, *mentalism* is the view that it is only mental states which play a role in justifying a belief; hence if two agents are alike mentally, they will be alike with respect to whether they are justified in having a belief.

I intend to remain neutral concerning whether the present notion of justification is an internal one. Nonetheless, the issue will arise at various points, in which case I will attempt to clarify why it is that my arguments do not rely on adopting either position.

So, if we then frame the questions in terms of justification, we have the following:

1) What position does Hero need to be in with respect to S and A in order for her to be justified in believing A and its consequences?

2) Given an answer to (1), what does S have to be like in order for it to be *possible* for Hero to be in such a position (and thus be in a position to have a justified belief in A and its consequences)?

The first of these breaks down into two separate questions, regarding what Hero knows (or or is justified in believing)[2] about the relationship between A and S, and what Hero knows about S itself. Firstly, given a proposed solution S, to what extent does Hero have to be able to know of each abstraction principle A whether A is acceptable according to this solution? Secondly, does Hero have to be able to know that S is the set of acceptable abstraction principles, or, at least, does she have to be able to know that it is successful (e.g. that it avoids consistency).

That is, question (1) can be divided into:

1a) What position does Hero need to be in with respect to the question of whether $A \in S$?

1b) What position does Hero need to be in with respect to the question of whether $S \in \mathsf{SUC}$?

Answers to these two question will then determine to some extent answers to the second question. Moreover, these answers can be given in terms of the various definability and representability properties that S may have in a particular language.

There will be some language $\mathcal{L}_H$ that Hero speaks prior to the introduction of abstraction principles, and some theory governing her use of that language. If questions concerning what Hero knows about S are to even get off the ground, it must be the case that $\mathcal{L}_H$ includes some method of referring to abstraction principles themselves. That is, Hero's language must include at least some fragment of some language in which syntax can be formalised. So, we may assume that it is some fragment of the language

general justified, to then object that they are nonetheless systematically false strikes me as, at worst perverse, and at best no more than outright scepticism (see Burgess and Rosen (2005) for a similar claim).

[2] I will refrain from adding this parenthetical remark to each instance of 'know' in this paragraph.

of arithmetic. Similarly, there will be some theory T_H governing Hero's use of this language. Presumably, such a theory will far weaker than, say, Peano arithmetic, since we are assuming that Hero does not have any knowledge of arithmetic. Since the aims of neo-Fregeanism are to show how Hero can have knowledge of *abstract* objects, T_H should be weak enough that any consequences of it can be interpreted as being about concrete objects, such as *tokens* of abstraction principles, rather than abstract *types*. I shall discuss in more detail what such a theory could be later, in section 4.4.1

Now, consider what answers may be given to (1a) and (1b). Suppose that it is required that Hero *knows* that an abstraction principle is acceptable in order to be justified in believing A and its consequences. Whether such a situation is *possible* will depend on various characteristics of S. So, it might be required that S is definable in $\mathcal{L}_H$, in which case it could be claimed that, at the very least, Hero must know what it is for some abstraction principle to be acceptable. It might also be required that S be representable in T_H, so that, not only can Hero know what it is for an abstraction principle to be acceptable, but is able to prove whether any given abstraction principle is acceptable or not.

We can likewise ask similar questions concerning SUC. What we require of SUC will depend on the answers given to (1b).

We thus have the following two questions corresponding to (2) in the same way as (1a) and (1b) correspond to (1):

2a) Must S be definable in $\mathcal{L}_H$? Must S be representable in T_H?

2b) Must SUC be definable in $\mathcal{L}_H$? Must SUC be representable in T_H?

4.3 The options

The way in which I intend to answer these questions is to consider various options concerning what properties S must have. That is, I shall consider primarily possible answers to question (2a). For each option, I will discuss what kind of answer to the first question would motivate such a choice (and the motivation for *that* answer). I will then consider firstly whether such an answer is plausible, and secondly whether there are likely to be any proposed restrictions which satisfy the requirements.

Finally, I will consider answers that might be given to the questions (1b) and (2b). I shall however be brief on this part of the issue, since the options to consider will correspond very closely to the options concerning (1a) and (2a).

4.3.1 Option 1: Provability

One immediate thought would be that, in order to be justified in accepting an abstraction principle, Hero must have justification for the claim that the abstraction principle is acceptable. Since the standard means of being justified in accepting some logical or mathematical claim is *proof*, Hero must then be able to prove that the abstraction principle is successful.[3] The answer to question (1a) would thus be: In order to be justified in believing an abstraction principle A, Hero must have proved that A is acceptable.

[3]Of course, it could be rejected that the claim that an abstraction principle is acceptable is either logical or mathematic. Nonetheless, in cases of the sorts of formally stated restrictions which are common, proof may still be expected.

An answer to the corresponding question (2a) then immediately follows. If this is to be possible, then it must be the case that, for any acceptable abstraction principle A, it is provable from T_H that $A \in S$. I.e., S must be representable in T_H (and hence also definable in $\mathcal{L}_H$).

This position is essentially the same as the first option considered by Ebert and Shapiro, under the title 'Ya really gotta know'—it is the option that 'Hero must be in a position to show that the conditions …are met in order to be credited with knowledge of the implicit definition in question' (p.425).[4] But this would, Ebert and Shapiro claim, be an impossible achievement, whatever the condition of acceptability might be. A necessary (though not sufficient) part of a condition of acceptability will be consistency, and so Hero must be able to prove that a given abstraction principle is consistent. But it is a consequence of Gödel's second incompleteness theorem that a theory (such as one which results from an abstraction principle) can be proved consistent only from the viewpoint of what is in some sense a *stronger* theory, and, as already noted, whatever T_H might be, it will surely be much weaker than arithmetic.[5]

In describing what is required for Hero to *know* that a principle is acceptable, there is an alternative to the requirement that it be provable. It could instead perhaps be required simply that it is *decidable*, in the sense that there is some effective decision procedure, such as an algorithm, for determining whether any abstraction principle is acceptable or not. In this case, the condition of acceptability would not be given in terms of some sentence that needs proving by Hero, but rather in terms of something like an algorithm that must give the correct answer. This approach, however, seems to suffer the same problem as requiring provability. For it is undecidable whether an arbitrary abstraction principle is consistent or not.[6]

So, the requirement that Hero must know seems to be far too demanding. It would require something that is simply impossible. Indeed, it seems that a similar requirement for any similar project would be too strong, resulting in some form of scepticism about many kinds of knowledge.

The situation is not, however, quite so simple; Ebert and Shapiro's objection needs extra finessing. Although *consistency* as a requirement is neither provable nor decidable, it does not follow that any restriction S suffers similarly. For although S must only contain consistent abstraction principles—so that it is a subset of the (undecidable and unrepresentable) set of consistent abstraction principles—it does not follow that S itself is unrepresentable or undecidable.[7] Consider for example the rather simple restriction

A weaker condition for Hero being justified that $A \in S$ may also be suggested, such as having good inductive evidence. Such a position will be similar to one that I consider later.

[4]Shapiro and Ebert characterise such a position as internalist. But it should be noted that, although the position would be acceptable to internalists of any stripe, it does not require internalism to be true in generality. For there is nothing about externalism to suggest that there are not certain kinds of belief—or ways of forming beliefs—in which the requirements for justification are internal; the situation of mathematics and proof seems plausible as such a case.

[5]It is not quite right that a *stronger* theory is required. For example, Gentzen's proof of the consistency of arithmetic requires a theory which is neither stronger nor weaker than arithmetic in any natural way. Nonetheless, whatever T_H might be, it is unlikely to be able to prove the consistency of arithmetic.

[6]Heck (1992) proves this result. The reason is that such a decision procedure could be transformed to a decision procedure to decide whether an arbitrary sentence of second-order logic is consistent, and no such decision procedure exists.

[7]Alternatively, this could be put in terms of the conditions which define the set S—as Ebert and Shapiro do—by saying that 'consistency is among the conditions' (p.425). In this case, consistency only needs to

on abstraction principles which says that HP is acceptable, but no other abstraction principles are, even if they are equivalent to HP (ie. $S = \{\text{HP}\}$). Then this entails consistency, in that all abstraction principles satisfying it (ie. just HP) are consistent. But this restriction will be both definable and representable in any reasonable choice of $\mathcal{L}_H$ and T_H. For the formula $x = \overline{\ulcorner \text{HP} \urcorner}$ will define S, and any (even very minimal) theory of syntax should be able to prove such an equality. (Similarly, because S is definable by such a simple formula, it is decidable.)

Whether this is a viable option will depend on whether Hero—as well as being able to prove of each $A \in S$ that $A \in S$—can also *prove* that S is a subset of the consistent abstraction principles (or, in terms of conditions, that the conditions entail consistency).[8] And whether Hero can prove this will depend on the answer to the second question, concerning whether Hero needs to be a position to prove whether the restriction S is *successful*, ie. whether $S \in \mathsf{SUC}$. The kind of motivation for expecting S to be representable will also presumably motivate this requirement as well. If it *is* required that Hero has to prove this, then there are two options. Either Hero will be acquainted with S via the aforementioned formula ($x = \overline{\ulcorner \text{HP} \urcorner}$), in which case a proof that $S \in \mathsf{SUC}$ will essentially just be a proof that HP is consistent. This will be impossible given a relatively weak theory T_H for the given reasons. Or, Hero will be acquainted with S by some other definition, which she *can* prove to be consistent (perhaps), but for which she can not prove that HP is a member. So, if both requirements are in place—that Hero is able to prove that $S \in \mathsf{SUC}$ and that $A \in S$—*then* the proposal that Hero must *know* is going to be too strong.

4.3.2 Option 2: Entitlement and imponderable solutions

In any case, it does seem that to demand that it be proven whether any abstraction principle is acceptable would be too strong, and would bring along with it an epistemological form of the bad company problem which could simply be rejected. Might we instead be able to motivate a position according to which much less is required of Hero? That is, may we be able to claim that, in order for Hero to be justified in believing an abstraction principle A, she need do no prior work in establishing whether or not A is acceptable (or, at least, no work so onerous as providing a proof)?

What could such motivation look like? One way might be to claim that, if $A \in S$, then this fact *itself* may serve as justification of a belief that A is acceptable, regardless of Hero's acquaintance with this fact. This view would be an extreme form of externalism about justification, since the justification for the belief is neither accessible to Hero, nor a part of Hero's mental life. It amounts to the claim that a belief that p is justified just in case p is true (albeit for a restricted class of propositions).

At best, however, this view is unmotivated. For it to be viable, something would need to be said about how it could be that the mere truth of a belief could result in that belief being justified, independently of how that belief was formed. There are certain kinds of belief for which it may be possible to motivate such a principle—such as beliefs

be among the conditions in that they materially entail consistency. But this entailment may fall short of provability.

[8] Note that it is important that what is at issue is whether Hero can *prove* this, and not simply that she knows it. For it seems that Hero will know that any acceptable abstraction principle is consistent in any case—she is supposed to be able to know that they are *true*, which surely entails consistency.

about one's own mental state—but this motivation is unlikely to also apply to the present case.

But, a more plausible approach is available which would have a similar result, and which would—depending on the precise details—be acceptable to both internalists and externalists. The problem with the previous option, it might be thought, is that it requires Hero to do the requisite work needed in order to become *certain* (or close to certain) that an abstraction principle is acceptable. Such a requirement of certainty is likely to lead to scepticism in any circumstances. Instead, it could be claimed that Hero may have an *entitlement* to believe that an abstraction principle is suitable as an implicit definition, where an entitlement is a kind of right to believe which does not require having *evidence* for that proposition. There have been a number of proposals concerning entitlements to beliefs,[9] but the proposal which is most relevant to the present task is that of Wright (2004a,b).

According to Wright, there are certain propositions which we may be entitled to accept without having evidence for their truth.[10] Such propositions he calls *presuppositions* or *cornerstones* of some cognitive project.

Wright defines what it is for a proposition to be a presupposition as follows:

> *P* is a presupposition of a particular cognitive project if to doubt *P* (in advance) would rationally commit one to doubting the significance or competence of the project. (Wright, 2004a, p. 163)

Then, we will be entitled to accept *P* under the following circumstances:

> (i) [T]here is no extant reason to regard P as untrue and
>
> (ii) The attempt to justify P would involve further presuppositions in turn of no more secure a prior standing, …, and so on without limit; so that someone pursuing the relevant enquiry who accepted that there is nevertheless an onus to justify P would implicitly undertake a commitment to an infinite regress of justificatory projects, each concerned to vindicate the presuppositions of its predecessor.
>
> (Wright, 2004a, p. 163)

Now, how could this motivate the claim that Hero is not required to know *anything* about the acceptability of abstraction principle *A* in order to be justified in believing *A*? And what kind of answer to the second of my questions—concerning what properties a proposed solution must have—can it be used to motivate?

First, we must be clear on what the cognitive project in question is, and what the cornerstone proposition is that Hero may be entitled to accept. Consider an abstraction

[9] See, for example Burge (1993); Dretske (2000); Peacocke (2004).

[10] Wright makes a distinction between belief in a proposition, and mere acceptance of the proposition, where the latter may fail to amount to the former. The details of such a distinction will not make a difference to my evaluation of the present option. In particular, it should make no difference whether Hero has an entitlement to believe that an abstraction principle is acceptable or merely an entitlement to accept that it is acceptable. I will tend to use 'accept' and 'believe' interchangeably.

Wright also seems to refrain from calling entitlement a kind of justification (and other authors making use of entitlements explicitly make a distinction between justification and entitlements). Again, I shall not be making such a distinctions; the way I am using 'justification' is such that, if there are entitlements, then they are a form of justification.

principle A, which purports to provide a definition for terms referring to certain kinds of abstract objects, Fs. It seems that it is a worthwhile cognitive project to make use of A in order to expand one's vocabulary (to include terms referring to Fs), and to discover truths about Fs. In this case, it seems that a presupposition of the cognitive project is that A is indeed capable of providing an implicit definition of Fs. Otherwise, our terms purporting to refer to Fs would be ill-defined, and thus the project would fail to be a competent one. [11]

Now, if the conditions (i) and (ii) obtain concerning the acceptability of A, it is possible to see how Hero may be justified in believing A. The story is similar to one which may be told as in option one; in order for Hero to be justified in believing A and its consquences, she must be justified in believing that it is acceptable as an implicit definition. But, as long as (i) and (ii) obtain, she *is* justified in believing that A is acceptable as an implicit definition, as a matter of entitlement.

So, what would a proposed restriction S have to look like for the foregoing to be the case? It seems that there would be no requirement whatsoever on what S is like. Consider, for example, a restriction which is neither representable in T_H, nor even definable in $\mathcal{L}_H$. That is, the restriction S can only be stated in a language that Hero does not understand. Some of the model-theoretic restrictions considered in 3.2.2 may fall into such a category. They are typically stated in a way that makes use of much set-theoretic vocabulary. But it is an assumption about Hero that she does not start off with an understanding of even arithmetical vocabulary, let alone the kind of sophisticated set-theoretic vocabulary needed to express, say, stability.

Hero in such a case will obviously not be able to accept the claim that A satisfies the particular restriction S, but she will be able to accept that A satisfies *whatever it takes to be suitable as an implicit definition*. Moreover, it will also be the case that this belief (or acceptance) will satisfy (i) and (ii) in more or less trivial ways. Since S is not even definable in Hero's language, Hero will not be able to know what it is for some abstraction principle to be unacceptable, and so can not be in possession (or be expected to be in possession) of any reasons to suspect that A is unacceptable. Moreover, due to the non-representability of S, there is no hope of Hero being able to prove that S is acceptable. Thus Hero will be entitled to accept that A is acceptable.

A similar situation arises if a slightly less extreme position were to be taken, according to which it is required that S be definable—so that Hero can state what it is for some abstraction principle to be acceptable—but with no requirements at all on representability. Then there may be restrictions S such that: there are abstraction principles S with $A \notin S$, but for which no evidence can come to light, from Hero's perspective, that shows that they are indeed unacceptable. An example might be Fine's requirement that an abstraction principle be *non-inflationary*, where an abstraction principle is non-inflationary if it does not require that there are more abstracts than there are objects. A non-inflation requirement can be expressed internally, since it is expressible in pure third-order logic. But whether an abstraction principle is non-inflationary will depend very much on how many objects there actually are, which can not be known prior to laying down an abstraction principle.

[11] This claim concerning what should be taken to be the candidates for being entitlements should be contrasted with an alternative application of entitlements to the neo-Fregean programme. For example, Pederson (2011) claims that it is the abstraction principles themselves (and, in particular, HP) which are candidates for being entitlements.

Call such restrictions *imponderable*. They are the restrictions which are neither positively nor negatively representable in T_H, and may even be undefinable in $\mathcal{L}_H$. That is, no possible evidence available to Hero can tell either for or against an abstraction principle's acceptability. The resulting position is then the following: in answer to question (1a), for Hero to be justified in believing A and its consequences, the conditions (i) and (ii) of entitlement must obtain.[12] In answer to question (2a), *any* restriction can then play such a role, whether ponderable or imponderable. In the case of imponderable restrictions, it will be the case that the conditions (i) and (ii) obtain automatically. The resulting picture—at least in the case of imponderables—is similar to that discussed by Ebert and Shapiro under the title 'if it's good it's good'.

There are a number of problems however with this position. The first, given by Ebert and Shapiro, is that, given some plausible assumptions about the behaviour of restrictions, it leads to it being possible for Hero to know many non-trivial mathematical truths (their example is Fermat's Last Theorem) without significant work. The second—which is similar to a problem that Weir (2003) raises for a particular proposed restriction (stability)—is that it makes it impossible to distinguish between the good abstraction principles and the bad.

The problem of easy mathematic knowledge as presented by Ebert and Shapiro (pp.429–430) is as follows. Let ϕ be some true sentence in the usual language of arithmetic, a standard proof of which may be highly non-trivial. Now, in pure second-order logic a sentence can be constructed which effectively says that any structure satisfying the second-order Dedekind–Peano axioms satisfies ϕ. Let $\phi[F, f, x]$ be the result of replacing each occurrence of 0 and s in ϕ by variables x and f of the appropriate type, and restricting the quantifiers to F. Then, where PA is the conjunction of the second-order Dedekind-Peano axioms, and $\mathrm{PA}[F, f, x]$ is formed in the same way as $\phi[F, f, x]$, let ϕ^* be the sentence $\forall F \forall f \forall x (PA[F, f, x] \rightarrow \phi[F, f, x])$. When ϕ is true of the natural numbers, ϕ^* will be a truth of pure second-order logic.

Now, consider the following abstraction principle:

$$(\mathrm{HP}_\phi) \qquad \forall F \forall G[\S F = \S G \leftrightarrow F \approx G \wedge (\phi^* \vee \forall x(Fx \leftrightarrow Gx))]$$

The important properties of HP_ϕ are as follows. (a) The abstracts introduced by HP_ϕ can be considered to include natural numbers in that—as with HP—a version of the Dedekind–Peano axioms can be proved by introducing the language of arithmetic as abbreviations. (b) Since ϕ is true of the natural numbers (and hence ϕ^* is logically true), HP and HP_ϕ are logically equivalent. (c) HP_ϕ entails ϕ (where the arithmetical vocabulary in ϕ is substituted according to the abbreviations used in (a)), and the proof is relatively simple—if ϕ^* were not the case, then HP_ϕ would effectively become BLV and so would result in inconsistency.

Now, *on the assumption that acceptability is closed under logical consequence*, HP_ϕ will be acceptable, since HP will presumably be deemed acceptable by any proposed restriction. So, under the present proposal, Hero can lay down HP_ϕ to gain both knowledge of the natural numbers and that ϕ is true of them. To do this, she does not need to prove that it is an acceptable abstraction principle.

Of course, the proponent of the view that Hero need not know anything about whether an abstraction principle satisfies a given restriction could simply reject the

[12]There is a further question concerning whether Hero must know that they obtain. An access internalist may require that she must, whereas others may not.

assumption that the correct restriction will be closed under logical consequence. The problem of easy mathematical knowledge simply shows that the correct restriction will be one that is *not* closed under logical equivalence, so that HP is acceptable whilst the equivalent HP_ϕ is not. This is essentially the reply that Hale and Wright (2009a, pp.478–481) give; they reject HP_ϕ on account of it being '*arrogant*', where arrogance is 'the situation where 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' (p. 465). It seems to me that such reply is decisive.

However, there is another problem with this position which does not rely on the assumption of closure. In particular, an issue arises for imponderable restrictions. Consider a restriction which is imponderable (and which is being relied on alone as demarcating the good abstraction principles from the bad)[13] For a concrete example, suppose that the restriction in question is simply the satisfiability of an abstraction principle on the universe. That is, for an abstraction principle

$$\text{(AP)} \qquad \forall F \forall G[\S F = \S G \leftrightarrow \phi(F, G)],$$

to be acceptable, its Ramsey sentence (a third-order sentence):

$$\exists \mathbf{f} \forall F \forall G[\mathbf{f}(F) = \mathbf{f}(G) \leftrightarrow \phi(F, G)]$$

must be true (with its quantifiers ranging over absolutely everything). This restriction is plausibly internally definable (since it is expressed purely in third-order logic), but appears to be neither positively nor negatively representable (since if it were known of some abstraction principle that its Ramsey sentence were true, there would be no need to stipulate it). It is thus imponderable.

Now, suppose that there are two theorists, Hero1 and Hero2, who lay down, respectively, HP and NP, and hence come to believe the consequences. According to the standard abstractionist story, in such a situation Hero1 comes to know HP, and thus is in a position to come to know the Dedekind-Peano axioms by laying down a few explicit definitions. By contrast, Hero2's belief is faulty in some way. But this difference can not be explained in terms of differing justificational status; for, given the story about entitlement, both Hero1's belief and Hero2's belief are justified, and justified in essentially the same way.

We thus have a situation in which both Hero1 and Hero2 are both justified in their resulting beliefs, despite them being inconsistent with one another. And indeed, anyone in a similar situation (i.e. laying down some abstraction principle or other) would be justified in their belief. This is situation is a rather unattractive one to accept. It is of course not flat out inconsistent; it may in general be possible for different people to be justified in their mutually incompatible beliefs. But such situations will presumably be ones in which the two people have access to different evidence. For example, if Hero1 and Hero2 both had different views of the same physical object they may form mutually incompatible beliefs, but still be justified in having those beliefs. The present situation is not like that, however; Hero1 and Hero2 have access to exactly the same information, and, moreover, they are justified in having their respective beliefs to the highest level possible.

[13] The assumption that the imponderable restriction is relied on by itself is not strictly speaking required. For, if a restriction is imponderable, so too will be the conjunction of that restriction with any other restriction.

The problematic nature of the situation could also be explained in terms of the notion of reliability. The situation is such that justified beliefs about abstraction are utterly unreliable. That is, being justified in believing an abstraction principle is in no way a reliable indicator of that abstraction principle being true. For there will be infinitely many unacceptable abstraction principles which are justified on such an approach (since all abstraction principles would be justified). One does not have to be a reliabilist[14] to see this as problematic. For presumably it is the case that we *should* accumulate beliefs in such a way that—barring any external factors which are out of our control—we have mostly true beliefs. But, with imponderable restrictions on the scene, this will not be the case.

Still, it might be thought that, as long as we can appeal to an external perspective from which to evaluate the beliefs of Hero1 and Hero2, there will be no problem. That is, since we know that the universe is infinite, we can explain why it is that Hero1's belief results in knowledge, whereas Hero2's belief is justified but mistaken in some sense. This is to take an external perspective in that we already assume a certain amount of mathematical theory, in which we can determine that there are infinitely many objects. But then how do we justify *this* perspective? Such a perspective essentially presumes platonism. But, if the aim of this epistemological discussion is to show how it is that platonism may be justified, this would clearly beg the question against those—such as nominalists—who would claim that it is not.

A final suggestion concerning imponderable solutions might be the following: At first, perhaps, a restriction will be imponderable for Hero. But then, having been entitled to accept that certain abstraction principles are acceptable, she builds up a body of mathematical vocabulary and theory. At this point, the restriction is no longer imponderable; Hero may kick down the ladder provided by her initial entitlement and *prove*, from her new perspective, that the initial abstraction principles were acceptable after all. But, on further inspection, this too will not solve the problem. For which abstraction principles may be proved to be acceptable will depend on which abstraction principles are laid down in the first instance. For example, on laying down NP, it will be possible to prove that the Ramsey sentence of NP is true, whereas the Ramsey sentence of HP is false, and conversely.

This problem is not, however, specific to this example (that is, of universal satisfiability); it will apply to any restriction which is imponderable. For in any such case, there could be a situation in which Hero1 and Hero2 lay down competing abstraction principles, which—due to the imponderability of the restriction—can not be decided between, as in this case.

So, for example, Weir (2003) raises a problem for *stability* which can be seen to result in the same kind of situation.

Weir raises a problem which he calls '*Embarrassment of Riches II*' (*Embarrassment of Riches* is his name for the bad company problem). He does not consider it as a specifically epistemological problem, but rather as an 'analogue of the ER objection simply [recurring] at a metatheoretic level' (p. 15). Nonetheless, the problem has—as a corollary—exactly the kind of epistemological problem that I have just raised.

The problem is the following:[15]

[14]Reliabilists about justification claim that a belief is justified just in case is has been formed by a reliable process. See, for example Goldman (1979)

[15]See page 3.1 for the definition of a distraction.

> Consider a bunch of theorists, each taking a distraction principle as the basis for their pure mathematics, but a different one, utilizing a different definition of Bad, in each case...
>
> ...They can all take over the same definition of stability and each can define 'acceptable' in the same way but relative to provability from their own abstraction principle. Moreover, from the standpoint of any one theory, each of the others is unstable either because it places a cap on the universe at some unacceptably low cardinality or because it has no set models at all. And since the five distractions are pairwise inconsistent, each can prove that every other is unacceptable.

The upshot, then, is just the same as that for universal satisfiability. In particular, we have a number of theorists, each with their own abstraction principle, but no way of deciding between them. Moreover, Weir's example shows that the 'ladder kicking' response will be no good here either; the notion of stability which arises for any set theoretic abstraction principle will be such that that abstraction principle passes, but others fail.

Likewise, similar arguments could be levelled at any other imponderable solution.

4.3.3 Option 3: Innocent until proven guilty

To avoid this problem, there is no need, however, to go all the way back to option 1, requiring that Hero first prove that her chosen abstraction principle is acceptable. An intermediate position can be arrived at by modifying the notion of entitlement slightly.

Entitlement permits a certain amount of fallibility in Hero's beliefs, and for her to 'fly without a safety net', as Shapiro and Ebert put it (p. 433). Her belief in the acceptability of an abstraction principle is innocent until proven guilty.

The problem with imponderable restrictions arises because it forecloses the very possibility of any abstraction principle to be proven guilty from Hero's perspective, even in the case of abstraction principles which we would, from our external perspective, declare guilty. That entitlement appears to permit imponderable solutions then seems to be to be more of a bug than a feature of the notion. For the notion of entitlement to be meaningful, we must require that there be such a possibility; imponderable restrictions must be ruled out.

I would suggest that the problem may be avoided by adding a new condition to entitlement that corresponds to this requirement. That is, in order for Hero to be entitled to accept a proposition P (in this case, that an abstraction principle is acceptable), as well as (i) and (ii) holding, it must be the case that, if P is in fact false, it is possible, in principle, for Hero to discover that it is false.[16]

But the upshot concerning questions (1a) and (2a) should be fairly clear. The answer to (1a) will be much the same as in option 2. But in answer to (2a), if it is to be possible for Hero to discover of unacceptable abstraction principles that they are indeed unacceptable,

[16]Some care would be needed to give a fully rigorous specification of such a requirement, and, in particular, the treatment of the conditional featured in it. A material conditional would not seem adequate, and something like a counterfactual conditional would face problems related to the fact that if an abstraction principle is acceptable, it is necessarily so. But it seems likely that such worries could be overcome.

we must have that S is negatively semi-representable in T_H. That is, S must be definable by a formula ϕ in $\mathcal{L}_H$, and, given some $A \notin S$, it must be provable from T_H that $\neg\phi(A)$.

Now that the modification to entitlement has been made, the problems with the previous two options no longer arise. Consider first the problem with option 1, that it runs afoul of Gödel's incompleteness theorems. This problem no longer arises. What the incompleteness theorems show is that it is not possible to prove of consistent abstraction principles that they are indeed consistent. But what is now required (if we just consider consistency) is a kind of converse. We need to be able to prove *in*consistency, and the incompleteness theorems do not threaten this. Indeed, it is clear how to prove that an inconsistent sentence is inconsistent without assuming much—if any—mathematics: Simply derive a contradiction from that sentence.

The present option will also not suffer from the main problem that I raise for the second option. Suppose again that we have two theorists, Hero1 and Hero2, and they both put forward competing abstraction principles, which can not both be acceptable. Suppose for example that Hero1's abstraction principle is acceptable and Hero2's is unacceptable. It may be that, at first, neither Hero1 and Hero2 are aware of any evidence which tells against their respective beliefs. But, if the notion of acceptability is negatively representable, then it will be possible to prove, from the initial perspective shared by Hero1 and Hero2, that Hero2's abstraction principle is unacceptable. When made aware of such a proof, Hero2 will no longer be justified in believing her abstraction. We thus do not reach the kind of standoff that resulted in the previous case.

4.3.4 Definability and representability of SUC

I have still not said much about questions (1b) and (2b), which relate to what Hero must know about whether S is successful as a restriction. Ultimately, we want to know to what extent we require SUC to be definable or representable in Hero's initial language.

There is not much new to say in this regard. The options will be much the same as before, and corresponding motivations and objections will apply.

As before, an immediate thought is that Hero must be able to prove of any S whether $S \in \mathsf{SUC}$ or not. The reason is that, in order for Hero to make use of a restriction, she first has to come by such a restriction. Unless some kind of divine revelation is imagined, in which Hero is told of some restriction from on high (or, more plausibly, by somebody approaching the neo-Fregean enterprise from an external perspective), the only clear way in which Hero can come to accept some restriction is by proving it to be successful.

However, unlike the first option considered previously, this position does clearly fall foul of Gödel's second incompleteness theorem. This would require that Hero be able to prove of any consistent set of abstraction principles that it it consistent. Since some sets of abstraction principles are sufficient to interpret arithmetic, this will not be possible unless Hero already has access to some particularly strong theory.

We can, however, give a similar answer to this question as we did to the previous question. That answer is that, as long as SUC is negatively representable, Hero will have an entitlement to accept that a given solution S is successful, absence evidence to the contrary. The resulting picture would be as follows: Hero may try out some particular restriction, say consistency, and be entitled to accept that it is successful. However, she may then come across reasons that the solution is *not* in fact successful (say, the mutual incompatibility of HP and NP), at which point her entitlement is defeated, and she is

forced to find some other restriction. In fact, this would be much the same as how the discussion of possible solutions has gone in the literature—restrictions are proposed, then possible defeaters are found, and so on.

4.4 Evaluating solutions

The situation is then as follows. I have argued that there must be an epistemological component to considerations about the bad company problem. I have furthermore argued that, in order for a proposed restriction on abstraction principles to be able to play the epistemological role required of it, it must be *negatively representable* in Hero's original theory T_H. Moreover, the set of successful restrictions (which we are assuming to simply be those which avoid inconsistency) must also be negatively representable in T_H.

Now, *prima facie*, there will be negative consequences to such a requirement. A number of solutions which have been proposed, including those which seem most promising, seem to be ruled out. Among these, it seems, would be the various model theoretic solutions, since they appear to require a substantial amount of mathematical language to state. Those restrictions such as non-inflation which involve the unrestricted domain may also be rejected since there does not seem to be an internally acceptable method by which Hero could rule out inflationary abstraction principles. At the other end of the scale, consistency appears to pass the test, although it fails to rule out all bad company.

But, to be sure of these consequence, a more systematic approach is needed. It is not clear, for example, whether any of the solutions based on consequences will pass the test. Or, it might be the case that some of those which appear at first to be ruled out, should not be, perhaps in a different guise. Two things are required. First, we need to be clear on what we should take to be Hero's language $\mathcal{L}_H$ and theory T_H. Secondly, we need to check each restriction to see if it is negatively representable, given such a choice. It might also be desirable to investigate whether there are any sets of abstraction principles which (1) avoid all instances of bad company and (2) satisfy the relevant definability and representability conditions. I will however argue that this latter is not a realistic goal.

4.4.1 Hero's language and theory

Before we can assess any of those restrictions which have been suggested, plausible candidates for $\mathcal{L}_H$ and T_H need to be settled on. The language and theory are ideally going to be very weak, since Hero is supposed not to have any knowledge of mathematics before laying down abstraction principles.

If Hero's language is impoverished enough that it cannot even refer to abstraction principles, then *no* proposed solutions will satisfy the requirements that I have claimed are needed if neo-Fregeanism is to fulfil its epistemological aims. No solution will be expressible in such a language.

So, such a language must have terms referring to certain linguistic entities, and as such, may as well be taken to include the basic vocabulary of arithmetic $(0, s, +, \text{etc.})$ (although it may make more sense to include operations which represent certain primitive syntactic interpretations such as concatenation rather than addition and multiplica-

tion).[17] This does not mean, however, that a full theory of arithmetic is being attributed to Hero, since that would ultimately be self defeating. It also does not need be the case that this attributes the capability of reference to numbers to Hero; by considering some codification of syntax, each arithmetical term can be interpreted, not as referring to some natural number n, but as referring to the linguistic entity encoded by n.[18] The use of a (weak) theory of arithmetic is merely for convenience.[19]

In order not to attribute any knowledge of an infinity of objects already to Hero, it must be the case that no sentences of $\mathcal{L}_H$ that she can derive by use of T_H alone can be seen to committing her to infinitely many objects. It may also be desirable that the objects which it *does* commit her to can be taken to be concrete. Such a constraint would arise if one thought that *all* abstract objects are to arise from abstraction principles, and so Hero must be assumed not to have any knowledge of any abstract objects prior to laying down abstraction principles.

One option which seems to fill these requirements is *Primitive Recursive Arithmetic* (PRA). The most notable feature of the language of PRA is that it does not contain any quantifiers, and this is why no sentence in it can be taken to be about infinitely many objects. Instead, it features a symbol for every *primitive recursive* function and predicate. Importantly, for the purposes of considering it as a theory of syntax, this will include function symbols and predicates that correspond to natural syntactic operations and properties. So, for example, there will be a function symbol corresponding to concatenation, and predicates such as 'x is an abstraction principle', 'x is a longer string than y' and 'x is a proof of y from z'. (It will also contain predicates and functions which do not correspond to natural operations on linguistic entities, but these can simply be ignored.)

The *theory* which goes with PRA is similar in many ways to a standard theory of arithmetic, but devoid of quantifiers. There are the two axioms:

1. $sx = sy \rightarrow x = y$
2. $\neg sx = 0$

then axioms governing each primitive recursive function. Eg. for addition:

3. $0 + y = y$
4. $sx + y = s(x + y)$

Finally, there is a *rule* of (quantifier-free) induction:

$$\frac{\phi(0) \qquad \phi(x) \rightarrow \phi(sx)}{\phi(x)}$$

[17] To see that any language of syntax will essentially contain the language of arithmetic, consider the following. A language of syntax will have at least one term for at leas one symbol, say '|'. It will also have the means of expressing concatenation of symbols. Then, '|' can serve as zero, and concatenating x with | can serve as a successor function. It is also clear that these will behave just as zero and successor behave.

[18] The coding would have to be such that every $n \in \mathbb{N}$ codes some linguistic item, as well as every linguistic item having a code or Gödel number (i.e. the coding function is a bijection). This is not typically the case for Gödel numberings, but, since there are no more natural numbers than linguistic items, such a coding will of course be possible.

[19] A language which is more transparently a theory of syntax rather than arithmetic would require, for example, a constant referring to each symbol in object language instead of a constant '0' and a concatenation function symbol instead of a successor function symbol.

Now, is PRA plausible as a candidate for $\mathcal{L}_H$ and T_H? Importantly, would it avoid commitment by Hero to an infinity of abstract sentence types? A parallel can be drawn between an affirmative answer to this question, and the claim—argued for by Tait (1981)—that PRA is the correct formalisation of finitism about arithmetic.[20]

For this to be the case, two requirements have to be met. Firstly, no sentence derivable in PRA must require the existence of infinitely may objects. Secondly, the vocabulary of PRA (and in particular the primitive recursive function symbols) must be intelligible without requiring a commitment to infinitely many objects. It would also be desirable for any sentence of PRA to be interpreted so as not to refer to abstract objects.

The first of these requirements is easily met. Since the language contains no quantifiers, each closed sentence must only refer to finitely many objects (namely, those for which a term denoting them appears in the sentence).

What of the second requirement? One may think that having a large array of function symbols may be problematic, since these symbols will denote functions from all natural numbers/types/possible tokens to all natural numbers/types/possible tokens. Since there are infinitely many of these, it might seem that to understand such a function symbol would require already a commitment to infinitely many objects. Tait (1981) argues that this is not the case, and in fact that the finitistically acceptable functions are precisely the primitive recursive functions. The reason is that for each such function symbol (and its corresponding axioms) 'the finitist can accept it as a construction' (p.533). In the case of considering PRA as a theory of syntax, each such function can be considered as a method of constructing a new linguistic item from existing items.

But could PRA it be interpreted as being only about *concrete* objects? Each theorem will be built up from identities of the form $\ulcorner s = t \urcorner$, with sentential connectives. The question would then be whether each of s and t could refer to a concrete sentence token. Perhaps this is possible. Where such a theorem is inscribed or uttered, each term could refer to *itself*. This would require in addition that '=' is not interpreted as numerical identity, but rather as whatever relation holds between different instances of the same type. There are no doubt many problems with this suggestion, but any attempt to eliminate reference to sentence types is surely going to face such problems. And any attempt to locate a theory of syntax which avoids all reference to abstract objects will likewise face such problems.[21] (Of course, the requirement that T_H be free of commitments to abstracta could just be dropped, and a non-abstractionist epistemology for linguistic types given.)

I will assume that this is more or less correct, and that we can take Hero's theory as being PRA. My arguments which follow would still, however, go through if a weaker theory were adopted. It also seems unlikely that a theory sufficiently stronger than PRA could be motivated.

[20] Adoption of PRA as Hero's original theory results in a situation quite similar in a number of ways to Hilbert's programme. His aim was to start off with finitistic mathematics (essentially, PRA), and use this to justify the acceptability of higher mathematics, where acceptability is the conservativeness (in the strong sense) of higher mathematics over the finitistic mathematics. Hilbert required a finitistic *proof* of conservativeness, thus essentially requiring conservativeness to be *positively* representable, which ultimately put an end to the programme at the hands of Gödel's theorems. That we are only requiring negative representability ensures that the same fate does not face us.

[21] Another alternative might be to go modal, and interpret the theorems of PRA as concerning the *possibility* of the existence of concrete sentence tokens.

4.4.2 Definability and representability in PRA

Now, which sets—and crucially, which potential solutions $S \subseteq \text{AP}$ to the bad company problem—are definable and representable in PRA? Sticking strictly to the requirement that S be definable, the requirement is that there is a quantifier-free formula $\phi(x)$ (with only x free) such that, for any $A \in \text{AP}$:

$$A \in S \text{ iff } \phi(\overline{\ulcorner A \urcorner}) \text{ is true}$$

A consequence of this is that any definable S is also fully representable.[22] Additionally, as a result, the only restrictions S which satisfy this requirement will be those which are primitive-recursive.

As a consequence of such a demand, very few restrictions would be deemed to be acceptable. For example, even simple proof-theoretic consistency is not definable (since then it would be recursive and decidable, which it is not).

However, it seems that some quantified formulas may be used as defining formulas—and as formulas for the purposes of representability—without incurring a commitment to infinitely many objects. Indeed, consistency (which is naturally defined as a universally quantified formula) seems like it should be negatively representable. For suppose some abstraction principle A is inconsistent. Then a proof of its inconsistency could simply be given by giving a proof of a contradiction from A. Such a proof would not require a commitment to infinitely many objects.

And in general, some generalities can be proven with PRA. A proof of $\forall x \phi(x)$ (where ϕ is quantifier-free) will be a proof of $\phi(x)$ (in PRA) where the proof does not depend on x. Similarly, a proof of $\exists x \phi(x)$ will be a proof of $\phi(t)$ for some term t. Tait (1981) argues that these are finitistically acceptable proofs of generalities (although the generalities themselves are not interpreted as being about infinitely many objects, but about giving a method of proving $\phi(t)$ for any particular term t).

Of particular interest will be generalities of the first sort. That is, I shall be concerned with restrictions S which are definable and negatively representable in the sense that there is some quantifier-free formula with parameters $\phi(x, y_1, \ldots, y_k)$ such that, for all $A \in \text{AP}$

(a) $A \in S$ iff $\forall y_1 \ldots y_k \phi(\ulcorner A \urcorner, y_1, \ldots y_k)$ is true

(b) If $A \notin S$ then for some terms $t_1, \ldots, t_k$, $\text{PRA} \vdash \neg\phi(\overline{\ulcorner A \urcorner}, t_1, \ldots t_k)$

There are two things to note about this requirement. Firstly, the first condition (definability) is sufficient for the second condition.[23] Secondly, this requirement corresponds to the requirement that S be *co-recursively-enumerable* (co-r.e.) (in that the complement is recursively enumerable), and, computationally, to the requirement that S be negatively semi-decidable (assuming Church's Thesis). So, there is a decision procedure such that, for $A \notin S$, the algorithm will affirm so, and halt. Alternatively—and equivalently—there is an algorithm which enumerates the abstraction principles not in S.

[22] Proof: Suppose S is defined by ϕ. Now, if $A \in S$, then $\phi(\overline{\ulcorner A \urcorner})$ is true, and since ϕ is just an identity between terms given by primitive recursive functions (or propositional combinations of these), this can easily be verified using the axioms governing those functions. Similarly, if $A \notin S$ then $\neg\phi(\overline{\ulcorner A \urcorner})$ which can also be similarly verified.

[23] Proof: Suppose $A \notin S$. Then, by (a), $\exists n_1, \ldots, n_k \in \mathbb{N}$ such that $\neg\phi(\ulcorner A \urcorner, n_1, \ldots, n_k)$ is true. Then, for similar reasons as before, $\neg\phi(\overline{\ulcorner A \urcorner}, \overline{n_1}, \ldots, \overline{n_k})$ will be provable.

It may also be worth considering higher levels of quantifier complexity, with more nested quantifiers. Recall that a formula in the language of arithmetic is Σ_0 and Π_0 iff it contains no quantifiers. Then, for each n, a formula is Σ_{n+1} (resp. Π_{n+1}) iff it is of the form $\exists x_1 \ldots \exists x_k \sigma$ (resp. $\forall x_1 \ldots \exists x_k \sigma$) where σ is Π_n (resp. Σ_n). So, the claim above is that we should be interested in the definability of a restriction by a Π_1 formula. But, it could perhaps be claimed that formulas of greater complexity could be allowed, although, in this case, the equivalence between definability, representability and computability no longer holds.

4.5 How do restrictions fare?

Now, if I am correct that a restriction must be negatively representable in Hero's theory, and the best candidate for Hero's theory is PRA, how would those proposed restrictions which I consider in the previous chapter fare? That is, which of them are negatively representable in PRA?

In this section, I shall consider these restrictions in turn. Rather than grouping restrictions in the way that I did in the last chapter—in terms of how they are specified—I shall group them by which sets they define. That is, if two proposed characterisations of acceptability define the same restriction (or similar restrictions), then I shall consider them together.

4.5.1 Consistency and satisfiablity

The simplest restrictions which I considered were those closely related to consistency. These were deductive and semantic consistency, as well as satisfiability (which is equivalent to semantic consistency). Are either of the two resulting sets negatively representable in PRA?

We have the following two propositions:

Proposition 4.1. *The set of deductively consistent abstraction principles is co-r.e., and so negatively representable in PRA.*

Proof. Consider the following decision procedure. Consider an abstraction principle A. Run through all possible proofs from A and halt if a proof of $\bot$ is arrived at. If A is inconsistent, then eventually a proof of $A \vdash \bot$ will be arrived at, and so the algorithm will halt. □

Proposition 4.2. *The set of semantically consistent abstraction principles (= the set of satisfiable abstraction principles) is not co-r.e., and so not negatively representable in PRA.*

Proof. Suppose that the set is is co-r.e., so that there is a decision procedure that halts and gives the correct answer if $A \vDash \bot$ (ie. A is not-satisfiable). But then consider the following abstraction principle:

$$(\mathrm{A}_\phi) \qquad \S F = \S G \leftrightarrow \forall x(Fx \leftrightarrow Gx) \vee \phi$$

As Heck (1992) notes, A_ϕ is satisfiable if and only if ϕ is. But then there is a decision procedure which, given any second-order sentence ϕ, will halt if and only if ϕ is unsatisfiable. But this can not be the case, since second-order satisfiability is not decidable. □

Furthermore, the previous proof also shows that satisfiability is not definable by any arithmetic formula of any quantifier complexity (since, if it were, it would mean that the set of second-order satisfiable sentences is arithmetical, which it is not).

We thus have one example of a restriction which can play the role required of it in Hero's epistemology. Alas, consistency, as is well known, is not sufficient to rule out all bad company.

4.5.2 Conservativeness and unboundedness

What of the restrictions based on the idea of conservativeness and the closely related restrictions based on boundedness; will any of these turn out to be negatively representable in PRA? Recall that these restrictions were two kinds of conservativeness (deductive and semantic), as well as boundedness.[24] Recall that conservativeness as stated using semantic consequence is equivalent to the notion of uniform unboundedness.

It turns out that semantic conservativeness is not negatively representable in PRA, for much the same reasons that semantic consistency was not:

Proposition 4.3. *The set of semantically conservative abstraction principles is not co-r.e. and so not negatively representable in* T_H

Proof. Consider some sentence of first-order arithmetic ϕ and let ϕ^* be the second-order sentence which asserts that every structure satisfying the second-order Dedekind–Peano axioms satisfies ϕ (as in the problem of easy mathematical knowledge). If ϕ is a truth of arithmetic, then ϕ^* is a logical truth. If ϕ is not a truth of arithmetic, then ϕ is incompatible with the statement that the Dedekind-Peano axioms are satisfied (and in particular it entails that the universe is finite).

Now, for each such ϕ, consider:

$$(A_\phi) \qquad \S F = \S G \leftrightarrow [\forall x(Fx \leftrightarrow Gx) \vee \phi^*]$$

If ϕ is true, so that ϕ^* is a logical truth, then A_ϕ will have models of every cardinality and hence (since it is purely logical) will be semantically conservative. Similarly, if ϕ is false, then A_ϕ will be non-conservative (since it entails that the universe is finite).

Now, suppose that we have an algorithm which enumerates the non-conservative abstraction principles. By considering only abstraction principles of the form $A_{\neg\phi}$, this will essentially be an algorithm which enumerates the truths of arithmetic. Since this is not possible, there can not be such an algorithm and so the set of conservative abstraction principles is not co-r.e. and so not definable in the relevant sense. □

What about *deductive* conservativeness? It certainly seems that there might be a chance, since it makes use of proof, which is decidable, rather than semantic consequence, which is not. What is needed is that, if an abstraction principle is non-conservative, then Hero must be able to come to know that it is non-conservative. The most natural way to do this, it seems, would be to exhibit some theory T and sentence ϕ such that $\phi^\S$ is derivable from A together with $T^\S$, but ϕ is not derivable from T alone.

The first of these is simple—to show that $T^\S, A \vdash \phi^S$, Hero simply needs to exhibit such a proof (just as in the case of consistency). The second however is not; if there were

[24] I will not consider the 'mixed' kind of conservativeness which I mentioned in section 3.2. I expect that the situation for that will be similar to the situation for SS-conservativeness.

some general method of showing that $T \nvdash \phi$, then there would be a general method of showing that a theory is consistent, which is not possible. So, conservativeness is not clearly co-r.e. Indeed, it is natural to formalise conservativeness as a sentence which has more quantifier complexity than just a universal generalisation. It would be something like:

$$\forall T, \phi[\exists p(p \text{ is a proof of } \phi^{\S} \text{ from } T^{\S}, A) \rightarrow \exists p(p \text{ is a proof of } \phi \text{ from } T)]$$

which, when put in prenex normal form, is a Π_2 type formula, rather than simply a universally quantified formula as required.

This is not to say, however, that there is not a formula of the appropriate kind formalising conservativeness, or that there is not a similar restriction which is definable by a formula of the appropriate kind. We should then see if there is such a formulation. For the sake of simplicity, I shall only consider the case for purely logical abstraction principles, but I expect things will goes similarly in the general case.

In trying to find an appropriate formulation, we might turn to boundedness. Recall that an abstraction principle A is unbounded if, given any cardinality κ, A had a model of cardinality κ. Uniform unboundedness is a modification of this idea to take into account non-logical vocabulary.

Now, clearly unboundedness makes use of set theory in its formulation, so is not definable as-is in $\mathcal{L}_H$. But many cardinality properties are definable in pure second-order logic, and this leaves room to try and construct a version of boundedness which is definable in $\mathcal{L}_H$. If there were some sentence (or class of sentences) which may be taken to mean something along the lines of 'the universe is bounded', then this can be taken to be just as undesirable a consequence of an abstraction principle as a flat out contradiction, and a restriction can be given correspondingly.

Boundedness constraints along these lines, making use only of a consequence relation, can be given as follows. First, we can define the set of sentences that characterize cardinalities precisely:

$$\Gamma = \{\phi : \exists \kappa \forall \mathcal{M}, \mathcal{M} \vDash \phi \text{ iff } |\mathcal{M}| = \kappa\}$$

Then a function from Γ to cardinalities can be given:

$$\text{Card}(\phi) = \text{the } \kappa \text{ s.t. } \forall \mathcal{M}, \mathcal{M} \vDash \phi \text{ iff } |\mathcal{M}| = \kappa$$

For any $\phi \in \Gamma$, let $\phi[F]$ be the result of restricting all the quantifiers in ϕ to F (so $\forall x(\ldots x \ldots)$ becomes $\forall x(Fx \rightarrow \ldots x \ldots)$ and so on). Then, the following can be noted. Given some abstraction principle A (or indeed any sentence) and $\phi \in \Gamma$:[25]

$$A \vDash \neg\exists F \phi[F] \text{ iff } A \text{ only has models with cardinality } < \text{Card}(\phi)$$

Then two variants of boundedness can be given, which I shall call *definable boundedness*:

[25] *Proof*: Suppose that $A \nvDash \neg\exists F\phi[F]$. Then A has a model $\mathcal{M} \vDash \exists F\phi[F]$. So, the domain of $\mathcal{M}$ has a subset of cardinality $\text{Card}(\phi)$. So, $\mathcal{M}$ has cardinality $\geq \text{Card}(\phi)$.

For the converse, suppose that for every $\mathcal{M} \vDash A$, $|M| < \text{Card}(\phi)$. Then no subset of the domain of $\mathcal{M}$ can have cardinality $\text{Card}(\phi)$. So, $\mathcal{M} \vDash \neg\exists F\phi[F]$.

Definition 4.1. A abstraction principle A is semantically/deductively definably unbounded iff

$$\forall \phi \in \Gamma, A \nvdash \neg \exists F \phi[F]$$

according to whether $\vdash$ is taken as deductive or semantic consequence.

These new notions have been motivated mainly by the desire to have a notion related to boundedness and conservativeness which has a better chance of being negatively representable in T_H. We can note the relationship between the various notions of unboundedness as follows (bearing in mind that we are restricting attention to purely logical abstraction principles):

$$\text{Unbounded} = \text{Conservative} \subseteq \begin{array}{c}\text{Semantically}\\ \text{definably bounded}\end{array} \subseteq \begin{array}{c}\text{Deductively}\\ \text{definably bounded}\end{array}$$

What does this say about whether any conservativeness constraint is co-r.e.? A similar kind of decision procedure can be given for testing whether an abstraction principle A is deductively definably bounded as for whether A is inconsistent. Simply run through proofs starting from A, then halt if a proof of $\neg \exists F \phi[F]$ for some $\phi \in \Gamma$ is arrived at. Or course, this requires that it be possible to recognise when some sentence is a member of Γ. That is, Γ must be recursively enumerable.

It is not clear that Γ is r.e., but there are certainly some large subsets of Γ which *are*. George (2006) proves various results about which cardinals are characterisable by explicitly constructing sentences that characterise certain cardinalities, and these explicit constructions can be used to construct a recursively enumerable subset Γ^* of Γ. So, as base members of Γ^* we have sentences ϕ_1 and ϕ_ω such that ϕ_1 characterises 1 and ϕ_ω characterises $\aleph_0$. Then, three recursion rules can be given, generating more members. For each $\phi \in \Gamma^*$, we can construct sentences ϕ^+, 2^ϕ, $\aleph_\phi$ which characterise $\text{Card}(\phi)^+$, $2^{\text{Card}(\phi)}$ and $\aleph_{\text{Card}(\phi)}$ (where $\text{Card}(\phi)$ is considered as an ordinal) respectively. Since these constructions are given explicitly by rules, the resulting set Γ^* will be recursively enumerable. As such, the set of abstraction principles:

$$B = \{A \in \text{AP} : \forall \phi \in \Gamma^*, A \nvdash \neg \exists F \phi[F]\}$$

is co-r.e. as required.

Now, it would be important to know if B as a restriction is restrictive enough so as to rule out inconsistent pairs of abstraction principles. It is not, however.

Clearly all unbounded abstraction principles are in B (since if it is provable that some abstraction principle is bounded, so not in B, it *is* bounded). Since boundedness does not rule out inconsistent pairs of abstraction principles (Weir, 2003, p.27), B will not either. Something beyond conservativeness and unboundedness is needed, such as stability. The question is: is a similar technique possible to construct some co-r.e. restriction related to stability, and which is similar enough that it rules out inconsistency?

Before considering stability, it is worth noting a possible argument against this approach to conservativeness. Hero is not supposed to have any knowledge of mathematics, and in particular any knowledge of cardinality. How then is she supposed to understand that some second-order sentence which is derivable from an abstraction principle says anything about the cardinality of the universe? In order to interpret the

sentences of Γ^* as being about the cardinality of the universe, one needs some specific model theory (and if the model theory were different, such as Henkin semantics, then the sentences of Γ would *not* characterise cardinalities, since Henkin semantics satisfies the Löwenheim–Skolem theorems).

But, just because Hero does not have knowledge of cardinal numbers as *objects* does not mean that she should not have knowledge of cardinality quantifiers. Just as it seems unproblematic that someone with a knowledge of first-order logic automatically has knowledge of first-order definable cardinality characterisations (such as $\exists x \forall y(x = y)$ and $\exists x \exists y \forall z(x \neq y \wedge (z = x \vee z = y))$), it should be unproblematic that someone (like Hero) with a knowledge of second-order logic also has knowledge of cardinality characterisations.

4.5.3 Stability and Irenicity

Finally, we move on to restrictions which do make claim to ruling out inconsistent pairs of abstraction principles, unlike the previously considered restrictions. Recall that an abstraction principle is stable if there is a cardinality κ such that A is satisfiable on all cardinalities greater then κ (strong and weak stability disagree on whether A may have models of cardinality less than κ). Stability succeeds as a restriction in that any set of stable abstraction principles is consistent (and, indeed, stable).[26]

Again, stability is a set-theoretic notion, and so can not be formulated in $\mathcal{L}_H$ in its current form. But is there an equivalent restriction which can be? The answer is no:

Theorem 4.4. *Stability is not co-r.e.n and so is not negatively representable in PRA.*

Proof. The proof is very similar to that of Theorem 4.3. Consider A_ϕ as before, for ϕ a sentence of first-order arithmetic. If ϕ is true of the natural numbers, A_ϕ has models of every cardinality, and so is stable. If ϕ is not true of the natural numbers (so that $\neg\phi$ *is* true of the natural numbers), A_ϕ is satisfied by every finite cardinality, but no infinite cardinality, and so is unstable.

Given an algorithm which enumerates the unstable abstraction principles, an algorithm can be given which enumerates the truths of arithmetic, which is not possible. This is since exactly one of A_ϕ, $A_{\neg\phi}$ will be unstable, and so will get listed. If A_ϕ is unstable, then $\neg\phi$ is true, and if $A_{\neg\phi}$ is unstable, then ϕ is true. □

Similar considerations will also show that stability is not representable with an arithmetical formula of any quantifier complexity.

There are two possible approaches to trying to find a similar restriction which is co-r.e. Firstly, we could consider the notion of irenicity. Recall that an abstraction principle is irenic if it is conservative and consistent with all other conservative abstraction principles. Semantic irenicity defines the same set as stability, which, as has just been noted, is not negatively representable in T_H, and thus does not help. But, since deductive

[26] Linnebo and Uzquiano (2009) claim that it is not successful, since there are proper classes of stable abstraction principles which taken together are inconsistent. Such a scenario will not, however, be a possibility under the present framework. Under the present framework, abstraction principles are taken to be certain sentences of a certain language. Hence, there are not proper class many abstraction principles *tout court*, let alone of some particular type. The way in which Linnebo and Uzquiano construct their proper class of abstracation principles is actually to construct a single instance of what I called an 'extended abstraction principle' in the previous chapter.

consequence is decidable, there may be more chance of the deductive version of irenicity being co-r.e..

Alas, this inherits the problem which faced deductive conservativeness, but amplified. The natural way to formalise such a statement will have more quantifier complexity than is required (more so even than conservativeness). Similarly, a natural thought on what would count as evidence against an abstraction principle being irenic does not appear to be decidable. To show that an abstraction principle is non-irenic, one would need to exhibit another abstraction principle B with which it is inconsistent—that much is simple—but then it must be verified that B is conservative.

A second approach is to find an analogue of the set B, but for stability. An abstraction principle A will be strongly unstable if and only if there are cardinalities $\kappa < \lambda$ such that A has models of cardinality κ but not of λ (or it has no models at all).[27]

To find an analogue of B_4 then, we require two recursively enumerable classes of sentences L and K such that:

$$(\forall \phi \in L)\exists \lambda \text{ such that, if } A \vDash \phi \text{ then } A \text{ has no models of cardinality } \lambda.$$
$$(\forall \phi \in K)\exists \kappa \text{ such that, if } A \vDash \phi \text{ then } A \text{ has a model of cardinality } \kappa.$$

We could then say that an abstraction principle A is *provably unstable* if there are $\phi \in L$, $\psi \in K$ such that $A \vdash \phi$, $A \vdash \psi$ and $\vdash \forall F \forall G(\phi[F] \wedge \psi[G] \rightarrow F \prec G)$, where $\prec$ means that there is a relation which maps the Fs one-to-one into the Gs, but not vice-versa.

Unfortunately, this is not possible. Although there is a suitable L (just take $\neg\phi$ for each ϕ in Γ), there is no such set K. For suppose K is non-empty, and take $\psi \in K$. By *ex falso quodlibet*, $\bot \vDash \psi$, but there is no cardinality κ such that $\bot$ has a model of cardinality κ (since $\bot$ has no models at all).

4.6 Conclusion

So then, none of the restrictions considered in the previous chapter will be successful in both avoiding the problem of mutually incompatible abstraction principles *and* the epistemological problem which I raised at the beginning of this chapter. At one end of the spectrum, consistency satisfies the requirement of being co-r.e., yet badly fails to rule out mutually incompatible abstraction principles. At the other end of the spectrum, stability rules out mutually incompatible abstraction principles, but there does not appear to be a restriction even in the vicinity which is co-r.e.. Conservativeness fares somewhere in the middle. There is a way of modifying boundedness to as to result in a co-r.e. restriction, but this requires weakening an already unsuccessful restriction.

How could the abstractionist find their way out of this problem? One way would be to bite the bullet concerning the conclusion of the epistemological argument.[28] That

[27] Proof that this is in fact equivalent to instability: Suppose A is unstable. If A is satisfiable, then there is a least κ such that A has a model of cardinality κ. Since A is unstable, there must be $\lambda > \kappa$ such that A has no models of cardinality λ (otherwise A would be satisfiable on all and only cardinalities $\geq \kappa$, and hence stable.)

Suppose that A is such that $\kappa < \lambda$, $\kappa \vDash A$ but $\lambda \nvDash A$. Then A is not stable. For suppose that it is, so that for some μ, A is satisfied by all and only cardinalities $\geq \mu$. But if $\mu \leq \kappa$, then $\lambda \geq \mu$ but A is not satisfied at λ. And if $\mu > \kappa$, then A is satisfied at κ, but $\kappa \ngeq \mu$.

[28] I will ignore the possibility of biting the bullet concerning inconsistency, and so accepting the existence of true contradictions.

is, she could decide that one of the proposed restrictions—perhaps stability—does in fact pick out the set of all and only acceptable abstraction principles. Someone like Hero would then be unable to tell, even in principle, which of the abstraction principles which face him are acceptable, and there will be no way of him finding out if he is wrong in his choice. This would lead to a choice between two positions which seem to be to be unsatisfactory. Either the abstractionist programme will have a crucial element of *luck* attached to it, so that Hero (and, ultimately, *we*) has to pick abstraction principles essentially at random, and hope to be lucky in choosing an acceptable one. Or, a degree of relativism may be introduced, whereby multiple starting points (e.g. HP and NP) may be deemed to be acceptable, each of which may give rise to different conceptions of acceptability (see, for example, Sider (2007, pp.26–27), who suggests something along these lines).

A second way out for the abstractionist would be to face the epistemological challenge head on, and seek an restriction on abstraction principles which rules out inconsistency (as stability does), *and* is co-r.e.. Since I have set the problem up in a somewhat formal manner, it might be hoped that there could be a rigorous investigation into whether such a set of abstraction principles exist. The problem is not, however, so simple. It is in fact trivial that there are such sets of abstraction principles; one such example is the singleton set which contains just HP, another would be the singleton set which contains just NP. These are consistent sets, as required, and they are also co-r.e. (and r.e. as well); it is simple to test whether, of a given abstraction principle, that abstraction principle is HP or not, and whether it is NP or not.

Such a restriction is, however, clearly *ad hoc* and not well motivated. A proposal which seeks to avoid both problems will have to be well motivated. And such a condition is probably not easily amenable to formal study.[29]

Given the failure of all of the rather wide range of restrictions to satisfy these conditions, however, I am pessimistic about the possibility of such a proposal being forthcoming. This does not, of course, constitute an *argument* against there being such a proposal. But, absent a concrete proposal of such a kind, it seems to me that the burden of proof lies on the static abstractionist to provide one.

4.7 Towards expansionist abstraction

There is another, more radical, way in which the problem may be avoided. That is to conceive of abstraction in such a way that the bad company problem does not arise at all. If all abstraction principles were compatible with one another, then the set of acceptable abstraction principles could be taken to be all of them. This set will clearly be decidable in the appropriate way.

As I mentioned in chapter 1, the analysis of the bad company problem in terms of inflation suggests that the problem may arise from the assumption that the first-order domain is kept fixed. In the remaining chapters of this thesis, I intend to develop an approach to abstraction—which I call *expansionist abstraction*—according to which

[29] That is not to say that there are not certain conditions, related to well-motivation, which can be given a precise formulation. For example, one may require that a restriction S be *maximal*, in the sense that, for any abstraction principle $A \notin S$, $A \cup S$ is inconsistent. This would certainly rule out the singleton restrictions just discussed.

the first-order domain may expand. I will argue that, on this approach to abstraction, the bad company problem does not arise, and hence the *epistemological* bad company problem does not arise either.

Part II

Expansionist abstraction

Chapter 5

The expansionist account of abstraction

5.1 Introduction

So far, my concern in this thesis has been what I called the *static* conception of abstraction. On this view, Hero's quantifiers range over a fixed domain of objects, which does not change as a result of his laying down of abstraction principles. All that changes is Hero's *knowledge* of the domain.

I argued that the bad company problem which this version of abstractionism faces has not been resolved, and expressed pessimism about the possibility of resolving the problem; any solution which has been suggested either fails to avoid contradiction, or is incapable of playing the role required of it in Hero's epistemology.

My aim for the remainder of this thesis is to develop an alternative way of looking at abstraction—an *expansionist* conception—that does not suffer from the bad company problem. My aim is relatively modest; I do not wish to claim to refute the static conception of abstraction, so that abstractionists are compelled to adopt the expansionist conception of abstraction. My aim is instead simply to provide an alternative conception of abstraction. The two conceptions can be assessed on their relative benefits and weaknesses. I would then claim that, assessed on these benefits and weaknesses, the expansionist conception comes out the stronger, in major part due to problem of bad company. The expansionist approach will, of course, face its own difficulties, and a major goal of the latter half of this thesis will be to face down those difficulties.

There are a couple of aims for this brief chapter. Firstly, I will recap and expand upon the characterisation of expansionist abstraction which I gave in chapter 1. Secondly, I wish to highlight the main questions that must be answered for the expansionist position to be viable.

5.2 Expansion and impredicativity

Expansionist abstraction is, to put it briefly, the claim that abstraction principles may serve to allow someone to expand their domain of quantification. It shares with static abstraction the claim that abstraction principles can be used to define an abstraction operator, and along with it the abstract terms which result from the application of this operator. It differs from static abstraction on issues regarding quantification and, in particular, how the abstract terms interact with the quantifiers. Static abstractionists claim that, if these abstract terms have referents, then these lie within the range of the same quantifiers by which the abstraction principle is stated. As such, new knowledge may be gained about this particular domain of quantification. Or, to avoid talk of domains, an abstraction principle allows one to gain new knowledge of propositions expressible using the same quantifiers as were used in stating the abstraction principle. As such, on the static picture, abstraction principles are *impredicative* definitions—they define abstract terms by means of quantification which involves the referents of those terms.[1]

Expansionist abstraction drops the assumption that the referents of abstract terms must lie within the domain of quantification used in the abstraction principle. Instead, it may be the case the referent of an abstract term may lie within a different, *wider* domain. Thus abstraction principles do allow one to gain new knowledge, but this new knowledge will consist in knowledge about a possibly wider domain of quantification. Or, again to try to avoid talk of domains, it allows one to gain new knowledge, but expressible only by means of interpreting the quantifiers differently, or by using the quantifiers in a different context. It also provides one with the means to interpret the quantifiers in such a way.

But, although expansionist abstraction stands in contrast to the impredicativity of static abstraction, it does not amount to the claim that abstraction principles must be understood in a *predicative* manner.[2] This would be the claim that, for any abstract term, the referent of that term does *not* lie within the domain of quantification of the quantifiers used in the abstraction principle.

There are, instead, two main differences between expansionist abstraction as I shall see it and predicative abstraction, simply understood. The first is that I will not insist that an abstract term must not lie within the domain of the quantifiers in terms of which it is defined. Instead, it will be left open whether this is the case or not. Consequently, it is left open whether the domain of quantification which results from an abstraction principle is indeed different from the domain in terms of which the abstraction principle is stated. In some cases, however, it will be possible to show that some referents of abstract terms *must* lie outside the original domain, and thus that there is a domain expansion.

Secondly, predicative abstraction has normally taken to be a one-off affair. That is, an abstraction principle is laid down, with resulting abstract terms which refer to 'new' objects, and that is the end of the matter. But it is possible to conceive of this

[1] This should be distinguished from another dimension of impredicativity in the abstractionist programme—in the second-order logic. The second-order comprehension principle is impredicative in that concepts may be defined by means of quantification over all concepts.

[2] It has been claimed by some that abstraction must be understood in a predicative manner by (e.g. Potter and Sullivan, 2005).

process being iterated, so that we quantify over these new objects, and then lay down the abstraction principle again resulting in yet more new objects. This process may then be continued indefinitely.[3] The version of expansionist abstraction which I intend to give is designed to allow such iteration.

5.3 Some case studies: HP, BLV and NP

It will be useful to look at some examples of roughly how expansionist abstraction may work, partly in order to make the idea clearer, but also partly because doing so will highlight some issues that may arise. The examples that I will look at are HP, BLV and NP.

5.3.1 HP

Consider Hero, who we are supposing to have no knowledge of mathematics, and no understanding of mathematical vocabulary. Suppose that his quantifiers range over a domain D. For the moment, nothing needs to be assumed about D; it may be severely limited (to say, just physical objects), or even absolutely unrestricted (although we shall see that there is a problem with absolutely unrestricted domains later).

Now, suppose that Hero lays down HP as an abstraction principle, in order to gain knowledge of the *number of* operator N. The effect is that N maps concepts F and G onto the same object if and only if (from our point of view), an equal number of the objects in D fall under F and G. Hero gains knowledge of these objects, and thus of a domain of quantification D' which ranges over them.

The first thing to note is that it is left open whether D' is more expansive than D. In particular, there is nothing to say that Hero was not already quantifying over the natural numbers, unbeknown to him. Such would be the case, presumably, were D absolutely unrestricted.

Secondly, in contrast to the static case, it need not be the case that D' is infinite, so that Hero gains knowledge of infinitely many objects. For consider the case where D is finite, containing n objects. Then, as noted in 1.3.2, there will be $n + 1$ numbers which Hero comes to know of (0 to n), and thus which lie in D'. Even if D is infinite, it will only be the case that Hero come to *know* that there are infinitely many objects in D' if he knows that there are infinitely many objects in D. In particular, the proof that there are infinitely many numbers will not be available to Hero, since that assumes that HP is impredicative.

But this does not mean that HP will only ever allow Hero to come to know of one additional object. For suppose that Hero knows that there are at least m objects in D; let us suppose that he names them a_1 to a_m. Then he will be able to prove that there are $m + 1$ numbers in D'. But he can then iterate the process, resulting in D'', in which Hero knows there are at least $m + 2$ numbers and so on. A problem will, however, arise at this point. The situation seems to be the following: given any natural number n, Hero is able to iterate HP so that he knows of n natural numbers. As such, there is a kind of *potential* infinity of natural numbers which he has access to: no matter how many

[3] There is an issue concerning whether this process could be continued, not just indefinitely, but genuinely *infinitely*, so that we reach transfinite stages of the process. I shall return to this issue later.

natural numbers he requires for some purpose, he will be able to iterate to get them. But it does not seem that he can iterate enough so as to get an *absolute* infinity of natural numbers all at once, at least, not without performing some kind of supertask. This is an issue that I will return to in later chapters.

5.3.2 BLV

The case of BLV is much the same as for HP, and many of the same issues still arise. In particular, the same issue concerning infinite iterations will arise. In the case of set theory this problem is particularly pressing. For it may be thought acceptable for Hero's knowledge of the infinity of the natural numbers to be merely potential, since it would at least be the case that, for any natural number, it is possible for Hero to expand his domain to include such a number. But the analogous situation in set theory is that, for any *hereditarily finite* set, Hero may expand his domain to include it. But this, it seems, is manifestly not enough. If there is to be an abstractionist set theory, it must at the very least be the case that Hero can expand his domain to include infinite sets.

Another major difference concerns the possibility of starting out with an absolutely unrestricted domain. This time, the abstraction operator will map concepts F and G onto the same object if and only if the same objects in D fall under F as fall under G. This will mean that—again, as noted in chapter 1—there will be 2^κ resulting abstracts if there are κ objects D.

So, the resulting domain, D' can not be the same as as D, but must be a proper expansion. Likewise, iterating this process results in a sequence of domains $D \subsetneq D' \subsetneq D'' \subsetneq \ldots$. Consequently, it seems that there is no possibility of D being the absolutely unrestricted domain. This raises a potential conflict, which I will discuss in section 5.4.

5.3.3 NP

Finally, something should be said about other examples of bad company, such as the nuisance principle NP. If the expansionist approach is to avoid the bad company problem, it must avoid any problems arising from these as well as flat-out inconsistent abstraction principles such as BLV. That is, it needs to be the case that the combination of, say, NP and HP is does not result in a contradiction.

Although I will not be saying much about this issue in the coming chapters, I will merely note why an analysis in terms of inflation suggests that no such problem arises. The mutual unsatisfiability of NP and HP arises from the fact that the two abstraction principles put different demands on the domain; HP requires the domain to be infinite, and NP requires that it be finite. And this can be accounted for in terms of inflation: HP inflates on on finite domains, since, given κ objects, it requires that there be $\kappa + 1$ numbers, and, for finite κ, $\kappa < \kappa + 1$. Likewise, NP inflates on infinite domains, since given an infinite domain of cardinality κ, it requires that there be 2^κ nuisances (see Wright, 1997).

But this will not be a problem on the expansionist account; if an abstraction principle requires that there be more abstracts than there are objects, the domain may simply expand so as to accommodate them. The resulting effect will be that, in finite domains, HP will drive domain expansion, and in infinite domains, domain expansion will be driven by NP.

5.4 Expansionist abstraction and absolute generality

Say that *generality absolutism* is the view that it is possible, in some sense, to speak about (i.e. to quantify over) absolutely all objects at once. *Generality relativism* by contrast is the opposing view. Generality relativists claim that it is not possible to quantify over absolutely everything at once. More needs to be said about the precise content of this distinction (with which I will be concerned in chapter 6). But hopefully the idea has enough intuitive appeal for present purposes.

I have not characterised distinction between static abstractionism and expansionist abstractionism in terms of absolute generality, but there clearly is a relationship. I have already mentioned, albeit briefly, that one important motivation for the static approach is the assumption of the availability of an unrestricted domain. The reason that we may take the domain to be static, and that we may take abstract terms to refer to objects within this fixed domain, is that the domain is the absolutely unrestricted one (for obviously then there would be no possibility of expansion).

This raises the question: Is there a similar, but converse, relationship between the expansionist account and absolute generality? That is, does the expansionist account required generality relativism as a motivating assumption? As mentioned in the previous section, there is *prima facie* reason to think that it may do.

One way that it might have been thought that expansionist abstraction is not committed to generality relativism would be that, since there is no requirement in general that the domain properly expands, the view is compatible with $D_1 = D_2 = \ldots$, with each of these domains being absolutely unrestricted. Then the expansionist position would neither be committed to generality absolutism (since there is no requirement that the domain must not expand) nor to generality relativism (since there is no requirement that is must expand).

But the properties of BLV mentioned in the previous section show that this idea is not tenable. For although there is no requirement in general that the domain must expand, it seems that the inflationary nature of BLV shows that it is required in some cases. (As I mentioned in the discussion of inflation in chapter 1, it may be possible to resist this, since the argument depends on particular features of the metatheoretic viewpoint that may be resisted. I will show later that, just as the inconsistency of BLV on the static approach can be replicated in the object language of abstraction, so too can the relationship between BLV and generality relativism.)

So, the expansionist account (if it is to accommodate abstraction principles such as BLV) must have it that each D_i is not the absolutely unrestricted domain. But there is another way in which one may separate the expansionist account from generality relativism. If there were a way to demarcate absolute domains from non-absolute domains, then it could perhaps be claimed that abstraction principles are only to apply to non-absolute domains. This strikes me, however, as problematic. There are two questions which the proponent of such a position must answer. These are, firstly, how is it that we are to demarcated two types of domain, and secondly, *why* is it that abstraction principles should not apply to absolute domains? Unless such questions can be answered (and it seems unlikely to me that they can be) then this approach too will not be feasible.

So then, it looks like the expansionist abstractionist is committed to generality relativism. This raises somewhat of an obstacle to expansionist abstraction. For generality

relativism faces a number of challenges itself. These challenges must then be met by the expansionist abstractionist if it is to be a viable position. My main aim in chapter 6 will be to clarify what these challenges are (especially as they apply to expansionist abstractionism) and to respond to them. My response to these challenges will actually be very similar to the response that static abstractionists give to potential challenges to their position, in that it makes essential use of the context principle.

5.5 Where do we go from here?

There are a number of issues which clearly need to be addressed by the proponent of expansionist abstraction. My aim for the next few chapters will be to address such issues. What is required?

My characterisation of expansionist abstraction in chapter 1 and this chapter has solely been an external one. This is in contrast to how static abstraction has been developed. On the static approach, abstraction principles have had a status much the same as axioms in an axiomatic system, and we have been interested in the axiomatic systems which result, such as Frege Arithmetic. Such an approach makes sense from the point of view of the epistemological aims of abstractionism. The abstraction principles are sentences which Hero knows through stipulation, and the theorems of the resulting axiomatic system are sentences which Hero can come to know by deduction. The external viewpoint can still be useful, but mainly for the purposes of various metatheoretic results concerning, for example, what is possible in such an abstractionist system.

Such an internal characterisation has been notably absent from my discussion of expansionist abstraction so far. I have not suggested an axiomatic system which can be seen as the one that Hero occupies. Some kind of axiomatic system is needed if there are to be any epistemological claims concerning the approach. That is, we need to identify certain basic principles, and give a story about how Hero could come to know them. Then, some deductive system must be identified, and it must be claimed that Hero can make use of this deductive system to extend his knowledge.

So, the first task is to develop such a deductive system. This will be my aim in chapter 7. Given such a deductive system, there will be further questions that may be asked from the external perspective. In particular, we would then be able to ask more precisely the question about whether the bad company problem can be avoided. In particular, we want to know whether the resulting axiomatic systems are consistent. We also want to know how much mathematics can be developed in such a situation. I.e. we want to know how strong a mathematical theory can be interpreted in such a theory.

These tasks are more or less purely technical. But there are also important philosophical questions that must be answered for expansionist abstraction to be viable. In particular, generality relativism must be defended, and particularly the form of generality relativism which is involved in expansionist abstraction. There are a number of issues that the generality relativist needs to explain. They must, for example, explain the metaphysics of domain expansion, and explain how abstraction principles might serve to bring one about. All this should be done without appealing to any mysterious ideas, such as the creation of mind-dependent objects. I will address these issues in chapter 6.

Finally, there is the issue of how to account for the apparent need to iterate domain expansion into the transfinite. This issue will arise in a more technical form in chapter 7,

in which the lack of such infinite iterations forms an obstacle to developing a particularly powerful system. My solution in that chapter is to introduce a *reflection principle*, which, I claim, has the same effect as an infinite iteration. I explore the technical consequences of such a principle in 7, and then discuss the possible motivation for such a principle (and some issues arising from it) in chapter 8.

Chapter 6

What do generality relativists need to explain?

As I argued in the previous chapter, the expansionist position will likely entail *generality relativism*, the view that in some sense it is not possible to quantify over *absolutely everything*. The aim of this chapter is to answer the first of the challenges raised at the end of the previous chapter—to defend generality relativism as a coherent position. There are two parts to this challenge. Firstly, it has been claimed by some that generality relativists can not even state their position without falling into self-contradiction. Secondly, even if a position can be stated, much more needs to be said so that the position does not seem utterly mysterious. The relativist must claim that domains of quantification must 'expand' in some way or other, and as such they owe an explanation of the metaphysics of domain expansion. In absence of such an explanation, the view risks being to committed to something like the claim that domain expansion involves something like the literal *creation* of objects.

Sections 6.2 and 6.3 concern the first of these challenges. In section 6.2 I shall articulate the challenge, and argue that, rather than being a problem purely for relativists, it is also a problem for their opponents—*generality absolutists*. In section 6.3, I will argue that the challenge can be met by formulating relativism in a modal language, which is a view argued for by Kit Fine. In contrast with Fine, who claims that the relevant 'postulational' modality must be taken as primitive, I argue that it may be explained in terms of other, more commonplace modalities. I then give such an explanation.

Sections 6.4 and 6.5 concern the metaphysics of domain expansion. I shall argue that the idea of moving to a more inclusive domain from a less inclusive domain can be made unmysterious by adopting a viewpoint according to which talk of *domains* is secondary to talk of the truth-conditions of sentences containing quantifiers. As such, the view arrived at relates to quantifiers and quantification in much the same way as orthodox neo-Fregeanism relates to singular terms and singular reference.

6.1 Introduction—Relativism and absolutism

Although the reason that I need to defend generality absolutism is that it is a consequence of expansionist abstraction, others have given independent reasons for accepting the view. It will be useful to very briefly survey these arguments here.

Various writers (e.g. Fine, 2006; Glanzberg, 2004, 2006) have claimed that generality relativism follows from Russell's Paradox.[1] The reasoning goes roughly as follows: Given some purported use of absolutely unrestricted quantification, we can consider a particular object—namely, the Russell set $\{x : x \notin x\}$—which must not lie within the range of the quantifier, on pain of contradiction. Hence, the purportedly absolutely unrestricted quantifier is not absolutely unrestricted after all. More fully, the argument (the following version is Fine's) goes like this:

> Let us use '$\exists$' and '$\forall$' for those uses of the quantifier that the universalist takes to be absolutely unrestricted. The critical step in the argument against him is that, on the basis of his understanding of the quantifier, we can then come to another understanding of the quantifier according to which there will be an object (indeed, a set) whose members will be all those objects, in *his* sense of the quantifier, that are not members of themselves. ... From [this], we can derive the extendibility claim:
>
> (E) $$\exists^{+} y \forall x (x \neq y)$$
>
> ... But the truth of (E) shows that the original use of the quantifiers $\exists$ and $\forall$ was not absolutely unrestricted. (Fine, 2006, p.22)

A similar argument, due to Williamson (2003), but which does not make reference to sets, concerns instead quantification over *interpretations* of a language. Suppose we have a candidate for an absolutely unrestricted quantifier. For a given predicate letter P in the language and any definite condition on objects, we must be able to consider an interpretation I of P so that *for all* objects x, P applies to x if and only if x satisfies that condition. Now, consider the condition which any x satisfies if and only if x is not an interpretation under which P applies to x, and consider the interpretation I which associates P with such a condition. On pain of contradiction, I must not fall under the quantifier involved in its own definition. Hence, the quantifier in question must not be absolutely unrestricted after all.

A third version of the argument may be given, using ordinals rather than sets or interpretations. The key step is in noting that, since *all* the ordinals are well ordered, they too should have an ordinal. Then, on pain of the Burali-Forti paradox, this will not lie within the range of the current quantifiers.

For the moment, I shall be unconcerned with whether these particular arguments are successful or not. Before individual arguments themselves can be assessed, a generality relativist already faces a number of challenges. Firstly, the relativist must develop the resources necessary to even state her position. In section 6.2 I shall show why a problem arises (and why it is also a problem for the absolutist), and consider some of the solutions which have been suggested—namely, introduction of a primitive modality and appeal

[1] As such, their argument is very similar to the kind of reasoning which shows that the expansionist position leads to generality relativism.

to 'open-ended' schemes. In section 6.3 I develop further the modal solution, filling in what I see as a number of deficiencies of the solution as offered by its main proponents.

Secondly, the relativist owes an explanation of the metaphysics of domain expansion. In section 6.4 I shall argue that such an explanation has not been sufficiently given. Various broadly fictionalist approaches suggest themselves (whereby the domain expansion is only an expansion according to some fiction). Such a fictionalist approach is however clearly unacceptable for the aims of the relativist. I shall instead make an alternative suggestion, whereby what happens when a domain expands is a shift in the meaning of the quantifiers, which is not to be explained in terms of antecedently understood domains. I shall defend this approach from various objections.

6.2 Articulating relativism

The first issue that must be addressed is how the relativist can best articulate her position. It clearly can not be stated in the same way as it naïvely is at the beginning of this chapter. That is, as:

(GR_1) It is not possible to quantify over absolutely everything.

For this characterisation *uses* absolutely unrestricted quantification! Indeed, it is trivially the case that it is possible to quantify over everything, since, by definition, *everything* is what we quantify over. This is not so much an *argument* against the relativist, however. That (GR_1) is trivially false must instead simply be an indication that it is not what generality relativism is getting at. For the same reason, the absolutist position can not simply be 'it is possible to quantify over absolutely everything'. Either this presupposes whatever the position is, or it states a mere tautology. Even in clearly non-absolute domains (such as the one in use when somebody about to embark on holiday asks 'is *everything* packed?'), the claim 'I am quantifying over everything' is trivial.

Without a clearly articulated position, there is nothing for the relativist and absolutist to dispute. Therefore, the need to develop the resources needed to articulate the relativist position is something that should also concern the absolutist. Once an adequate formulation of relativism has been come to, the absolutist position can be understood simply as the negation of relativism. Similarly, given an adequate articulation of absolutism, relativism should simply consist in the negation of this position.

Of course, if the absolutist is permitted to make use of absolutely unrestricted quantification, they can state their position in a similar way to (GR_1). It would simply be, in this case, 'it is possible to quantify over *absolutely everything*'. Various writers have claimed that this is all that is needed on the part of the absolutist (e.g. Williamson, 2003). But this will not do. What is at issue is whether some linguistic device with certain properties is in good standing, and what is needed to articulate the two positions is some characterisation of these properties. But it is surely inadequate, or at least highly unsatisfactory, to attempt to *use* the very notion itself in characterising its behaviour. Suppose that somebody puts forward some new linguistic device—a new logical connective, say—and then refuses to explain what she means by this new connective other than by using the connective itself (say, by giving a homophonic semantics). In this situation such an explanation would be clearly lacking, and there would be no way to conduct a debate about whether the new connective is in good standing or not.

To give an example, suppose that somebody starts to use the word 'schtonk' as a logical connective, and then, when asked what it means, says that 'ϕ schtonk ψ' is to be true just in case ϕ is true schtonk ψ is true. That is, 'schtonk' means schtonk! And suppose then that we are to decide whether schtonk is in good standing as a connective (in the same way as, say, 'and' and 'or' are), or whether it is not (in the same way as 'tonk' is not). Clearly we do not yet have enough information in order to decide.

Why should the situation be any different in the current case, where the linguistic device in question is the absolutely unrestricted quantifier? It might be claimed by the absolutist that the situation is different since, unlike 'schtonk', absolutely unrestricted quantification is used all the time in everyday speech. But this too will be unsatisfactory. For one thing, absolutely unrestricted quantification is uncontroversially *not* in use all the time in everyday speech; the vast majority of uses of quantification are clearly contextually restricted. And in the rare other cases (e.g. in metaphysics), it is precisely whether the quantification involved is absolutely unrestricted which is at issue between the relativist and the absolutist. To simply take it that in these cases the quantification is absolutely unrestricted is simply to beg the question against the relativist.

So then, in order to get to the heart of what is at issue in the debate, a formulation of each position must be articulated in a neutral manner, that is, in a language that both the relativist and the absolutist can speak and accept as intelligible. Of course, in this situation, although each side will find the other's position intelligible (in that is will be a perfectly meaningful sentence of the language), they will regard it as (necessarily) false.[2]

Nonetheless, it seems that the onus should be on the relativist to provide a characterisation, since it is they who are putting forward the claim to be discussed.

One option that may be considered is to give more force, somehow, to the modifier 'absolutely'. As it is, it is unclear how 'absolutely everything' is to differ from 'everything' simpliciter. Similar problems, however, seem to arise. Williamson (2003, p.416) points out that even the full phrase 'absolutely everything' may be contextually restricted in the usual way; someone may exclaim (in the same situation as before) 'Of course I'm late—you left me to pack ABSOLUTELY EVERYTHING!'. In any case, there would still be the need to articulate clearly what precisely is meant by 'absolutely'.

A more neutral characterisation of the two positions may be arrived at by considering more carefully what the extendibility arguments purport to show. The claim is that, whatever quantifier is being used, or whatever context the quantifier is being used in, it is always possible to arrive at a new quantifier meaning which is wider in scope. This can be expressed more formally, by indexing quantifiers to contexts—so that '$\exists_C$' expresses the existential quantifier as used in context C—and allowing quantification over contexts[3]:

$$(\mathrm{GR}_{\mathrm{contexts}}) \qquad \forall C_0 \exists C_1 \exists_{C_1} x \forall_{C_0} y (x \neq y)$$

[2] A distinction needs to be drawn here between a relativist finding *absolutism* intelligible, and finding *absolutely unrestricted quantification* intelligible. The first concerns whether a certain position is an intelligible position—and it surely shall be if a neutral formulation can be arrived at (since it is just the negation of a position that the relativist herself accepts). The second concerns whether a certain linguistic device is intelligible.

[3] Similar possibilities are considered by Williamson (2003) and Fine (2006), and deemed unsuccessful for the same reasons. What follows in considering alternatives follows approximately the same line as these two papers.

This needs to be distinguished from a similar (natural language) formulation considered by Williamson (2003, p.430):

(6.1) For any context C_0, there is a context C_1 such that not everything that is quantified over in C_1 is quantified over in C_0.

As Williamson points out, this is immediately self-refuting in the same way as (GR_1) is. For (6.1) must be stated in some context C. Then, instantiating C_0 in (6.1) by the context of utterance, C, it follows (expressed in C) that

(6.2) There is a context C_1 such that there is something which is quantified over in C_1 which is not currently being quantified over.

Instead, ($GR_{contexts}$) should be compared to the following formulation, which Williamson also considers:

(6.3) For every context C_0, there is a context C_1 such that 'Not everything is quantified over in C_0' is true as uttered in C_1 (where 'C_0' as uttered in C_1 refers to C_0)

which is not clearly unacceptable.

Now, is ($GR_{contexts}$) acceptable as a neutral formulation of relativism? Unlike (GR_1), it is not trivially false. Nor does it make use of absolutely unrestricted quantification—the quantifiers are *restricted*, to *contexts*. However, the extendibility arguments, if successful, show not merely that quantification over absolutely all *objects* is somehow illicit, but that even quantification restricted to certain sortal concepts—namely, over sets and over interpretations—suffers similarly. That is, given some quantifier purportedly over absolutely all Fs, we can specify some F which does not fall under that quantifier.

Without a guarantee that this is not the case for *contexts* themselves, there is similarly no guarantee that ($GR_{contexts}$) successfully expresses what it aims to express. Indeed, the concept of context seems to be precisely the kind of concept which exhibits such behaviour, given the similarity between contexts and interpretations.

Fine (2006) claims that this is indeed the case. Under the assumption that the quantifier over contexts is absolutely unrestricted, it is claimed that it follows that there is a context for which the quantifier is absolutely unrestricted. Fine's argument is as follows. Consider a context C_0 where we interpret the quantifier as follows:

$$\exists_{C_0} x\phi \text{ iff } \exists C \exists_C x\phi$$

But then, the existence of C_0 contradicts ($GR_{contexts}$), since it follows that

$$\forall C_2 \forall_{C_2} x \exists_{C_0} y(x = y).$$

Now, this relies on the premise that there *is* such a context, and it is not entirely clear why there should be. Nonetheless, with no guarantee that there is no such context, or no similar argument to be made, formulating relativism as ($GR_{contexts}$) is rather tenuous. It would be desirable to find some better way of expressing relativism which is less susceptible to suspicion. There are two main approaches that have been suggested in the literature. The first is to state the position in some schematic form, where the relevant schema is not to be understood as the generalisation over its instances. This is closely related to the idea of a statement being given with 'typical ambiguity'. The second approach is to adopt a primitive modality in place of the quantifier over contexts.

6.2.1 Schemes and systematic ambiguity

The problem with ($GR_{contexts}$) is that we can not be sure that, in the context in which it is uttered, the quantifier over contexts is neither restricted, nor extendible. In shifting to a new context, we may include more contexts within the range of our quantifiers, and, for all ($GR_{contexts}$) tells us, an utterance of ($GR_{contexts}$) in this new context may be false. What is desired is that it be stated that ($GR_{contexts}$) be true *as uttered in any context*. Clearly, simply stating such a generalisation ($\forall C'((GR_{contexts})$ is true as uttered in $C'))$ will not do; the problems will simply reemerge. Instead, the schematic approach suggests that we do away with attempting to generalise over contexts by use of quantification, but instead simply lay down:

($GR_{schematic}$) $\exists C' \exists_{C'} x \forall y (x \neq y)$

where this is understood as being *systematically ambiguous*. That is, it is recognised as being true no matter what context it is uttered in.[4] Importantly, however, this should not be understood as being the same as asserting the universal closure (over contexts); the universal closure over contexts is just ($GR_{contexts}$), which we have seen is unacceptable.

This approach does, however, suffer from a number of weaknesses. For one thing, it would be desirable to say more about what a systematically ambiguous sentence is supposed to amount to. What (if anything), for example, is being asserted when a scheme is laid down as being systematically ambiguous? And, if there is nothing being asserted, what is the nature of this non-assertorial speech act which is involved?

Secondly, systematic ambiguity comes at the expense of expressive power. We can not, for example, negate a systematically ambiguous statement. This is a particular problem if we wish to characterise absolutism as the negation of relativism. Nor can we embed a systematically ambiguous statement into other sentences. I believe that the next and final approach that I will consider will avoid such difficulties by, in effect, generalising the notion of systematic ambiguity.

6.2.2 Postulational modality

The second possibility is to introduce a modality, roughly representing the 'possibility of extending the domain'. This is suggested (though eventually rejected) by Williamson (2003, p.431), and taken up fully by Fine (2006). Fine states the idea behind using a primitive modality as follows:

> Under the modal formulation of the limitivist [relativist] position, we take seriously the thought that any given interpretation *can* be extended, i.e. that we can, in principle, come up with an extension. Thus, in coming up with an extension we are not confined to the interpretations that fall under the current interpretation of the quantifier over interpretations. (p.30)

So, $\ulcorner \Diamond \phi \urcorner$ expresses that interpretation of the quantifiers occurring in ϕ can be extended so that ϕ is true, and similarly, $\ulcorner \Box \phi \urcorner$ expresses that ϕ is true no matter how the quantifiers are interpreted. Fine calls the modality in question *postulational possibility*.

[4] The systematic ambiguity approach is adopted by, amongst others, Glanzberg (2004), Hellman (2006) and Lavine (2006)

This then results in something like a generalisation of systematic ambiguity in the quantifiers. Since $\Box\phi$ says, in effect, that ϕ is to be taken to be true no matter how the quantifiers in it are interpreted, this has the effect of laying down ϕ in a systematically ambiguous manner.

A way of expressing the relativist position modally does not fall out of this simply. What is required in order to say that the current quantifier is extendible is that it is postulationally possible that there be an object which is not under the range of the *current* (or perhaps, actual) quantifier. This would require that something like an actuality operator be added to the language, and can then be expressed as

$$\Diamond\exists x(@\forall y(y \neq x)).$$

However, this will not be enough; it simply states that the *present* domain of quantification is extendible. A stronger claim is needed to the effect that this is postulationally necessary, but with $\forall y(y \neq x)$ being evaluated outside the scope of the possibility operator (as above), but within the scope of the outermost necessity operator. This would require enhancing the language with a more sophisticated method of indicating scope. Such a method is supplied by an extension of modal logic given by Hodes (1984b), who adds an operator $\downarrow$, whose effect is to exempt what follows from the scope of the *innermost* modal operator.[5] Then, generality relativism can be expressed as:[6]

(GR) $$\Box\Diamond\exists x(\downarrow\forall y(y \neq x))$$

The question immediately arises as to how this new modality should be understood. A suggestion that may present itself is that the modality should be understood as a quantifier over meanings, interpretations, contexts or domains of quantification. So, $\Diamond\exists x\phi$ means that *there is* an interpretation I of the quantifier so that $\exists x\phi$ is true under that interpretation. Indeed, Fine considers a move like this before introducing modality. But this would mean that the modal approach simply collapses to one similar to that of ($\mathrm{GR}_{\mathrm{contexts}}$). And we have already seen that such an approach will fail.

Fine goes on to reject a number of possibilities of how to understand the modality in terms of other, more familiar, concepts. So, postulational modality is not to be understood as a 'circumstantial' modality, like metaphysical and physical modalities. It concerns instead 'possibilities *for* the actual world, and not merely possible alternatives *to* the actual world' (p. 33).

Nor, Fine claims, can the modality be defined in terms of more familiar modalities. So, for example, the postulational possibility of a proposition can not be defined as 'the metaphysical possibility of our specifying an interpretation under which the proposition is true' (p.34).

So, how can postulational possibility be explained? Fine seems to claim that an understanding of such a possibility is implicit in the possibility of the domain expanding by means of demonstrating the existence of a Russell set, as in his original argument against absolutism:

[5] Alternative scoping operator could be used, such as those discussed by Parsons (1983b, Appendix).

[6] Fine does not go down this route, but instead uses quantification over interpretations to allow the claim to be expressed. Introducing quantification over interpretations seems to me to be more problematic than this minor extending of the expressive power of the modal logic.

> [I]t seems clear that there is a notion of the required sort, one which is such that the possible existence of a broader interpretation is indeed sufficient to show that the narrower interpretation is not absolutely unrestricted. For suppose someone proposes an interpretation of the quantifier and I then attempt to do a 'Russell' on him. Everyone can agree that if I succeed in coming up with a broader interpretation, then this shows the original interpretation not to have been absolutely unrestricted. Suppose now that no one in fact does do a Russell on him. Does that mean that his interpretation was unrestricted after all? Clearly not. All that matters is that the interpretation should be possible. But the relevant notion of possibility is then the one we were after. (pp.34–5)

I fail to see how this could amount to an explanation of the new modality.[7] Leaving aside for the moment what an *interpretation* of a quantifier supposedly is—which Fine deals with later—it is unclear how an explanation of the relevant kind of interpretation is supposed to lead on to an explanation of the modality. Firstly, a supporter of absolutism will presumably not accept that someone, in attempting to 'do a Russell', *can* succeed. For the absolutist will refuse to acknowledge that from our understanding of sets and the quantifier we can come to an understanding of a quantifier so that it includes in its range a universal set (that is, universal with respect to our current quantifier). Of course, it is the conditional claim that *if* a Russell manoeuvre is possible *then* this would show that the domain is not absolutely unrestrictedthat needs to be accepted, and this may be accepted without accepting the antecedent. Nonetheless, surely the sense in which it is required that such a specification of an interpretation be possible is a more common metaphysical (or perhaps even physical) possibility. Or, at least, Fine gives us no reason to suppose that this it is not.

Fine thus intends that the modality must be understood primitively. But stipulating that postulational modality should be understood primitively is problematic in a way which it is not for circumstantial and other modalities. In these cases, such as physical and metaphysical possibility, it is at least plausible to take them as already understood. It is also reasonable to make use of them without first requiring further explanation of them. The reason is that they are (at least apparently) in common usage, and so therefore presumably understood by competent speakers of a language.[8] But the same can not be said of postulational modality, which only *explicitly* features in Fine's work. Fine claims (Fine, 2005, p.108) that it is actually widespread, for example where apparent postulation is used in mathematics. But the burden presumably is on him to show this, and that these uses are not best accounted for either by eliminating any apparent modality, or by means of other, more familiar, modalities. For this reason, if postulational modality is to play a crucial role in discussion, it seems reasonable to first demand that some explanation of it in other terms be given. In any case, if an explanation of postulational modality can be given in more primitive terms, then surely it is desirable to do so.

[7] Perhaps it is not intended as a full explanation, but simply as a partial explanation. Nonetheless, surely something more than a partial explanation is required.

[8] This is not to say that they must be taken as primitive altogether, so that they cannot be explained in terms of something more simple. That would be an extremely controversial claim, and not one that I wish to commit myself to. All I intend to claim is that such an explanation is not required of somebody before they are entitled to use these commonplace modalities.

In the following sections, I wish to suggest such an explanation, which is one that Fine explicitly rejects. It is that ϕ is postulationally possible just in case it is (metaphysically) possible to interpret the quantifiers occurring in ϕ so that ϕ is true under that interpretation. I claim that a sufficiently careful version of this avoids Fine's criticisms of it, as well as other criticisms which may be raised against it.

6.3 Possible interpretation

Fine considers very briefly a definition of postulational possibility in terms of metaphysical possibility, before rejecting it:

> Nor can we take the postulational possibility of a proposition to consist in the metaphysical possibility of our specifying an interpretation under which a proposition is true. For one thing, there may be all sorts of metaphysical constraints on which interpretations it is possible for us to specify. More significantly, it is not metaphysically possible for a quantifier over pure sets, say, to range over more pure sets than there actually are, since pure sets exist of necessity. So this way of thinking will not give us the postulational possibility of there being more pure sets than there actually are. (p.34)

I will argue that these two objection are misguided, or at least would be against the best formulation of such an account. The first objection—that there may be metaphysical constraints on specification of interpretations—will be avoided by modifying the definition to avoid the dependence on *specifications* of interpretations. The second ('more significant') objection—concerning the metaphysical necessity of the existence of sets—will be seen to rest on a use/mention (or perhaps *de re*/*de dicto*) confusion once a more precise formulation is given.

Before getting to a more precise formulation, let me say something briefly about relationship between interpretations, contexts, and the need to specify either of them. An alternative to considering these issues in relation to interpretations—so that a domain expansion is a matter of changing one's interpretation of one's quantifiers—is to consider them in relation to contexts, whereby a domain expansion is a matter of shifting to a different context of quantification.[9] I do not think that there is much in the way of difference between these approaches. For each context of quantification, there will be a unique interpretation of the quantifiers—namely, the interpretation which they receive in that context. And for each interpretation, presumably there will be a context in which the quantifiers are interpreted according to that interpretation.

Now, in order to change context, or to reinterpret parts of one's language, there is no need to *specify* what the interpretation is to be. Indeed, many speakers of a language (which will of course be an *interpreted* language) would not be able to specify how their language is to be interpreted (whatever that would amount to), except perhaps homophonically (or in some cases by means of a translation into another language, say by specifying how a sentence in English is to be interpreted by giving the French equivalent). So, the issue here is not *specifying* interpretations or contexts, but simply being in such and such a context, or interpreting an expression in such and such a way.

[9]The contextual approach is taken by Glanzberg (2001, 2004, 2006); Parsons (1974).

Removing the reference to specifications then, a first attempt at a definition of postulational possibility might be:

(PP1) $\ulcorner\Diamond\phi\urcorner \overset{\text{def}}{\sim} \ulcorner$it is possible to interpret the quantifiers so that $\phi\urcorner$[10]

There is obviously an issue at this point concerning what to make of the construction 'it is possible to *A* so that *B*', where *A* stands for some action, and *B* a possible outcome of such an action. This does not clearly have a straightforward gloss in terms of the possibility of some proposition. But in this it is no different to many other modal notions, such as counterfactual conditionals, and is in any case a commonly used construction. A gloss might perhaps involve considering only possible worlds in which facts concerning the action *A* change. For the moment, I shall simply assume that this kind of construction is understood. Now, in the case of (PP1), it is unclear whether Fine's first objection would apply. What would be required for the objection to work, presumably, is for there to be some interpretation such that it is metaphysically *impossible* to interpret some sentence according to that interpretation. It is not just required that the interpretation be *unspecifiable*; it must be impossible for a being to use a sentence interpreted in that way. But what could such an interpretation possibly be? And in what sense could an interpretation which can never be used be properly called an interpretation? Perhaps one could envisage the kind of thought experiment in which, for example, a shy god immediately strikes down anybody who considers interpreting the quantifiers so that their domain includes the god. Although it may not be possible to rule out such an example *a priori*, I feel comfortable ruling it out on empirical grounds (and the mere remote metaphysical possibility of such a scenario will not suffice to bring back the objection).

(PP1) does, however, suffer insurmountably from another problem, which stems from the fact that the definition involves *using* ϕ, and, in particular, using ϕ in the same context in which one wishes to assert $\ulcorner\Diamond\phi\urcorner$. Consider an example concerning, not the interpretation of quantifiers, but of ordinary predicates.[11] Consider the following question: were we to apply 'leg' to dogs' tails, how many legs would a dog have? The answer, is, of course, four. Nothing that we do concerning how we interpret words can affect facts about the anatomy of dogs. What we *can* say concerning this example, however, is that were we to apply 'leg' to tails, then 'dogs have five legs' would be true. The difference here is that we are evaluating the truth of a sentence under a counterfactual situation in which some of the words in that sentence are interpreted differently.

This then also applies to interpretations of quantifiers. There are no more sets than there actually are (trivially). Moreover, we can not change this simply by interpreting some words differently (just as we can not change the fact that dogs have four legs by changing the meanings of our words).[12] So then, under the definition given by (PP1),

[10] The precise relation '$\overset{\text{def}}{\sim}$' between the expressions named on the left and right hand sides here should be taken as schematic. There are various options for how a definition may be understood. For example, the expressions may be taken to be part of an object language, then the definition may be taken as stating (in the metalanguage) that they are to have the same truth conditions. Or the right hand side might be taken to be in the metalanguage, stating the truth conditions for the sentence named on the left hand side, and so on.

[11] This is an example used by Hirsch (2002). The original question is often attributed to Abraham Lincoln (though presumably not to make a point about use and mention!).

[12] Some kinds of entities might plausibly be taken to arise from our change in linguistic practice. For example, one might hold that a sufficiently major reinterpretation of the language in some literary work might result in a *new* literary work. But this surely does not apply to sets.

the postulational possibility that there be more sets than there actually are is ruled out immediately. This is perhaps the second objection that Fine had in mind. But note that the metaphysical *necessity* that there be no more sets than there actually are is not needed. It is certainly metaphysically possible for there to be more tables than there actually are, but more tables can not be conjured into existence simply by us interpreting our quantifiers differently. Hence the definition rules out the postulational possibility of there being more tables that there actually are by definition.[13] The necessary existence of sets does, however, rule out the option that we may indeed create sets just by reinterpreting our language. Note that this objection does not in any way stem from disallowing the actuality operator involved from exempting evaluation from the scope of the possibility. It simply concerns the inability of speakers to change the world by changing the meanings of their words. For example, under the use of actuality which I have in mind, it would be perfectly correct to say that it is possible to build furniture so that there are more tables than there actually are.

As already suggested, the problem stems from *using* ϕ in the right hand side of the definition.[14] Reinterpreting quantifiers will not change facts about the world. But it *will* change facts about the truth value of quantified sentences. This suggests then the following improvement on (PP1), which mentions, rather than uses ϕ:

$(\text{PP2}_{\text{INT}})$ $\quad \ulcorner \Diamond \phi \urcorner \quad \overset{\text{def}}{\sim}$ ⌜it is possible to interpret the quantifiers so that 'ϕ' is true under that interpretation⌝

Or, in terms of contexts, rather than interpretations:

$(\text{PP2}_{\text{CONT}})$ $\quad \ulcorner \Diamond \phi \urcorner \quad \overset{\text{def}}{\sim}$ ⌜it is possible to shift context so that 'ϕ' is true as uttered in the new context⌝

Similarly, use can be made of a satisfaction predicate, in place of a truth predicate, to deal with formulas with free variables.

This formulation, I believe, avoids the objection stated above to (PP1), and, in particular, avoids Fine's second objection. Consider some context C_1, in which we are (trivially) correct in saying 'everything falls under the range of the quantifiers as used in C_1' (note the *mention* of the quantifiers). It is not self-refuting to claim that it is possible to shift context in such a way so that it is correct to utter 'there is something which did not fall under the quantifiers as used in C_1'. This simply then amounts to the postulational possibility of there being more things than there actually are. Nor is it self refuting if we consider only the kinds of things which exist of necessity. What *would* be self-refuting would be to claim that it is possible to shift to a context C_2 in such a way that it is correct to utter 'there is something which does not fall under the range of the quantifiers as used in C_2'. That amounts to saying that it is possible to shift to a context in which it is

[13] Not that Fine, or anybody else, wants to defend the postulational possibility of there being more tables than there actually are. Nonetheless, the position should not be ruled out as immediately, and in this way. An alternative example might to consider, not tables, but impure sets which have tables in their transitive closure. It is metaphysically possible for there to be more of these than there actually are (if there were more tables), and one might also wish to defend the postulational possibility of there being more of these than there actually are. But the postulational possibility is ruled out in the same way as before.

[14] Hellman (2006) suggests that the key to formulating positions concerning absolute generality is to be careful in when to use and when to mention quantification. His suggestion is similar to what I am proposing, though not using modality (at least, not explicitly).

correct to utter 'there is something which is not anything'.[15]

Although (PP2) avoids the objections which (PP1) faces, it nonetheless creates a number of new difficulties which must be avoided. Firstly, it may be objected that, since the definition of $\Diamond\phi$ involves predication of sentences, it is at best misleading to treat the modality as an *operator*, rather than itself a predicate. A second, related, issue, concerns the appearance of a truth predicate in particular. Since we will want to nest the postulational modal operator, it seems that we will need to account for the truth predicate applying to sentences which themselves contain a truth predicate. Without care, there will be a risk that something like the liar paradox may surface. Thirdly, there is a difficulty when ϕ contains an actuality operator, or other similar scoping operators (this may have been suspected when names for contexts were introduced in the example above instead). For, if, for example, ϕ is 'there are more sets than there actually are', plugging this directly into the definition will fail to get the required result. For 'there are more sets than there actually are' is false in any context and under any interpretation. Finally, a worry may remain concerning whether the possibility of reinterpretation simply reduces to the existence of an interpretation (just as, for example, the provability of a mathematical claim might be thought to reduce to the existence of a proof).

I think that all these difficulties can, however, be overcome.

6.3.1 Predicates and operators

The right hand side of the definition (PP2) essentially involves predicating something of the sentence ϕ. However, what is being claimed to be defined is *not* a predicate of sentences, but an operator. The difference between these categories is an important one. A language which allows predication of sentences must have terms in it which refer to sentences, but operators require no such thing. Moreover, predication of sentences is in some sense inherently more risky than use of operators. For example, the naïve truth schema for a truth *predicate*

$$\mathrm{Tr}\ulcorner\phi\urcorner \leftrightarrow \phi$$

is famously inconsistent (as long as the background theory allows self-reference). But the corresponding principle with a truth *operator*

$$\mathrm{T}\phi \leftrightarrow \phi$$

is utterly harmless (and in many ways, rather dull—the truth operator is just the identity truth-function). The same holds for predicate versions of familiar unary and binary operators; naïve predicate versions of negation (falsehood), conditionals and modality are also inconsistent (Deutsch, 2010; Montague, 1963). Because of these important differences we should not smuggle predicates of sentences into operators, and this—so the objection goes—is just what (PP2) does.

[15] This will not be the case if the notion of interpretation/context is such so as to allow for the quantifiers '$\forall$' and '$\exists$' to take on meanings which should not count as universal and existential quantifiers respectively. For example, if '$\exists$' and its natural language counterparts meant 'most', then 'there is something which is not anything' would come out as true as long as there are at least 3 objects; it would mean that, for most x, it is not the case that most things are identical with x. I shall assume that whatever notion of interpretation or context is used, it will rule out deviant interpretations such as these. I will come back to the issue of just what will count as an interpretation later.

It is not, however, immediately clear that predications of sentences should not feature in the definition of an operator. Consider, for example, the usual Tarskian semantics for the ordinary truth-functional operators, which can be taken in some sense to be a definition of them. So, for example, for conditionals we have (given in some metalanguage):

$$\ulcorner \phi \to \psi \urcorner \text{ is true} \quad \Leftrightarrow \quad \psi \text{ is true or } \phi \text{ is not true}$$

Similarly, a definition in the spirit of (PP2) can be treated as such a metalinguistic specification of truth conditions (see footnote 10), where the metalanguage features vocabulary for metaphysical possibility:

$$\ulcorner \Diamond \phi \urcorner \text{ is true} \quad \Leftrightarrow \text{ it is possible to interpret the quantifiers so that } \phi \text{ is true.}$$

This definition suffices also for nested modal operators. Just as the definition for the conditional will yield—with two applications—the condition that $\ulcorner \phi \to (\psi \to \theta) \urcorner$ is true just in case θ is true or either of ϕ or ψ is not true, the definition for '$\Diamond$' will give relatively simple truth conditions for, eg. $\ulcorner \Diamond\Diamond\phi \urcorner$ as:

$$\begin{aligned} \ulcorner \Diamond\Diamond\phi \urcorner \text{ is true} \quad \Leftrightarrow \quad & \text{it is possible to reinterpret the quantifiers} \\ & \text{so that it is possible to reinterpret the quantifiers so that } \phi \text{ is true.} \end{aligned}$$

It might, however, be demanded that there be a definition of '$\Diamond$' in the object language, or that somehow giving a metalinguistic definition is inadequate. For example, if ascending to a metalanguage is itself the result of a reinterpretation of the language (so that, for example, the domain of quantification includes a set representing the domain of the object language quantifiers), there may be worries of circularity. I do not think that this is much of a worry. Postulational possibility involves the reinterpretation of quantifiers. But, the metalinguistic version of (PP2) does not use quantification at all, so any issues concerning the interpretation of the quantifiers in the metalanguage simply do not arise.

I believe then that these concerns can be overcome in such a manner. But in any case, worries of this kind will not be specific to just the case of postulational possibility, as I have defined it, but must also apple to a number of modal operators which are much more commonplace. Consider, for example, propositional attitude reports that report *de dicto* attitudes to propositions involving proper names. To give an example, suppose that, of the following two sentences, we judge (1) as true, but (2) as false.

1.) Lois Lane believes that Superman is strong. (B(Strong(Superman))

2.) Lois Lane believes that Clark Kent is strong. (B(Strong(Kent)))

These are treated in doxastic logics by means of an operator on sentences (in a way similar to that shown above). But any explanation of the truth value of these sentences will have to involve mentioning the names appearing in the sentences—for this is all that differs between them—and thus involve at some level predication of linguistic items. And this will be the case not just for some specific position with regards to propositional attitude reports, but must hold for any position which allows (1) and (2) to have different truth values.[16]

[16] An exception would be *naïve Russellians*, who claim that co-referential terms are intersubstitutable *salva veritate* even in belief contexts. As such, (1) and (2) will be materially equivalent, with the differences between them being pragmatic, rather than semantic.

6.3.2 Scoping operators

It will be desirable to embed scoping operators within postulational modalities. Indeed, a very natural expression of generality relativism involves actuality—namely, that there (postulationaly) could be something which is not actually anything. More generally, I will want to make use of the scoping operator '↓' which exempts what follows from the scope of the innermost enclosing modal operator (rather than exempting what follows from the scope of *all* enclosing modal operators). But this raises a problem for how such scoping operators are supposed to interact with postulational modalities.

Sentences involving actuality can not simply be inserted into the definition to get the desired result. For example, consider the statement of relativism: there could be something which is not actually anything. Formalised, this is: $\Diamond\exists x@\forall y(x \neq y)$. But the definition tells us that this is true just in case it is possible to reinterpret the quantifiers so that '$\exists x@\forall y(x \neq y)$' is true. But this is plainly false. No matter how the quantifiers are interpreted (ruling out 'deviant' interpretations), '$\exists x@\forall y(x \neq y)$' will be false; the actuality operator no longer lies within the scope of a modality for the purposes of evaluating this sentence, and so is redundant—the sentence is effectively the contradiction '$\exists x\forall y(x \neq y)$'.

It should be noted that this problem is again not unique to postulational modality. It applies too to other operations which result in something like a *de dicto* evaluation. So, for example, we may wish to consider scoping operations in relation to belief. It is common to ascribe to another a mistaken belief in something which does not, in fact, exist (for example, the charge levelled against platonists by nominalists, or against theists by atheists). This is best formalised very similarly to the statement of relativism, as $B_a\exists x@\forall y(x \neq y)$.[17] But this should not be taken to be a claim that a believes the contradiction that there is something which is not actually anything.

A typical way to explain scoping operators is within a possible worlds semantics. Either a particular world is specified as the actual world, or which worlds have been 'travelled through' are kept track of. The result of a scoping operator is then to take one back to the actual world, or to some world which has been travelled through. It should be clear that such an approach is not available here. The very reason for adopting a modality is to avoid quantification over interpretations or contexts, which would be the best candidates for possible worlds. (Similarly, it seems like the possible worlds approach would not do for belief operators either.)

This problem can be solved if '@' and scoping devices more generally are thought of as 'quasi-syntactic'. That is, they do not express any meaning themselves in a sentence, but simply indicate how other parts of a sentence ought to be evaluated. Other symbols which would fit into this category would be parentheses. These too do not carry any meaning of their own, but simply indicate in which order to evaluate other parts of a sentence. For example, $\ulcorner\downarrow\phi\urcorner$ should not be thought to have any meaning itself. Or, perhaps, $\ulcorner\downarrow\phi\urcorner$ should mean exactly the same as ϕ in the same way as $\ulcorner(\phi)\urcorner$ means exactly the same as ϕ.

[17] Of course, if each belief is taken seperately, then there is no need to make use of the actuality operator. It is rather what the two beliefs have *in common* (according to an atheist nominalist) that requires the actuality operator.

A formulation of what it is that the beliefs have in common could also be given without recourse to an actuality operator if a second-order formulation is adopted. Then, the commonality can be expressed as '$\exists F(B_a(\exists xFx) \wedge \neg\exists F(Fx))$'. But it seems desirable to be able to make such a claim in first-order logic.

This idea surely needs more in the way of spelling out. I do so by giving a deductive system which, I claim, does justice to the idea that when $\ulcorner\downarrow\phi\urcorner$ appears within the scope of a modal operator, it signifies that ϕ should be evaluated as if it did not lie within the scope.

Consider, for example, the formula $\ulcorner\Box(\phi \wedge \downarrow\psi)\urcorner$. A deductive system governing the scope operator should allow one to 'extract' ψ from the scope of the necessity operator. Indeed, this formula will clearly be equivalent to $\ulcorner\Box\phi \wedge \psi\urcorner$. The deductive system must give us a systematic way of deducing equivalences like these. But, we should not expect to find an *equivalence* in every case (otherwise there would be no need to have the scoping operator in the first place).

What is needed is a way to 'look inside' the scope of an instance of '$\Box$', and then, when confronted with a subformula prefixed by $\downarrow$, treat this as *not* lying within the scope. This can be done by means of a deductive system which makes use of *labels*. These are used throughout a deduction to keep track of when we are looking inside the scope of a modal operator, and when we can exempt a formula from that scope. In order to look inside the scope of a modal operator, we may simply *ignore* it, but add a label indicating that it is currently being ignored. This label will then indicate that either the modal operator must be appropriately reinstated later, or, if $\downarrow$ is encountered, the operator may be ignored permanently, by removing the label.

I give the details of such a deductive system in appendix A.

6.4 Interpretations and the metaphysics of domain expansion

The definition of postulational modality that I have given involves the notion of an interpretation, or at least the act of interpretation of a piece of language.[18] But what is the relevant notion of an interpretation, and in particular the interpretation of a quantifier? And what is supposed to be going on when one reinterprets the quantifiers? The question concerns how quantifiers succeed in 'latching on' to the world. How is this determined, and how is it supposed to change?

Interpretations of quantifiers are often thought of as being very closely related to *domains*, so that the meaning of a given pair of quantifiers just is the associated domain of quantification. The account of interpretation that I will give will consider talk of domains to be purely secondary in role to more fundamental aspects of the meaning of quantifiers, namely the truth conditions of whole sentences in which they appear. Nonetheless, it is very useful to talk of domains in relation to the meanings of quantifiers (though leaving it open how domains are themselves to be understood). The main question of this section then becomes how to understand the shift in domain that must occur when reinterpreting one's quantifiers. Of particular interest will be cases in which the domain *expands*.

First, I would like to place some constraints on how such a shift in domain could be understood, at least for my purposes. I wish to consider only broadly realist accounts in the shift of interpretation. So, for example, a shift in domain should not be accounted for in a fictionalist manner, so that a shift in domain consists in enhancing a fiction, or

[18] I will put aside for the moment the contextual version.

moving to a new fiction. And a shift in domain should not involve the 'creation' of new mind-dependent objects to add to the domain.

Fine suggests two ways in which the effective expansion of a domain could be understood: as a lifting of restrictions, or by means which does not involve the lifting of restrictions (the latter of which he calls *creative* or *expansive* (Fine, 2005, p.103)). The first of these is the standard way of specifying the interpretation of a quantifier, and shifts in such interpretation. A domain D for a quantifier is specified in some metalanguage which itself has a *wider* domain of quantification, as a restriction of the wider domain. Truth conditions for sentences involving the quantifier are given in the usual way—i.e. $\ulcorner \forall x \phi \urcorner$ is true under such an interpretation iff every object in D satisfies ϕ. It should be clear that this standard notion of the interpretation of a quantifier can not suffice for present purposes. For it presupposes what is, in effect, absolutely unrestricted quantification. And postulational possibility would then be simply definable in terms of this absolutely universal domain. $\ulcorner \Diamond \phi \urcorner$ could simply be defined as $\exists D(\phi^D)$, where ϕ^D is the restriction of the quantifiers in ϕ to D.[19]

This means that the shift in interpretation, if it is to be understood in terms of domains, must be explained in terms of domain *expansion*. That is, moving to a larger domain can not be thought of in terms of a lifting of restrictions on a larger domain. Instead, an interpretation with a larger domain must be reached by presupposing only the domain from which it is an expansion.

But how are we to make sense of the idea that a domain may expand without this being the case of lifting restrictions? There are two questions which must be answered—one broadly epistemic, and one concerning the metaphysics of domain expansion.[20] The epistemic question concerns how we can come to understand an expanded domain, and the metaphysical question concerns how such an expansion of an unrestricted domain can come about—what must the world be like in order for it to be possible to expand the domain? These are not independent questions; what is needed in order to come to an understanding of a new domain may include being able to know that the appropriate metaphysical conditions obtain.

Answers to these questions are immediately available when considering either the placing of or the lifting of restrictions. If we understand a particular domain of quantification, we can understand a restriction of it in the usual way (assuming that we have a proper understanding of the restricting predicate). Similarly, there is nothing metaphysically problematic. If we can quantify over some particular domain, then we can surely quantify over a restriction of it. What is required to be able to quantify over ϕs is that *there are* ϕs. This 'there are' can simply be understood as being the larger (perhaps absolutely unrestricted) quantifier that we are restricting.

Similarly, if domain expansion is thought of in terms of a lifting of restrictions, similar answers can be given. For suppose we start with a domain D' which is a restriction of a domain D. Then we can easily come to understand an expanded domain, as long as it too is a restriction of D. The situation in which it is permissible to expand to a domain which includes ϕs just is when there *are* ϕs (in the sense of the quantifier attached to D).

[19] There are several options as to how D itself may be understood. It could be taken to be a plural variable, a second-order variable, or perhaps (although this last suggestion would have some limitations) as a set variable. The restriction on quantifiers in ϕ would then be given in a way appropriate to whichever choice is made.

[20] These roughly correspond to two question which Hale and Wright (2009b, p. 191) raise which may be asked about the use of abstraction principles.

But these answers are not available for domain *expansion*. The target is as follows. Suppose we have a domain of quantification, D, which is not understood as the restriction of a larger domain (we might say *unrestricted*, if we are careful to distinguish this from *unextendable*) . How can we then come to quantify over a domain which includes objects which are not in D (such as, for example, the domain D itself, considered as a set)? How do we gain an understanding of such a quantifier, and what justification in there for this move being legitimate? Fine writes that this expansion is to come about via *postulation* of a certain kind:

> [T]he change in interpretation of the domain of quantification is somehow given by the condition $\forall y(y \in x)$. But rather than thinking of that condition as serving to define a new predicate by which the quantifier is to be restricted, we should think of it as serving to indicate how the range of the quantifier is to be extended. Associated with the condition $\forall y(y \in x)$ will be an instruction or 'procedural postulate', $!x\forall y(y \in x)$, requiring us to introduce an object x whose members are the objects y of the given domain. In itself, the notation $!x\forall y(y \in x)$ is perhaps neutral as to how the required extension is to be achieved. But the intent is that there is no more fundamental understanding of what the new domain should be except as the domain that might be reached from the given domain by adding an object in conformity with the condition. Thus $!x\forall y(y \in x)$ serves as a positive injunction on how the domain is to be extended rather than as a negative constraint on how it is to be restricted. (Fine, 2006, p.37)

This tells us how, on Fine's view, a domain expansion is to be effected—via postulation—and what effect it should have—that there be an object satisfying such and such a condition. Given a logic governing procedural postulates (which Fine (2005) promises), it will also provide a route for learning about the consequence of such an expansion. This then gives a partial answer to the epistemological question. But we still require an answer to the metaphysical question. When is it legitimate to make such a postulation and go on to speak as if the domain has expanded? Moreover, what justification is there for postulation *ever* being legitimate?

For Fine, a postulation is legitimate just in case it is consistent to make the stipulation, and moreover, that it is consistent with a set of 'postulational constraints' governing the vocabulary in the postulates. For example, the constraints governing postulation about sets are:

Extensionality $\forall x\forall y[Sx \wedge Sy \wedge \forall z(z \in x \leftrightarrow z \in y) \rightarrow x = y]$

Sethood $\forall y[\exists x(x \in y) \rightarrow Sy]$

Set-rigidity $\forall y\forall F[Sy \wedge \forall x(x \in y \rightarrow Fx) \rightarrow \Box\forall x(x \in y \rightarrow Fx)]$.

But this does not explain *why* postulation is legitimate in these cases, if ever. For one thing, consistency generally falls well short of truth; it is consistent to suppose that the earth is flat, or that there are talking donkeys, or to postulate electrons into existence, for example. And it is not clear how consistency with a set of constraints improves this matter.

Weir (2007) raises a number of challenges along these lines, and which surely have to be met by the proponent of domain expansion. I shall claim that Fine's responses to these challenges is not adequate. But I will argue that they can indeed be met, by adapting neo-Fregean ideas about singular terms to quantifiers.

Broadly, there are two relevant criticisms that Weir raises.[21] Firstly, there are some apparent counterexamples to the claim that consistency and compatibility with constraints is sufficient for the legitimacy of postulation. These are examples where it appears that it would be permissible to postulate the objects of some physical theory, such as electrons. The example is as follows. Suppose that T is some physical theory, and $\theta(x_1, \ldots, x_n, R_1, \ldots, R_m)$ is the result of replacing each non-logical constant in T by variables of the appropriate kind. The *Carnap conditional* of T is then

$$\exists!(x_1, \ldots, x_n, R_1, \ldots, R_m)\theta(x_1, \ldots, x_n, R_1, \ldots, R_m) \rightarrow T$$

which plausibly serves as an (non-existentially committing) implicit definition of the terms in T. But then, if this is taken to be a postulational constraint, postulating the existence of the objects involved in the theory will be permitted.

Hence, some non-*ad hoc* criteria for what count as acceptable constraints must be given. But, even if some non-*ad hoc* restriction can be made, this will not result in a justification for why it is then legitimate to expand one's domain via postulation. Surely, an opponent will argue, in order for it to be legitimate for us to interpret our quantifiers so that $\ulcorner \exists x \phi \urcorner$ comes out as true, it is not sufficient that it simply be consistent that there are ϕs. What is required instead is *existence*; there must *be* some ϕs. There must be a domain which contains ϕs. But this will not ever be the case when we are considering a genuine expansion. For then any domain which we have access to, and can use to ask the quantificational question 'are there ϕs', will not contain ϕs. In the case of the main opponent—the absolutist—the domain in question will be what they perceive to be the absolutely unrestricted domain. More must be said in order to answer *why* a condition such as consistency could serve as a sufficient condition for domain expansion, and why demands for more (such as *existence* in some already given sense) can be rejected.

There are two broadly anti-realist proposals which Weir suggests could make sense of such undemanding sufficient conditions (although each has problems of its own). This is his second objection—that domain expansion must rely on some form of anti-realism. The first is that the domain expansion involves the literal *creation* of new objects. For example, it is plausible that one could expand the domain of absolutely all tables simply by building more tables.[22] It is also plausible that some abstract objects can be created, such as fictional characters, works of literature and the like. But this is much more implausible in the case of sets. For one thing, it makes the existence of sets a contingent and tensed matter. But in any case, it is not the kind of realist metaphysics which is being sought.

The second anti-realist picture is a fictionalist one. On this picture, we are not expanding a genuine domain of quantification, but rather just changing what we can say within a fiction. So, a domain expansion is akin to adding new characters to a story, and the resulting claim that there is something satisfying some condition is asserted

[21] Weir also criticises Fine's approach for its use of second-order logic. I shall not discuss that criticism here.

[22] Whether this example works will depend on some contentious issues in metaphysics. For example, it will depend on whether an absolutely unrestricted quantifier must include future objects, and whether the rearrangement of matter can result in the creation of new objects.

within an implicit 'according to the fiction' operator.[23] While this may be a plausible approach, it is again not the realist picture we are after. Moreover, it is unclear what advantages a fictionalist reading of domain expansion would have over a more direct form of fictionalism along the lines of Field (1980). In particular, it will presumably inherit any drawbacks that more standard forms of fictionalism face.

Regardless of their individual merits and demerits of these positions, neither satisfy the requirement that domain expansion be made sense of in a broadly realist fashion. Neither do they do justice to the idea that it is the *interpretation* of a quantifier that is changed. What is to occur in a domain expansion is not to result from a change in how the world is (via creation), or a change in what is true according to some fiction. Rather, all that happens is a change in the truth-value of sentences involving quantification, brought about through a change in *meaning* (to some extent) of the quantifier. We do not change how the world is, but merely how it is being described.

The position is thus similar in some respects to that of *quantifier variance* (cf. eg. Hirsch, 2009). According to such a view, the meaning of quantifiers can vary, and in such a way that no one meaning is privileged over others, either metaphysically—by, for example, representing more accurately the *real* structure of the world—or logically—so that all other quantifiers are to be understood as restrictions of one particular quantifier. Both positions also share difficulties. In particular, they both face the challenge of explaining these differences and changes in meaning. And this challenge must be met without negating their position by presupposing some wider perspective (either an absolutely unrestricted domain, or metaphysically fundamental quantifier) in which to do so.

The most natural way to explain such a change would be in terms of domains, and hence by domain expansion. But it is then hard to see how *that* might be explained but by means of talking of change of meaning. But this is obviously circular. A different problem is to explain why these separate meanings should not be taken in the discussed fictionalist manner, but as genuine quantifiers which are used, at face value, as quantifiers. The problem is even more pressing when the opponent claims to be quantifying over absolutely everything. For any expanded quantifier meaning will appear not to have a domain (for otherwise it would be expressible as a restriction).

So, the two problems are this:[24]

(a) How are the changes in meanings to be accounted for? What aspects of the meaning of the quantifier is is that *varies*?

(b) How is it that all the supposed quantifiers are *quantifiers*? What aspect of meaning is *shared*?

I believe that answers to these questions can be given by rejecting talk of domains, at least, as being fundamental to the meaning of a quantifier. And this can be done by adopting something like the syntactic priority thesis (cf. Wright, 1983) to quantifiers as opposed to just singular terms.

[23] If one adopts realism about fictional entities, then this view will collapse to the first, creationist, view.

[24] These two challenges are raised by Hale and Wright (2009b) and others against quantifier variantists.

6.5 The context principle, the syntactic priority thesis, and quantifiers

Frege (1884, p.x) urges us 'never to ask for the meaning of a word in isolation, but only in the context of a sentence.' This principle has usually been used to apply to singular terms, notably by neo-Fregeans, but if the principle is correct, it is surely correct for any category of words, quantifiers included. If this is right then this will allow for a way to explain a shift in the interpretation of a quantifier not in terms of its domain (and hence, in some sense, removed from the context of a sentence), but simply in terms of the truth conditions of sentences involving quantification.

More needs to be said about how this is to be done, and how it will avoid the worries laid out in the previous section. That will be the aim of this section.

If the context principle is to be of any use in explaining the meanings of words in terms of the sentences in which they occur, we can not explain what category some term belongs to (such as a singular term, or a quantifier) in terms of its semantic function. So, for example, we can not state that σ is a singular term if it refers (or purports to refer) to an object, or that it is a quantifier if it serves to generalise over a domain. Instead, an account must be given in non-semantic terms. The claim that the syntactic behaviour of some term is prior in some sense to its semantic function is the *syntactic priority thesis*. Stating this principle (together with a form of the context principle), Wright (1983) writes:

> The question of whether a particular expression is a candidate to refer… is entirely a matter of the sort of syntactic role which it plays in whole sentences… questions concerning its reference should be addressed by… reflection on the truth-conditions of sentences of the appropriate kind. (p.51)

The corresponding principle for quantifiers will then be:

> The question of whether a particular expression is a candidate to *quantify* is entirely a matter of the sort of syntactic role which it plays in whole sentences. Questions concerning its *domain* should be addressed by reflection on the truth-conditions of sentences of the appropriate kind.

In order to apply such a principle, we must then give some syntactic or inferential account of what it is for some term to be a quantifier of a certain kind.[25] I propose that we simply say that some symbol Σ is an existential quantifier just in case it obeys the free-logical inference rules for existential quantifiers:[26]

$$(\Sigma\text{-I})\ \frac{\phi(x/t)}{\Sigma x\phi} \qquad (\Sigma\text{-E})\ \frac{\Sigma x\phi \qquad \begin{array}{c}[\phi(x/t)]\\ \vdots \\ \psi\end{array}}{\psi}$$

[25] Such a task for singular terms is carried out by Dummett (1973, ch. 4) and Hale (2001b,c)

[26] This will not suffice for natural language quantifier phrases. In this case, grammatical considerations are not sufficient to distinguish quantifiers from singular terms (both are noun phrases), and the natural language counterparts of the quantifier rules are not sufficient either. Nonetheless, it seems plausible that further tests could be given to distinguish the two, similar to those given by Dummett and Hale.

and similarly for universal quantifiers.

It is important that if we wish to allow for different quantifiers having different domains, and if we wish to allow for empty singular terms, the rules must be free. This is because otherwise, for any term t, and any quantifier, t must have a referent that lies within the domain of that quantifier.[27]

Now, if we adopt something like the context principle and the syntactic priority thesis for quantifiers, it seems that we can shrug off demands to *explain* shifts in interpretation by means of domains, and to answer the two questions (a) and (b) raised earlier about these shifts in meaning (p. 109). Recall that these were: what aspect of meaning is is that is preserved amongst quantifiers, and what aspect of meaning is it that varies? The aspect of meaning that is *preserved* between quantifier interpretations is their inferential behaviour—this is what sets them apart as *quantifiers* (of a certain kind). More importantly, we can explain how it is that the meanings can vary, without taking a perspective which views them as restrictions, and without making appeal to dubious ideas such as creation. What changes about the meaning of a quantifier is simply the truth conditions of sentences containing them. Any further demand to explain *this*, perhaps in terms of domains, is to get the order of explanation the wrong way round. It is not the change of truth-conditions that is to be explained in terms of change in domain. Rather, the change in domain is to be explained in terms of change of truth conditions.

Where does this then leave us with respect to the metaphysical and epistemic questions that need answering about domain expansion raised in the last section? Recall that the challenges are: (1) to explain how one can come to an understanding of a new interpretation of a quantifier without presupposing an understanding of larger domain of which it is a restriction; (2) to answer what the world has to be like for such a domain expansion to be successful in generalising over a domain, and justifying that answer (without presupposing something like absolutism).

To the first question, an answer similar to the neo-Fregean position concerning singular terms may be given. For them, abstraction principles serve as a stipulated implicit definition of a class of singular terms (or rather, of the abstraction operator which is common to that class of singular terms), and an understanding of those singular terms derives from knowing the truth of the implicit definition. For quantifiers then, perhaps something like implicit definitions can be used to stipulate the truth conditions of sentences involving a quantifier. Perhaps Fine's procedural postulates could serve as a method of stipulating such truth conditions. I would favour, however, a more uniform way of doing so, via abstraction principles. Consider the following abstraction principle:

$$\text{(AP)} \qquad \forall F \forall G[\S F = \S G \leftrightarrow \Phi(F, G)]$$

We wish to be defining a *new* quantifier which, in effect, has §s in its domain[28], so we do not assume that $\S F$ has a referent which lies within the domain of any quantifier appearing in Φ. Nor do we need to assume that it does not – it is not a matter of stipulation that the new quantifier will have a new domain (although that may, in some circumstances, be a consequence).

[27] See, for example, Turner (2010), who puts free logic to a similar use.

[28] The domain talk here is easily eliminated. What is required is that $\exists x \exists F(x = \S F)$ is true with the new quantifier.

Now, we can define a new pair of quantifiers $\exists^{\S}/\forall^{\S}$ by stipulating the truth conditions of sentences containing it as follows:

$$(*) \qquad \exists^{\S} x\phi \leftrightarrow \exists F\phi(\S F)$$

where in some cases the appearance of § on the right hand side may be eliminated by use of (AP). For example, consider the claim that there are[§] at least two things:

$$(6.4) \qquad \exists^{\S} x \exists^{\S} y(x \neq y)$$

By $(*)$ this is true just in case the following is true:

$$(6.5) \qquad \exists F \exists G(\S F \neq \S G)$$

The abstraction operator can then be eliminated via (AP), to get:

$$(6.6) \qquad \exists F \exists G \neg \Phi(F, G)$$

Since (6.6) contains only old vocabulary, an understanding of (6.4) can be gained just given an understanding of the old vocabulary. Since it may be the case that the new quantifier $\exists^{\S}/\forall^{\S}$ may exceed any of the quantifiers appearing in Φ, this may then give us a means of understanding a quantifier without presupposing an understanding of a wider quantifier.

What then of the second question: In what circumstances is such a method of implicit definition justified, and what reason is there to believe that it is *ever* justified? The answer is again similar to that given by neo-Fregeans concerning singular terms. There will be some limits on what implicit definitions are acceptable, but these will not be a matter of, as Hale and Wright (2009b) put it, '*hitting off* reference to a range of entities qualified to play the role that the principle defines' . Rather, the requirements on implicit definitions will be logical/linguistic as opposed to metaphysical—requirements such as conservativeness, harmony and so on. In the case of definitions of quantifiers in the way that I have suggested, an option is available which is not available for the standard approach to abstraction. That option is a strong form of conservativeness—roughly speaking, the requirement that the definition should not result in any new consequences expressed in the old language. This is not available on the standard approach. For example, HP entails that there are at least two objects ($\exists x \exists y(x \neq y)$), which is expressible without using the newly introduced abstraction operator, and which does not follow without HP. Although the present expansive approach will also (if it is to be successful) entail that there are at least two objects, this will be a claim expressed using the *new* quantifiers, and hence does not violate strong conservativeness.

And this position—that the conservativeness of some implicit definition is sufficient for the possibility of interpreting the quantifier in such and such a way[29]—is itself justified given the context principle and the syntactic priority thesis. For what else could be required? The implicit definition simply serves as a prescription of how to use a new piece of language, and if this new piece of language does not interfere with the old language (through violations of conservativeness), what is to stop us? And it can not be claimed that the new piece of language may fail to be a genuine quantifier on the

[29]Or, at least that whatever sufficient conditions there are do not involve the metaphysical cooperation of the world in some way or other.

grounds of there not being any domain for it to generalise over. For, by the syntactic priority thesis, all there is to being a genuine quantifier is to have the correct inferential behaviour.

6.6 Conclusion

My aim in this chapter had been to defend generality relativism. I should emphasise that the aim has not been to argue directly in favour of the view. Instead, it has been to flesh out what I think the view must amount to, and, in doing so, to respond to what I see as some of the most pressing objections to the view.

My first aim was to respond to the charge that relativism is either self-refuting or inexpressible. There were two responses which I gave to that—one negative, and one postitive. The first, negative, response was to argue that the very same problem faces everyone in the debate, including the absolutist. The second, more positive response, was to provide a way to articulate relativism, by building on work of Fine's, and claiming that relativism can be expressed using the notion of *postulational possibility*. In contrast to Fine, I argued that it is not acceptable simply to take the modality as primitive. I supplied such an explanation, in terms of the possibility of reinterpreting vocabulary.

My second aim was to supply a metaphysics of domain expansion which does not suggest any mysterious faculty of creation of objects. Such a metaphysics (or, perhaps, anti-metaphysics) consisted in taking the meaning of a quantifier to derive, not primarily from a presupposed domain of quantification, but rather from the truth conditions of sentences in which it appears.

Chapter 7

Abstraction with domain expansion

7.1 Introduction

In this chapter, my aim is to answer the second of the challenges raised in chapter 5. That is, it is to develop a formal theory of expansionist abstraction. This will be done by investigating which components of the standard approach to abstraction involve the assumption of a fixed domain, and then modifying these components. In section 7.2 I shall discuss what changes are required in order to remove the assumption of absolute generality. This will involve using the modality of postulational possibility and adapting abstraction principles for a modal setting. Sections 7.3–7.7 will deal with the consequences of this background logic, and in particular for a set theory based on a modal version of Basic Law V. In particular, I will show that when an appropriate background logic is adopted, Basic Law V is consistent and can account for all of standard set theory.

7.2 A suitable logic for domain expansion

We saw in chapter 2 that the assumption of a fixed domain (and, in particular, an absolutely unrestricted domain) plays a key role in motivating the characteristic E!-I rule of a *negative* free logic. Recall, this is the rule:

$$(E!\text{-I})\ \frac{\phi(x/t) \quad \phi \text{ is atomic}}{\exists x(x = t)}$$

It is worth going over how this is so. A particularly important class of instances of the E!-I rule is when the singular terms involved are abstract terms. In such a circumstance, the definition of such a term may involve a quantifier. In cases such as these, E!-I says that, if an abstract term t occurs in a true atomic sentence, then one may infer that a referent of t lies in the domain of the very quantifier which is involved in defining t. And, as I argued, this seems plausible when the domain of quantification is absolutely unrestricted.

But no such motivation is available when we allow that the domain may expand. For in that case, why should we expect that the referent of t should lie within any particular domain?[1] Consider, for example, a commonplace case of where there is room for a domain to expand—when the quantifiers in question are contextually restricted to salient objects. Suppose that a family is discussing the state of their packing for a holiday. In such a context, it is likely that the quantifiers may be restricted to, say, just those things which might be expected to be packed, and thus that 'everything is packed' might be true in this context. But, there will be singular terms which refer, and which may appear in true atomic sentences, yet do not fall under the range of the quantifier. So, for example, it could presumably be true, in such a circumstance, that the kitchen sink is white. It does not follow, of course, that the kitchen sink is packed! That is, the following inference:

P1.	Everything is packed.
P2.	The kitchen sink is white.
C.	The kitchen sink is packed.

is not valid. In the context under consideration, the two premises are true, yet the conclusion false. It would, however, be validated by a negative free logic.

Of course, in the course of such a discussion, an utterance of P2 is very likely to cause the context to *shift*, so that the quantifier now ranges over the domain which includes the kitchen sink. But that is beside the point. Unless changes are made to the logic to take into account the possibility of a context shift from one line of a proof to the next, it must be a presupposition that we keep the context fixed.

What we *can* say in such a situation is that the context *can* shift so as to include a referent of the singular term in question. From the truth of an atomic sentence featuring a singular term t, one can infer that it is *possible* to shift context, or interpretation of the quantifier, so that t has a referent lying within the new domain of quantification. This can be expressed in the form of an inference rule making use of postulational modality as follows:

$$\frac{\phi(t/x) \quad \phi \text{ is atomic}}{\Diamond \exists x(x = t)}$$

This rule will still not quite suffice, however. Many such terms under consideration will be *non-rigid*, and their reference may vary as the interpretation of the quantifiers varies. For example, 'the set of everything' will denote a different set depending on the interpretation of 'everything'. Suppose that t is such a term. We do not want to say that the truth—in the present context—of an atomic sentence containing t allows us to reinterpret the quantifiers so that it includes the referent of t *in the new context*. Rather, we want to say that this allows us to reinterpret the quantifiers so that it includes the referent of t in the *old* context (the one in which an atomic sentence containing t is true). In other words, we want to exempt t from the scope of the possibility operator. This can be done by allowing the scoping device $\downarrow$ to apply to terms as well as formulas. The revised ($\Diamond E$!-I) rule will then be:

[1] We would expect this if our language was carefully restricted in such a way so that the referent of any term will lie within a particular domain of quantification. Consider, for example, the usual language of arithmetic, in which every term refers to an object in the (restricted) domain of natural numbers.

$$(\Diamond E!\text{-I})\ \frac{\phi(t/x) \quad \phi \text{ is atomic}}{\Diamond\exists x(x = \downarrow t)}$$

So then, in order to assess the impact of dropping the insistence on absolutely unrestricted quantification, the ($E!$-I) rule must be replaced by ($\Diamond E!$-I). This then gives us the basis for developing a formal theory of abstraction principles where the domain may expand. But, in order to make such a replacement, there are two main tasks that must be carried out. Firstly, changes need to be made to the form of abstraction principles in order to deal with modality. Secondly, a suitable background second-order modal logic which incorporates ($\Diamond E!$-I) must be developed.

7.2.1 Modalised abstraction principles

Abstraction principles in the form:

$$\text{(AP)} \qquad \forall \nu_1 \forall \nu_2 [\S(\nu_1) = \S(\nu_2) \leftrightarrow E(\nu_1, \nu_2)]$$

fail to give a fully general criterion of identity when considering modal contexts. The reason is that, even when stated as a necessity:

$$\text{(AP}\Box\text{)} \qquad \Box\forall \nu_1 \forall \nu_2 \big[\S(\nu_1) = \S(\nu_2) \leftrightarrow E(\nu_1, \nu_2)\big],$$

they fail to give *trans-world* identity conditions. Or, to avoid talk of worlds, given ν_1 and ν_2 in the abstraction domain, AP$\Box$ fails to determine whether ν_1 would have the same associated abstract were circumstances one way than ν_2 would have were circumstances some different way. This failure is pointed out by Hale and Wright (2001b, p.358) in the case of DE:

> To presuppose that the Direction Equivalence may be appealed to in reasoning in an arbitrary hypothetical scenario of lines and their properties is merely to assume that the principle is a necessary truth. *That* much is presumably alright… But that only yields that the Direction Equivalence holds *of* any world. What [is needed] is that the scope of the principles encompasses not merely relations of parallelism within a world, but relations of, as it were, *trans-world parallelism*. [2]

The problem is that in AP$\Box$, everything is evaluated within the scope of the necessity operator, and so evaluated at the same world. What is needed then, is for some part of the abstraction relation E to be evaluated outside the scope of the necessity, by making use of scoping devices, such as an actuality operator. To develop such a method in full generality to cover any abstraction relation would be a large undertaking, and I shall not attempt to do so here. Instead, I shall simply concentrate on the specific cases of BLV and HP.

[2] Worries could of course be raised about whether some absolute notion of transworld parallelism can be made sense of, independent of some specified *fixed* frame of reference (consider, for example, thought experiments where the entire universe is rotated 45 degrees). However, in the present case, where the modality only concerns changes in interpretation and context, the worry seems less pressing. In any case, certain relations—in particular coextensiveness and equinumerosity—certainly do seem to make sense considered as transworld relations.

We shall need to allow the actuality operator '@' to apply to singular terms as well as to formulas. This is because, as already noted, abstract terms are likely to be non-rigid, in that they have different referents depending on the circumstances in which they are evaluated.[3] By allowing this, we can formulate a modal version of BLV as follows:

(BLV@) $\quad \Box\forall F\forall G(\varepsilon F = @\varepsilon G \leftrightarrow \forall x(Fx \leftrightarrow @Gx))$

This says that for any concepts F and G, the set of Fs in some circumstances (or, in the case of postulational possibility, under some possible interpretation of the quantifiers) is the set of Gs under the actual circumstances iff everything that is an F under those circumstances in actually a G and vice versa.

But there are still two things wrong with this. The intended effect of an abstract term such as 'εF' or 'NF' is to refer to the set or number (or whatever) of *all and only* the Fs, and so implicitly involving the domain of quantification of the context in which it is used. This is what is stipulated in the non-modal versions of abstraction principles. For example, the criterion for F and G having the same set is $\forall x(Fx \leftrightarrow Gx)$, i.e. that *all* the Fs are G and *only* the Fs are G. It also corresponds to the actual usage of such terms; suppose, for example, that the context is such that the quantifier phrase 'all bottles' is restricted to just those bottles which are in this room. It seems plausible that, in this same context, the singular term 'the set of bottles' will refer to the set of just those bottles in this room.

But BLV@ does not do this. The criterion of identity should be that every F is an actual G and vice-versa, where 'actual G' is to be understood as only encompassing things which fall under the *actual* quantifier. But the relation $\forall x(Fx \leftrightarrow @Gx)$ does not do this, since the quantifier throughout is evaluated within the scope of the necessity operator. Instead, the relation should be $\forall x(Fx \leftrightarrow @(E!x \wedge Gx))$, where $E!x$ is an abbreviation of $\exists y(y = x)$.[4]

The other problem with BLV@ is that it is tied to actual circumstances. We want instead to state its necessitation. This can be done by using the scoping operator $\downarrow$ instead of @ (since otherwise @ would exempt what follows it from all enclosing modal operators). We then have the following:

(BLV$\downarrow$) $\quad \Box\Box\forall F\forall G(\varepsilon F = \downarrow\varepsilon G \leftrightarrow \forall x(Fx \leftrightarrow \downarrow\exists y(y = x \wedge Gy))$

Similarly, a suitable version of HP can be given:

(HP$\downarrow$) $\quad \Box\Box\forall F\forall G(NF = \downarrow NG \leftrightarrow$
$\exists R[\forall x(Fx \rightarrow \exists!y(\downarrow(E!y \wedge Gy) \wedge Rxy)) \wedge \forall x(\downarrow(E!x \wedge Gx) \rightarrow \exists!y(Fy \wedge Ryx))])$

[3] If we are interested in abstraction principles in a more general modal setting (e.g., in relation to metaphysical possibility), then there will be even more examples of this. For example, 'the set of red things' will denote different sets depending on the circumstances.

[4] This would of course not be a problem if the logic was a negative free logic. Since then, $@Gx$ would entail $@E!x$. But rejecting the ($E!$-I) rule characteristic of negative free logic is central to the current approach.

In fact, if a stronger comprehension principle were adopted than the one that I do adopt, then BLV@ would be inconsistent. Suppose that there are concept terms I and E such that for any term t, $\Box(Et \leftrightarrow \exists x(x = t))$ and $\Box It$ (so E is the concept 'falls under the quantifier', and I is essentially the concept 'is self-identical'). Then BLV@ will give us $\Box(\varepsilon I = @\varepsilon I)$, $\varepsilon I = \varepsilon E$ and $\Diamond(\varepsilon I \neq @\varepsilon E)$. But then these together yield a contradiction.

7.2.2 Logic

In order to investigate abstraction with the possibility of domain expansion, we need a suitable background logic in which we can lay down the modalised abstraction principles. Such a logic must feature the ($\Diamond E!$-I) rule, as discussed at the start of this section. But there are still more details to be given. Firstly, since both the ($\Diamond E!$-I) rule and the modalised abstraction principles feature the scoping operator $\downarrow$, more needs to be said about how this works. Secondly, appropriate principles need to be given for the behaviour of the modal operator. I shall call the resulting logic EL (for *expansion logic*). Full details of the logic—proof theory, model theory and a soundness result (though no completeness result)—can be found in Appendix A. This section will discuss the most distinctive features, and the motivation for them.

Proof theory for $\downarrow$

The backtracking operator $\downarrow$ was first introduced by Hodes (1984b), for which he provides a possible worlds semantics. However, for the present purposes, a proof theory is required (or, at the very least, highly desirable). Why is this? There are a few reasons. One was mentioned in the previous chapter, in 6.3.2; a proof theory might help to explain the sense in which the operator may be seen as *quasi-syntactic*—not having any meaning of its own, but instead signifying how other parts of a sentence are to be evaluated (similarly to parentheses). But there are two more reasons simply relying on the model theory would be problematic.

Firstly, the model theory essentially replaces the use of modal operators by quantification over worlds. But the worlds are—in the setting of postulational modality—perhaps best identified with contexts or interpretations of the quantifiers. But it was precisely because of the problems with expressing generality relativism by quantifying over contexts or interpretations that the modal approach was adopted. Even worse, the model theory for quantified modal logic involves quantification over one 'super'-domain, which contains within it every object which could be quantified over. But the denial of the possibility of such a quantification is central to the present approach. So, any reliance on the possible worlds semantics should be eliminated (except in an instrumental role to prove, via soundness, that certain proof-theoretic consequences do not obtain).

Secondly, one of the main reasons to develop a theory of expansionist abstraction in an internal manner, rather than the external manner in which it was characterised in chapter 5, was to give a story about how Hero's knowledge of mathematics may develop from abstraction principles. Since the object language now contains the $\downarrow$ operator, it must be claimed that it is the kind of thing that Hero can understand. But, since Hero is assumed to have no prior knowledge of mathematics, he can not understand the operator by means of the semantics, since the semantics is set theoretic.

I provide the full details of a natural deduction style system for the $\downarrow$ operator in appendix A.1. A key feature of the system, which requires defending, is that formulas in a proof are *labelled* by a finite sequence of natural numbers; the various inference rules permit changing these labels in certain ways. So, for example, a simple deduction of ϕ from $\Box\downarrow\phi$ could proceed as follows:

1) $\Box\downarrow\phi\,;-$ (Assumption)
2) $\downarrow\phi\,;0$ (1, $\Box$-E)

3) $\phi\,;-$ (2, ↓-E)

Here, the label is everything following ';' in a formula. '-' denotes an empty label. The (□-E) rule permits adding to the label of a formula (as occurs in the step from 1 to 2), and the (↓-E) rule permits removing from a label.

Now, the presence of such labels in the proof theory may raise a worry. At first, these labels look very much like they are playing the role of referring to worlds in the semantics (or—perhaps worse—to sequences of worlds). But then the possible worlds semantics again appears to be playing a crucial role, which, as I have already noted, it must not. How are the inference rules then going to be justified?

Now, even if the labels in proofs do have to be interpreted as making reference to world sequences, this is not disastrous. For, as noted, it would be plausible to take the worlds as being contexts or interpretations. And reference to such things might seem relatively innocuous. What *is* problematic is *quantification* over contexts, which merely using labels does not do. Nonetheless, the use of labels need not even imply so much as this.

Labelled sentences should not be taken to be some special kind of sentence which can be asserted (and so perhaps referring to worlds in the process). They never appear as the conclusion of a proper deduction, not even sentences labelled by the empty label. Proof theoretic consequence, as defined (definition A.3, p. 166), is a relation simply between sets of sentences and a sentence, all of which are sentences of a language which does not contain any vocabulary referring to or quantifying over worlds. Labels will only appear during the course of a proof. The use of labels should only be thought of as something of a bookkeeping device, to keep track of how some sentence may become exempt from the scope of the modal operators which enclose it. Thus, labels need not have any semantic interpretation any more than, say, line numbers, or the horizontal and vertical lines common in a number of versions of natural deduction.[5]

The inference rules which may be used during a proof do, however, feature labels in their premises and conclusion. How then can they be justified, if their premises and conclusions can not be taken to be assertions of any kind? It can not be in terms of validity, since labelled sentences are not the kind of things which can be said to be true or false.

The inference rules in question which need motivation are those for ↓ and □. These are (where $\vec{s}$ is a finite sequence of natural numbers, and $\vec{s}-$ is the result of removing the last element of $\vec{s}$):

$$(\Box\text{-I})\ \frac{\phi\,;\vec{s}}{\Box\phi\,;\vec{s}-} \qquad (\Box\text{-E})\ \frac{\Box\phi\,;\vec{s}}{\phi\,;\vec{s},n}$$

$$(\downarrow\text{-I})\ \frac{\phi\,;\vec{s}}{\downarrow\phi\,;\vec{s},n} \qquad (\downarrow\text{-E})\ \frac{\downarrow\phi\,;\vec{s}}{\phi\,;\vec{s}-}$$

with a restriction on the (□-I) rule that the premise may only depend on assumptions $\psi\,;\vec{t}$ such that $\vec{s}$ properly extends $\vec{t}$.

[5] In fact, the labels play a very similar role to the vertical lines indicating a *strict subproof* in the Fitch style systems of Fitch (1952); Hazen (1990); Siemens (1977).

Consider first the rules associated with ↓. The intended meaning for ↓ is that it exempts what follows it from the scope of the innermost modal operator from which it is not already exempt (if it already contains some occurrence of ↓). And so, it can be appended to any formula, together with a note (the label) signifying that there should be no overall effect when a modal operator is applied. The label is then simply signifying that a modal operator is currently being ignored for some purposes. So, in this way (↓-I) and (□-I) act in a kind of harmony with each other, with a use of each in turn resulting in what is essentially the same assertion—ϕ and $\Box\downarrow\phi$. Similarly, this process can be iterated, appending $\downarrow\downarrow$ to a formula, together with a note signifying that there should be no overall affect when *two* modal operators are applied, allowing $\Box\Box\downarrow\downarrow\phi$ to be derived from ϕ. (↓-E) and (□-E) then simply operate as duals of (↓-I) and (□-I) respectively.

This may make it seem peculiar that labels need to be *sequences* of natural numbers, rather than just a single natural number. After all, if they are just to note how many modal operators are permitted to be added, then surely a single number will do, namely the number of modal operators that may be added? But there may be need, considering the restriction on the (□-I) rule, to introduce two assumptions with different labels which are of the same length.

For example, consider the derived (◊-E) rule:

$$(\Diamond\text{-E})\ \frac{\Diamond\phi\,;\vec{s} \qquad \begin{array}{c}[\phi\,;\vec{s},n]\\ \vdots\\ \psi\,;\vec{t}\end{array}}{\psi\,;\vec{t}}$$

which is subject to the restriction that $\vec{s},n$ does not appear in any other assumption on which $(\psi\,;\vec{t})$ depends (this restriction follows from the restriction on the (□-I) rule). Now, it must be possible to introduce separate labels of the form $\vec{s},n$ if using this rule more than once at a time. For example, to prove $\Diamond\phi\vee\Diamond\psi\to\Diamond(\phi\vee\psi)$, two such subproofs are needed.[6]

Modal principles

It also needs to be considered what principles the modality will satisfy, and how it will interact with quantifiers. The modal logic governing ◊ and □ will be that of S4.2, which

[6]The proof is:

$$\dfrac{\Diamond\phi\vee\Diamond\psi \qquad \dfrac{\dfrac{\dfrac{}{\Diamond\phi\,;-}\,(1) \qquad \dfrac{\dfrac{\dfrac{}{\phi\,;0}\,(2)}{\phi\vee\psi\,;0}}{\Diamond(\phi\vee\psi)\,;-}}{\Diamond(\phi\vee\psi)}\,(2) \qquad \dfrac{\dfrac{\dfrac{}{\Diamond\psi\,;-}\,(1) \qquad \dfrac{\dfrac{\dfrac{}{\psi\,;1}\,(3)}{\phi\vee\psi\,;1}}{\Diamond(\phi\vee\psi)\,;-}}{\Diamond(\phi\vee\psi)}\,(3)}{\Diamond(\phi\vee\psi)}\,(1)}{\Diamond(\phi\vee\psi)}$$

(which makes use of derived rules for ∨ as well).

means adding the following axioms:[7]

(T) $\phi \to \Diamond\phi$

(4) $\Diamond\Diamond\phi \to \Diamond\phi$

(G) $\Diamond\Box\phi \to \Box\Diamond\phi$

In the context of postulational modality, these are very plausible. (T) says that if ϕ is true under the current interpretation, then it is possible to interpret the quantifiers so that ϕ is true. This is clearly correct, since the current interpretation serves as a witness to the possibility. (4) says, in effect, that if it is possible to make ϕ true by expanding the domain *twice*, then it is possible to expand the domain so that ϕ is true. Again, this is correct—it is still possible to interpret the quantifiers so that ϕ is true, even if it requires two shifts. (G) is the hardest to motivate. If $\Box$ and $\Diamond$ were to be interpreted as quantification over something like domains, then it could be done easily. Suppose that there is a domain D (or interpretation I, or context C etc.) so that for all D' expanding D, ϕ is true interpreted on D'. Now consider an arbitrary domain D''. Then we can expand the domain to $D'' \cup D'$, which is an expansion of D', and hence ϕ holds on it. This does not by itself justify (G), since the modality is *not* to be understood as quantification. However, it is plausible that, since the modality is in some sense a generalisation of quantification over domains, (G) should hold for it.

But we should not expect further rules or axioms. In particular, we should not expect postulational possibility to satisfy the characteristic S5 axiom:

(5) $\Diamond\phi \to \Box\Diamond\phi$

This states that, if it is possible to expand the domain so that ϕ is true, then, no matter how we expand the domain, it will always be possible to expand so that ϕ is true. But this will not be the case. Suppose that the domain is restricted to one which contains only two objects, so that $\exists x \exists y(x \neq y \wedge \forall z(z = x \vee z = y))$ is true. Then (by the (T) axiom), this is postulationally possible. But, it will not be the case that, however we expand the domain, this sentence will be postulationally possible. For if we were to expand the domain to one in which there are three objects, it will never be possible to expand it further so that there are just two objects again.

First- and second-order quantification

As mentioned previously, a key feature of the logic is the ($\Diamond E!$-I) rule and the restricted free-logical quantifier rules for first-order quantification. In addition, we will adopt the Converse Barcan Formula:

(CBF) $\exists x \Diamond\phi \to \Diamond\exists x\phi$

which guarantees that the domain can only expand, and not contract. This can be seen in some ways simply as a restriction on which reinterpretations are going to be considered possible.

[7] Actually, when $\downarrow$ is added to the language, these need restricting. Roughly speaking, they are restricted to formulas where any instance of $\downarrow$ is 'cancelled out' by a modal operator enclosing it. See appendix A.1.5 for full details.

For the second-order quantifiers, the rules will be the usual non-free rules, as well as both the Barcan and Converse Barcan Formulas. We will also have a strong modal comprehension principle:

(Comp) $$\exists F \Box \forall x (Fx \leftrightarrow \phi).$$

These reflect a natural thought that any formula will define an intensionally equivalent concept, which does not depend on the existence of objects for its existence (see, e.g. Williamson (forthcoming) for arguments that we should expect second-order modal logic to feature these principles).

Model theory

As was the case for static abstraction, it will be useful to have a model theory for pragmatic purposes. In order to adapt it to the present purposes, the model theory for static abstraction must be adapted so as to allow for modal operators. This can be done by having a possible worlds semantics, as is usual for modal logics.

A model will be a 4-tuple $\mathcal{M} = \langle W, D, \delta, I \rangle$. W is a set of worlds, which informally can be said to represent various possible contexts or interpretations of the quantifiers. D is a set of all 'possible objects'. δ is a function $\delta : W \to \mathcal{P}(D)$; for each $w \in W$, $\delta(w)$ is the domain of the first-order quantifier at that world/context/interpretation. Finally, I is an interpretation function, which will provide an interpretation for the abstraction operator.

The domain of the second-order variables, D_2, is fixed across worlds as the set of all *intensions*, which are functions $f : W \to \mathcal{P}(D)$. This represents the idea that the extension of a property may vary from world to world.

Given this interpretation of the second-order quantifiers, the interpretation of an abstraction principle will be a function $I(\S) : W \times D_2 \to D$. This allows a given property to have a different abstract associated with it at different worlds.

Given these ingredients, satisfaction of a formula in a model can proceed as normal. The only slight complication is to allow for the scoping operator. I give full details of how this works in appendix A. For most purposes in this chapter, it should be clear what effect it will have.

Remark on 'nonexistent' objects

An important feature of the logic is that it supplies the ability to refer to objects which do not fall under the domain of quantification. This may seem slightly odd, but it is in fact a crucial feature of generality relativism. A common feature of various extendibility arguments is that, for a given domain of quantification, they exhibit an object which does not lie in that domain. And this must be done by referring to one such object. One way to do this will be by using abstract terms, but the features of the logic will also allow for this in other ways—in particular, by using variables in certain ways.

One model theoretic way to consider reference to objects which do not lie in the domain of quantification—which is shared by any modal logic with variable domains—is if some formula with a free variable is evaluated at a world w with respect to a variable assignment a such that $a(x) \notin \delta(w)$ (where $\delta(w)$ is the domain of the world w). But, as already mentioned, the semantics should be treated as purely instrumental. In addition,

it might be claimed that the only assertable parts of the language are sentences, with no free variables. However, the presence of the scoping operator allows one to partially mirror this kind variable assignment in the object language without recourse to the model theory. For example, consider (GR), which, recall, is the sentence $\Diamond\exists x{\downarrow}\neg E!x$. Here, for the purpose of evaluating the $\neg E!x$ part, x is essentially assigned an object which does not lie in the domain of quantification. This will also occur during the course of proofs, when there are free variables around.

It should also be noted that a consequence of this (and this is not specific to this particular system) is that a formula $\phi(x)$ with a free variable will not be equivalent to its universal generalisation. In particular, I will state some principles and results in free variable form, and others with variables quantified out. It should be noted that a free variable formulation is stronger than the corresponding quantified formulation.[8]

It is tempting to paraphrase this ability to refer to objects which do not lie in the present domain of quantification as reference to objects which 'do not exist'. This may seem even more perplexing (and even suggest a certain kind of Meinongianism). But as long as 'a exists' is understood as 'there is something which is a', and it is understood that there is no absolutely unrestricted quantification, then the claim that it is possible to refer to objects which do not exist should no longer be so mysterious. It is simply the claim that a just does not fall under the present domain of quantification. Care should probably be taken in using phrases like 'a exists' for this reason. Nonetheless I will use it throughout in this latter sense, tied directly to quantification.

7.3 The consequences of Expansion Logic

Now, having laid out the logic, it is possible to see the effects of adding modalised abstraction principles. I will mainly concern myself with BLV, since that is the abstraction principle most likely to cause worries about inconsistency. It is also the abstraction principle which seems most likely to provide for a foundation of set theory. So, then, what does the theory which results from adding BLV$\downarrow$ to EL look like?

(Certain results will consist in saying that some sentence ϕ is derivable from BLV$\downarrow$ in EL. For these, I have typically given a formal proof of ϕ from BLV$\downarrow$ in appendix B, and given a rough outline of the formal proof here.)

7.3.1 The consistency of BLV$\downarrow$

Since Basic Law V is infamously inconsistent with a background of a standard second-order logic, it will be important to establish that BLV$\downarrow$ is consistent with the background logic being EL. (Note that BLV$\downarrow$ is not a weakening of the non-modal version, since the non-modal version is a straightforward consequence of it.)

Recall, the modal version of BLV is:

(BLV$\downarrow$) $\qquad \Box\Box\forall F\forall G\big[\varepsilon F = {\downarrow}\varepsilon G \leftrightarrow \forall x\big(Fx \leftrightarrow {\downarrow}(E!x \wedge Gx)\big)\big]$

[8] An open formula might not be taken to be the kind of thing which is assertable, and so not comparable to any closed sentence. But a principle stated as an open formula (or more likely, a scheme of open formulas) can be taken, not as an assertion, but as a principle licensing the use of that open formula in a proof. If a sentence corresponding to an open formula is required, then, for $\phi(x)$ a formula, the sentence $\Box\Box\forall x{\downarrow}\phi$ will have the same effect.

The model theory developed and the soundness theorem can be used to show that EL+BLV↓ is consistent, by constructing a model. It can be seen that a model $\mathcal{M} = \langle W, \delta, R, I\rangle$ satisfies BLV↓ if and only if, for every $w_1, w_2 \in W$, with $w_1 R w_2$, and for every $f, g \in D_2$,

$$I(\varepsilon)(w_1, f) = I(\varepsilon)(w_2, g) \text{ iff, for every } d \in \delta(w_2), d \in f(w_1) \cap \delta(w_1) \Leftrightarrow d \in g(w_2).$$

By extensionality of sets (in the metatheory), the right hand side is just $f(w_1) \cap \delta(w_1) = g(w_2) \cap \delta(w_2)$.

So, the condition is:

$$I(\varepsilon)(w_1, f) = I(\varepsilon)(w_2, g) \text{ iff } f(w_1) \cap \delta(w_1) = g(w_2) \cap \delta(w_2).$$

We now construct such a model. Let $W = \mathbb{N}$, $R = \leq$ (the usual ordering on $\mathbb{N}$) and $\delta(n) = V_n$, the nth finite level of the cumulative hierarchy (so $V_0 = \varnothing$ and for every n, $V_{n+1} = \mathcal{P}(V_n)$). So, $D = \bigcup V_n = V_\omega$. Finally, let $I(\varepsilon)(n, f) = f(n) \cap V_n \in V_{n+1}$.

It is easy to check that this model satisfies all the requirements needed. Hence BLV↓ is consistent.

7.3.2 Some rigidity properties

An important kind of property of formulas of EL will be various kinds of *rigidity*. Roughly speaking, rigid formulas will be those which have a fixed extension, in some sense. In particular, it is often claimed that set membership should have certain rigidity properties (e.g. Fine, 1981; Linnebo, ms; Parsons, 1983b), and rigidity of formulas will become important later on when recovering parts of set theory.

There are two broad classes of rigidity properties which might be considered. In EL, it is perfectly possible for a formula $\phi(x)$ to be true at a world even if the free variable x is assigned an object not in the domain of that world. This will almost certainly be the case in any modal logic with varying domains (consider for example the formula $\Diamond\exists y(y = x)$), but the presence of ↓ allows this to be expressed to some extent in the object language. For example, that $\phi(x)$ is satisfied at a world by an object which is not in the domain of that world can be expressed by $\Diamond\exists x{\downarrow}(\neg E!x \wedge \phi(x))$.

Then, the two kinds of rigidity are as follows. The first requires that a formula ϕ have the same extension of *existing* objects falling under it at each world. Model theoretically, ϕ is rigid in this sense at a world w if for all w' such that wRw', $\{a \in \delta(w') : w' \vDash \phi(a)\} = \{a \in \delta(w) : w \vDash \phi(a)\}$. This can be expressed in the object language by:

$(\mathrm{R}_0^x\text{-}\phi)$ $\qquad\qquad \Box\forall x(\phi \leftrightarrow {\downarrow}(E!x \wedge \phi))$

For example, the formula '$x = x$' is not rigid in this sense (its extension grows), but '$\downarrow\exists y(y = x)$' is.[9]

[9] This is the kind of rigidity which is normally used, partly because it can be expressed without using the ↓ operator, whereas the second can not (or so I suspect). It can be expressed by the following three properties (see eg. Parsons (1983b, pp.299–302)):

(R-ϕ) $\qquad \phi(x) \rightarrow \Box\phi(x)$

(R-$\neg\phi$) $\qquad \neg\phi(x) \rightarrow \Box\neg\phi(x)$

(BF-ϕ) $\qquad \forall x(\phi \rightarrow \Box Fx) \rightarrow \Box\forall x(\phi \rightarrow Fx)$

These can be shown to be equivalent to R_0-ϕ.

But another kind of rigidity requires that ϕ have the same extension at every world, where this is understood as including not just objects in the domain of that world, but objects in the domain of other worlds as well. Model theoretically, ϕ is rigid in this sense at a world w if for all w' such that wRw', $\{a \in D : w \vDash \phi(a)\} = \{a \in D : w' \vDash \phi(a)\}$. This can be expressed proof theoretically as:

$(\mathrm{R}_1^x\text{-}\phi)$ $\qquad\qquad \Box\forall x(\phi \leftrightarrow \downarrow\phi)$

For example, '$x = x$' *is* rigid in this sense (although it would not be in a negative free logic), but $\exists y(x = y)$ is not.

7.3.3 Defining set-theoretic notions

Set membership is often taken be rigid. We should thus like to decide which of the rigidity properties is most appropriate, and then show that we can prove that this rigidity follows from BLV↓. That is, that one of the following hold:

$(\mathrm{R}_0\text{-}\in)$ $\qquad\qquad \Box\forall x(x \in y \leftrightarrow \downarrow(E!x \wedge x \in y))$

$(\mathrm{R}_1\text{-}\in)$ $\qquad\qquad \Box\forall x(x \in y \leftrightarrow \downarrow(x \in y))$

However, the language of abstraction which we are currently dealing with does not feature a primitive membership relation. So, before dealing with the question of whether these principles follow, a membership relation must be defined (and preferably in a natural way).

A natural definition is the following (which is similar as that given in other discussions of abstractionist set theory such as Boolos (1989)):

$$s \in_0 t \stackrel{\text{def}}{=} \exists F\exists x(x = s \wedge t = \varepsilon F \wedge Fx).$$

(The subscript will become clear in a moment.) However, due to the presence of non-denoting terms in the language (or, rather, terms which denote an object which does not lie in the present domain of quantification), this will fail to give the expected results in some cases. The definition requires that, for '$a \in b$' to hold in some context, a must be in the domain of quantification, and b must be a subset of the domain. That the first is required is built directly into the definition,[10] and that the second is required follows from the quantification implicit in the 'ε' operator (as discussed in section 7.2.1). This is undesirable; membership should be like identity in not requiring that the objects involved fall under any particular domain of quantification. In particular, as with identity, we should expect membership to be rigid in the second sense discussed. But with this definition, $\Box\forall x(x \in y \rightarrow \downarrow(x \in y))$ may not hold.

So, it should not be required for $x \in y$ that x lies in the present domain of quantification, or that y is a set of objects taken from the present domain of quantification (just as it is not required for the truth of $x = y$ that x and y fall under the current domain of quantification). Instead, it should merely be required that it is possible to interpret the quantifiers so that x falls under them, and that y is a set formed from objects which fall

[10] This is in contrast to how membership is usually defined from a set abstraction operator. But the requirement is needed in order that the only objects which are members of 'εF' are those which *actually* exist, since εF is supposed to be the set of only such objects.

under the range of some quantifiers under which x falls. That is, the following revised definition should be adopted:

$$s \in_1 t \stackrel{\text{def}}{=} \Diamond \exists F \exists x (x = \downarrow s \wedge \varepsilon F = \downarrow t \wedge Fx)$$

The use of '$\downarrow$' is to ensure that if 's' and 't' are non-rigid, the same objects are picked out by 's' and 't' within the scope of the modal operator as outside.

This definition behaves just as we would expect it to when $y = \varepsilon F$ for some F:

Proposition 7.1. BLV$\downarrow \vdash y = \varepsilon F \rightarrow (x \in y \leftrightarrow (E!x \wedge Fx))$ *(for either* $\in_0$ *or* $\in_1$*.)*

Proof. The right to left direction is trivial: From $(E!x \wedge Fx \wedge y = \varepsilon F)$ we get $x \in y$ by an application of the (T) axiom (for $\in_1$) and existential generalisation.

For the left to right direction: $x \in y$ implies there is possibly some G such that $y = \varepsilon G$ and so on. Then BLV$\downarrow$ can be used to compare G and F, which gives the required result.

See page 175 for a formal proof. □

This has a perhaps more natural (though weaker) result as an immediate corollary:

Corollary 7.2. BLV$\downarrow \vdash \forall x \forall F (x \in \varepsilon F \leftrightarrow Fx)$

We can now prove the appropriate form of rigidity for $\in_1$:

Proposition 7.3. BLV$\downarrow \vdash \mathrm{R}_1\text{-} \in_1$

Proof. The left to right direction of $\mathrm{R}_1\text{-} \in$ follows fairly directly from a use of the (4) axiom. The right to left direction can be proved by defining a concept in such a way that it witnesses the truth of $x \in y$ appropriately. This direction makes use of the (G) axiom.

See page 175 for a formal proof. □

Since the resulting set theory will be one which allows non-sets, it will be necessary to define a sethood predicate. Two alternative definitions can be given, analogous to $\in_0$ and $\in_1$:

$$\mathrm{Set}_0(t) \stackrel{\text{def}}{=} \exists F (\varepsilon F = t)$$
$$\mathrm{Set}_1(t) \stackrel{\text{def}}{=} \Diamond \exists F (\varepsilon F = \downarrow t)$$

I will take the second of these as the definition that I am principally interested in. Again, we can prove the appropriate form of rigidity:

Proposition 7.4. BLV$\downarrow \vdash \mathrm{R}_1^x\text{-}(\mathrm{Set}_1(x))$

Proof. The proof is essentially the same as (or rather, a part of) the proof of proposition 7.3, since $x \in y$ entails $\mathrm{Set}(y)$. □

The relationships between the two alternative definitions of sethood and the two definitions of membership will be useful to examine. Partly, this will be because such results will be useful later on. But also, there may be a worry about the appearance of modal vocabulary in the definitions. It certainly does not seem that mathematicians have anything modal in mind when using such notions; the non-modal definitions appear much more plausible as an account of mathematics. It will therefore be important

to see under what circumstances the alternative definitions coincide. I claim that these are circumstances which always obtain in the course of ordinary mathematics.

It is clear that $\mathrm{Set}_0(x) \to \mathrm{Set}_1(x)$ and that $x \in_0 y \to x \in_1 y$. But we want to know when the converse holds.

Firstly, it can be shown that if $x \in_1 y$, $x \in_0 y$ will hold just in case x refers to an object in the present domain, and when $\mathrm{Set}_0(y)$:

Proposition 7.5. $\mathrm{BLV}{\downarrow} \vdash x \in_1 y \to (x \in_0 y \leftrightarrow (E!x \wedge \mathrm{Set}_0(y))$

Proof. See page 176 □

Then the question arises—given y such that $\mathrm{Set}_1(y)$, under what conditions will $\mathrm{Set}_0(y)$ hold? The answer is: just in case any expansion of the domain will not include members of y:

Proposition 7.6. $\mathrm{BLV}{\downarrow} \vdash \mathrm{Set}_1(x) \to [\mathrm{Set}_0(x) \leftrightarrow \Box\forall y(y \in x \to {\downarrow}E!y)]$

Proof. The left to right direction makes use of BLV↓ by comparing concepts F and G which both have x as a set in different contexts. Then, the relation $\forall y(Fy \leftrightarrow {\downarrow}(Gy \wedge E!y)$ allows one to derive ${\downarrow}E!y$ at the appropriate point. For the right to left direction, we define a set by using the concept F such that $\Box\forall y(Fy \leftrightarrow y \in x)$. Together with $\Box\forall y(y \in x \to {\downarrow}E!y)$, it can be proved that $x = \varepsilon F$.

See page 176 for a full formal proof. □

7.3.4 Set comprehension

As well as principles governing how set-theoretic vocabulary behaves modally, it is important to know what sets there are, or what sets there *can* be. It is a fairly direct consequence of BLV↓ and ($\Diamond E!$-I) that we have:

$$\forall F \Diamond \exists y(y = {\downarrow}\varepsilon F)$$

That is, every concept defines a set.

But we might also ask what *formulas* define a set. That is, what formulas are such that there is (or could be) a set x such that $\forall y(y \in x \leftrightarrow \phi)$? In one sense, just as every concept defines a set, every formula defines a set. That is, for every formula ϕ, the set of all and only the ϕs can be introduced into the domain:

$$\Diamond \exists x \forall y(y \in x \leftrightarrow {\downarrow}(E!x \wedge \phi(x)))$$

The modified right hand side here means that the set concerned only contains objects which fall under the *actual* domain of quantification.

But we can do better than this. It possible to give sufficient conditions for a formula not just to define a set, but to continue to define the same set as the domain expands. This gives a comprehension principle which is first order and does not feature ↓ in the condition which defines the set:

Proposition 7.7. $\mathrm{BLV}{\downarrow} \vdash \Diamond \mathrm{R}_0^x\text{-}\phi \wedge \Box \mathrm{R}_1^x\text{-}\phi \to \Diamond\exists y{\downarrow}(\mathrm{Set}\, y \wedge \Box\forall x{\downarrow}(x \in y \leftrightarrow \phi))$.

Proof. The basic idea is that, when ϕ is rigid, we can let $y = \varepsilon F$, where F is the concept defined by ϕ. Then we can prove that $\Box(y = \varepsilon F)$ (which makes use of rigidity). This means that y necessarily has the ϕs falling under it, which is what is required. The possible existence of y is then given by the ($\Diamond E!$-I) rule.

See p.178 □

7.4 Modal abstraction and the iterative conception

Other modal approaches to set theory (e.g. Linnebo, ms; Parsons, 1983b; Studd, 2012) have been motivated to various degrees by the iterative conception of set. On this view, the sets are built up in stages, starting from the empty set, and at every stage, as many sets as can be introduced are introduced. That is, at any stage, the very next stage consists of all subsets of the preceding stage. This process then carries on into the transfinite by taking unions at limit stages.

Since the present approach—which is not explicitly motivated by the iterative conception—is similar to these other modal approaches, it will be useful to compare it to them.

The present approach is similar in many ways to the iterative conception. If sets are introduced into the domain according to the principles discussed in the last section, then they will be introduced in stages. And, for any collection of objects, it is possible to introduce a set of them. However, the present approach differs in a couple of respects, in not having consequences which might be thought to be essential features of the iterative conception. These differences affect how various set-theoretic axioms are eventually proved.

One aspect of the iterative conception is that the elements of a set are *prior* to the set itself. That is, before a set can be introduced into the domain of quantification, all of its elements must be present in the domain of quantification. A similar, weaker condition is that elements of a set can not be introduced into the domain of quantification *after* that set has been introduced into the domain of quantification. This can be expressed as:

$$E!y \wedge \mathrm{Set}(y) \rightarrow \Box\forall x(x \in y \rightarrow \downarrow E!x)$$

Such a principle might be called *weak* priority, and versions of it can be derived in those modal set theories mentioned above.

This is not, however, a consequence of BLV↓ and the background logic. This is not because BLV↓ allows one to introduce sets into the domain before introducing elements of that set into the domain. Indeed, as the comprehension principles of the previous section show, any sets introduced in such a way do satisfy the weak priority principle. It is simply that it has not been ruled out that sets could be introduced in some other way as well.

Serving as an alternative to this priority principle, however, is a simple corollary of proposition 7.6:

Corollary 7.8 (of propn. 7.6)**.**

$$E!y \wedge \mathrm{Set}(y) \rightarrow \Diamond\Box\forall x(x \in y \rightarrow \downarrow E!x)$$

Proof. Because $\mathrm{Set}(y)$, we have $\Diamond\exists F(y = \varepsilon F)$. Since $\Box(\exists F(y = \varepsilon F) \rightarrow \Box\forall x(x \in y \rightarrow \downarrow E!x)$ by proposition 7.6, we have $\Diamond\Box\forall x(x \in y \rightarrow \downarrow E!x)$ □

This principle states that, given any set, it will always be possible to expand the domain to one which includes all of its elements.

Another feature of the iterative conception of set is a kind of maximality condition that, at each stage, *all* of the subsets of the previous stage are formed. A similar principle

can be stated in the present system. It is:

$$\Box(\exists x{\downarrow}\forall y(y \neq x) \rightarrow \forall F\exists y(y = {\downarrow}\varepsilon F))$$

This says that, whenever the domain is expanded, then it is expanded by introducing sets of every concept (but the sets that those concepts had at the previous stage). Again, this is not a consequence under the present approach. Although for every concept, it is possible to introduce a set for that concept, it does not follow that sets for all of the concepts can be introduced at once. A principle considered later, in section 7.6, will allow one to derive a similar condition, which will suffice for the purposes which maximality is needed.

7.5 Interpreting set theory

Since the aim is to develop as much set theory as possible with this approach, it needs to be seen which results of standard set theory (and in particular, which *axioms*) can be proved by adopting BLV↓. Clearly this is not simply the case of going straight ahead and attempting to prove axioms of ZFC,[11] since the language of standard set theory and the language of abstraction are very different. Instead, the aim is to *interpret* standard set theory in the system resulting from EL and BLV↓, making use of definitions, translations and so on. One difference between the languages is the difference in primitive non-logical vocabulary—'$\in$' (and perhaps a sethood predicate) for standard set theory, and the abstraction operator on the present approach. This is easily dealt with by making use of a natural definition of $\in$, as above. The other difference is that the present approach features modal vocabulary, whilst the standard language does not.[12] Given this, there are several routes one could take in trying to interpret non-modal set theory. These are: (a) we could ignore modality, and simply try to prove the axioms of ZFC as they are; (b) we could aim to prove that, taken together (in some sense), all the consequences of ZFC (or some subset of them) are (postulationally) *possible*; (c) some wholesale translation of non-modal set theory into modal set theory could be undertaken, and then we could aim to prove the translations of the axioms of ZFC.

The first of these approaches, (a), is clearly hopeless. It amounts to trying to prove that the axioms of ZFC are true on any world of any possible worlds model. Or rather (to avoid over-reliance on the talk of possible worlds), that any way of interpreting the quantifiers will have them ranging over some suitably large universe of sets, before any reinterpretation or domain expansion. This is clearly undesirable, and the fact that the model mentioned previously (p. 125) has an empty domain for one world shows that it is unattainable.

What is wanted instead is not that the quantifiers (under whatever context) range over a suitably rich domain of sets, but that we could shift context and reinterpret the quantifiers so that they range over such a domain. This is the approach (b). The thought is that, in doing set theory, set theorists operate under some context in which they quantify over some domain of sets.[13] In that case, the aim of abstraction—as presently

[11] Actually, I will be interested in set theories with urelements, since BLV↓ allows for urelements.

[12] If 'standard' set theory is first-order ZFC, then another difference will be the presence of second-order quantification. This need not be an issue, since the target could just be taken to be second-order ZFC instead.

[13] Alternatively, it might be thought that assertions made in the language of set theory should be taken to

conceived—is to show how, on the basis of some characterisation of the concept of set via BLV↓, it is possible to interpret the quantifiers so that they range over a suitably rich domain of sets. That is, the aim is to show how the statements made by set theorists are postulationally possible.

How is this aim to be fulfilled? First, it needs to be shown that the axioms of ZFC (or some of them) are possible. In particular, they should be jointly possible. Fortunately, second-order ZFC is finitely axiomatisable, and can therefore be axiomatised by a single sentence, namely the conjunction of these finitely many axioms, which I shall write Z. Hence, the aim is to show that BLV↓ $\vdash_{\mathsf{EL}} \Diamond Z$. Consequences of the axioms can then be accounted for as follows. If $ZFC \vdash \phi$, for some sentence ϕ, then evidently $\vdash_{\mathsf{EL}} \Box(Z \to \phi)$. Hence BLV↓ $\vdash_{\mathsf{EL}} \Diamond\phi$, and even BLV↓ $\vdash_{\mathsf{EL}} \Diamond(Z \wedge \phi)$.

Unfortunately, this shares the feature of approach (a) that, short of strengthening the background logic, very little can be achieved. For example, the empty set axiom can be proved (i.e. $\Diamond\exists x[\mathrm{Set}(x) \wedge \forall y(y \notin x)]$), but little else. It can be proved that it is possible to expand the domain in useful ways—for example, for any x and y, it is possible to expand the domain to contain their pair. But this is not the pairing axiom, since it can not be proved that the domain can be expanded so that for any x and y in the new domain, their pair also lies in that domain. The same goes for, for example, singletons.

That (b) will not result in much set theory can also be seen from the existence of the model in section 7.3.1. For, in that model, every domain is V_i for some $i \in \mathbb{N}$, which is finite. Since these domains represent the domains of quantification which are possible, it follows that we can not prove that it is possible to expand the domain to an infinite one. This is essentially the problem attested to in 5.3, that there is nothing to say that it is possible to iterate the process of domain expansion into the transfinite.

What about (c)? This approach would involve the wholesale translation of sentences of non-modal set theory into the modal abstractionist theory, by recursively giving some translation rules. This will generally allow more power than the approach in (b), since there is no restriction on what the translation may be. The aim is to have some translation $\phi \mapsto \phi^*$ such that:

- BLV↓ $\vdash_{\mathsf{EL}} \phi^*$ for every axiom[14] ϕ of ZFC,
- If $\Gamma \vdash \phi$ then $\Gamma^* \vdash \phi^*$, where $\Gamma^* = \{\phi^* : \phi \in \Gamma\}$.

This approach is taken by Parsons (1983b) and Linnebo (ms). It is also implicit in various presentations of Kripke semantics for intuitionistic set theory (eg. Lear (1977)). Parsons' translation is to take a formula ϕ, translate into an intuitionistic language via a negative translation[15], and then into the modal language via a standard translation of intuitionistic logic in classical modal logic (by, e.g. boxing every subformula). Linnebo's translation is to replace each occurrence of $\forall$ with $\Box\forall$ (and likewise each occurrence of $\exists$ with $\Diamond\exists$).

Considering translations such as this provides for precisely measuring the strength of the resulting system, in the usual way in which the strength of systems are measured.

be systematically ambiguous, intended to be about any suitable domain of sets. This will also be covered by approach (b).

[14]It might instead be acceptable to only succeed in interpreting most of the axioms of ZFC, although this would be less desirable.

[15]For example, simply double-negating every subformula is such a translation.

However—without the assistance of further argument—such approaches risk failing to be faithful to the meaning of the non-modal language. Both the languages involved are interpreted, at least to some extent (although there is room for disagreement over what the interpretation is), so each assertion in each language presumably expresses some proposition. So, if a translation $\phi \mapsto \phi^*$ is to be of any interest beyond measuring consistency strength, it must map sentences onto those which mean the same thing. It will not be the case that any translation is allowed. For an example (for a different translation), it is well known that real analysis (and in fact almost all ordinary mathematics) can be interpreted in second-order arithmetic. But it would not be open for neo-Fregeans to claim that Hume's Principle can be used as a foundation of almost all mathematics. The reason is that, in interpretations of analysis in arithmetic, real numbers are taken to be certain *second-order* entities. But under the intended interpretation of second-order logic as used by the neo-Fregeans, these are *concepts* whereas real numbers are *objects*. Similarly, one could perhaps imagine a translation from the non-modal set theory to the abstractionist language which completely changed the logical structure. Such a translation would not say anything about how much set theory can be proved from BLV↓.

That is not to say, of course, that *any* wholesale translation will not be faithful in this way. Of particular interest will be whether translations similar to those of Parsons and Linnebo are acceptable in this way, since these appear to allow the most set theory to be interpreted.[16] It seems to me that these translations will not do.

As a first assumption, it seems fair to take the surface grammar of set theory completely at face value. In particular, the quantifiers are indeed quantifiers, which range over some domain or other. But the translations do not map quantified statements onto quantified statements with the same structure. For example, using Linnebo's translation would map $\exists y \forall x (x \notin y)$ onto $\Diamond \exists y \Box \forall x (x \notin y)$. It could be claimed that the compounds '$\Box\forall$' and '$\Diamond\exists$' *are* quantifiers, so that a quantified statement has been mapped to a quantified statement of the same form. But if this is the case, then it is possible to interpret the simple quantifier symbols themselves as having such a meaning. But this them just goes back to approach (b), since the claim is still that it is possible to interpret the quantifiers such that $\exists y \forall x (x \notin y)$ is true. In any case, it seems that such an approach could be problematic from the current perspective. If $\Box\forall$ and $\Diamond\exists$ are indeed quantifiers, then it looks very much like they would be absolutely universal quantifiers, but we have seen that absolutely universal quantification is in conflict with basic law V.

Alternatively, it could be claimed that the original assumption about the surface grammar of mathematics should be challenged, and in particular the claim that in set theory there is some particular domain over which the quantifiers range. Indeed, it has often been claimed that assertions of set theory should be taken to be systematically ambiguous. That is, they should be taken as true no matter what domain of sets the quantifiers range over. But this interpretation of set theory does not warrant arbitrary translations, but rather just approach (b). The claim that some sentence is to be taken as true no matter how the quantifiers are interpreted (which is the claim that a systematically ambiguous assertion is intended to convey) is formalised in modal language as $\Box\phi$. And when the assertion is only intended to be ambiguous over suitably large, set theoretic

[16] I should note that my only concern (for the moment) is whether these translations are acceptable for *my* purposes. Nothing I say will concern whether these translations are appropriate for the purposes that Parsons and Linnebo put their translations to.

interpretations, this is best formalised as something perhaps like $\Box(Z \to \phi)$, just as in approach (b). Then the additional component of (b), proving $\Diamond Z$, simply ensures that this is not vacuous.

Perhaps more radical claims could be made about the intended meaning on non-modal set theory. However, I do not wish to make such claims. As such, I am limited to approach (b) to explain how much set theory is recoverable from this modal abstractionist approach. Since—as I have already noted—approach (b) does not allow for much set theory at all, more principles concerning postulational modality must be considered.

7.6 Reflection

It looks then that we do not get much set theory from this approach, at least, if we only attempt an interpretation of the same kind as type (b). But, it might be possible to justify and add some additional principles which do allow us to do so. The idea will be to suggest some principle which has the same effect as permitting transfinite iterations of domain expansion.

I will suggest that we adopt such a principle which, for reasons which will become clear, will be called a *reflection* principle. My aim in this section will be to specify such a principle, to give a brief motivation for it, and to investigate the result of adding such a principle. I will postpone a more detailed defence of the principle until chapter 8, since the issues that arise are somewhat involved.

Before stating this principle, some additional notation will be useful. Where ϕ is a sentence not involving the modal operators, let $\phi^{\Diamond}$ result from ϕ by replacing each instance of $\exists x$ with $\Diamond\exists x\downarrow$, and each instance of $\forall x$ with $\Box\forall x\downarrow$. Similarly, where ϕ is again a sentence not involving modal operators, let $\phi^{\downarrow}$ result from ϕ be prefixing every abstract term with $\downarrow$.

Now, consider the following inference rule:

$$\text{(Refl)}\ \frac{\phi^{\Diamond}}{\Diamond\phi^{\downarrow}}$$

From an external point of view, we can view this rule as follows: Given some progression of domains $D_1, D_2, \ldots$, the sentence $\phi^{\Diamond}$ says that ϕ holds over the entire domain $D = \bigcup_i D_i$. It is, in some sense, a 'potentialised' sentence; it talks not just about actual objects, but possible objects as well. The conclusion of the rule then just says that we can expand our domain to one in which includes all the possible objects which are involved in the truth of the premise. It collapses the potential into the actual. (The conclusion has to feature $\phi^{\downarrow}$ rather than just ϕ to ensure that the abstract terms featured in it still refer to the same objects as they do in the premise.)

Obviously this way of talking about things will not do as a motivation. For one thing, it does not make sense to talk about all postulationally possible objects if we have adopted relativism; to do so amounts just to talking about absolutely all objects. Even if this weren't a problem (so that talk of all possible objects were eliminated), motivation is still lacking. What is there to say that such a reinterpretation of the quantifiers is indeed possible?

The problems of motivating such a rule and distinguishing it from a motivation for absolutism are large. As such I will not say more about that motivation here, postponing such a discussion for chapter 8. I will instead confine myself to discussing the consequences of such a rule.

Informally, such a rule may allow us to break the transfinite barrier, so to speak. Suppose that we have an abstraction principle that allows us to expand our domain finitely, but arbitrarily many times, so that we have a sequence $D_1, D_2, \ldots$ (we might for example have the D_i as in section 7.3.1, so that $D_i = V_i$). The reflection principle will allow us to, in effect, move to an ωth stage, $D_\omega = \bigcup_i D_i$. From here, we can then start iterating again, to get $D_{\omega+1}, D_{\omega+2} \ldots$ and so on.

This kind of reasoning will also often be internalisable to the object language. So, for example, suppose that we can deduce that, necessarily for any object, it is possible to expand the domain to include a singleton of that object. This corresponds externally to the fact about the sequence of D_is that for any object a in D_i, there is some $j > i$ such that the singleton of a is in D_j. Reflection will allow one—internally—to derive that it is possible to expand the domain so that, for any object, there is a singleton of that object.

We can say more precisely the effect that (Refl) has by comparing it to reflection principles in standard non-modal set theory. These are schemas of the form:

$$\text{(Refl}_{\text{ZF}}\text{)} \qquad \phi \rightarrow \exists \gamma \phi^{V_\gamma}.$$

where γ ranges over the ordinals, and ϕ^{V_γ} results from ϕ by restricting the quantifiers in it to the γth stage of the cumulative hierarchy. (Refl$_{\text{ZF}}$) and (Refl) can not be compared directly, since the languages involved are quite different. Moreover, the definitions of $\in$ and Set which we are working with involve a modal operator, and so (Refl) does not yet apply to any formulas involving them. But it is simple to extend (Refl) to involve such formulas (without assuming any strengthened principles). Suppose the $(\cdot)^\Diamond$ operation were extended to apply also to formulas involving $\in$ and Set, with these treated as primitives. Then it is simple to check that the extended version of the reflection principle is valid.

Now, it is possible to see the relationship between (Refl) and non-modal reflection principles. Let $\mathcal{M}$ be a 'standard' model of BLV$\downarrow$, like that in section 7.3.1, based on V_α for some ordinal α. So, $W = \alpha$; for $\beta < \alpha$, $\delta(\beta) = V_\beta$; $D = V_\alpha$; D_2 is all functions $f : \alpha \rightarrow \mathcal{P}(V_\alpha)$; for $f \in D_2$ and $\beta < \alpha$, $I(\varepsilon)(f, \beta) = f(\beta) \cap V_\beta$. Then the following relate satisfaction of a formula at a world with satisfaction of formulas on V_α.

Lemma 7.9. *Let $\phi(\vec{x}, \vec{F})$ be a formula of a language with primitives 'ε' and '$\in$' (so a non-modal formula of the language of* EL*, but with '$\in$' treated as primitive) where $\vec{x}$ and $\vec{F}$ are lists of the free variables in ϕ. Then, for any $\beta < \alpha$, $\vec{a} \in V_\alpha$ and $\vec{X} \subseteq V_\alpha$:*

$$\mathcal{M}, \beta \vDash \phi^\Diamond(\vec{a}, \vec{X}') \textit{ iff } V_\alpha \vDash \phi^{[\beta]}(\vec{a}, \vec{X})$$

where:

- *For any $X \subseteq V_\alpha$, $X' : \alpha \rightarrow \mathcal{P}(V_\alpha)$ such that for any $\beta < \alpha$, $X'(\beta) = X$.*
- *$\phi^{[\beta]}$ is a formula in the standard language of (second-order) which results from replacing occurrences of εX by $X \cap V_\beta$.*[17]

[17] Some care is perhaps needed here. '$X \cap V_\beta$' will not be a term in the language of set theory. It can be eliminated in the usual way in which such terms are, and then treated as a formula with V_β being an additional parameter.

Proof. By induction on formula complexity:

Atomic formulas: There are several cases to consider: there are three kinds of atomic formula—$a = b$, $a \in b$ and Fa—and for each of these, the terms may be a variable or an abstract term. In every case, $\phi^\Diamond$ will just be ϕ. The interesting cases are for $a \in b$ and for when a term is an abstract term. For the first:

$$\begin{aligned}
\beta \vDash a \in b \quad &\Leftrightarrow \quad \beta \vDash \Diamond \exists F \exists x (a = x \wedge b = \varepsilon F \wedge Fa) \\
&\Leftrightarrow \quad \exists \gamma \geq \beta, \exists f \in D_2 \text{ s.t. } a \in V_\gamma, b = f(\gamma) \cap V_\gamma \text{ and } a \in f(\gamma) \\
&\Leftrightarrow \quad a \in b \\
&\Leftrightarrow \quad V_\alpha \vDash a \in b
\end{aligned}$$

For the second, we have, for example:

$$\begin{aligned}
\beta \vDash a = \varepsilon X' \quad &\Leftrightarrow \quad a = X'(\beta) \cap V_\beta = X \cap V_\beta \\
&\Leftrightarrow \quad V_\alpha \vDash a = X \cap V_\beta \\
&\Leftrightarrow \quad V_\alpha \vDash (a = \varepsilon X)^{[\beta]}
\end{aligned}$$

The situation is similar for other atomic formulas containing abstract terms.
Connectives: For conjunction, we have:

$$\begin{aligned}
\beta \vDash (\phi \wedge \psi)^\Diamond \quad &\Leftrightarrow \quad \beta \vDash \phi^\Diamond \text{ and } \beta \vDash \psi^\Diamond \\
&\Leftrightarrow \quad V_\alpha \vDash \phi^{[\beta]} \text{ and } V_\alpha \vDash \psi^{[\beta]} \\
&\Leftrightarrow \quad V_\alpha \vDash (\phi \wedge \psi)^{[\beta]}
\end{aligned}$$

The situation is similar for negation.
First-order quantification: We have,

$$\begin{aligned}
\beta \vDash (\forall x \phi)^\Diamond \quad &\Leftrightarrow \quad \beta \vDash \Box \forall x \downarrow (\phi^\Diamond) \\
&\Leftrightarrow \quad \text{for all } a \in V_\alpha, \beta \vDash \phi^\Diamond(a) \\
&\Leftrightarrow \quad \text{for all } a \in V_\alpha, V_\alpha \vDash \phi^{[\beta]}(a) \\
&\Leftrightarrow \quad V_\alpha \vDash (\forall x \phi)^{[\beta]}
\end{aligned}$$

Second-order quantification: First note that, for a non-modal formula, and $f : \alpha \to \mathcal{P}(V_\alpha)$, whether $\beta \vDash \phi(f)$ and whether $\beta \vDash \phi^\Diamond(f)$ depend only the value of f at β (this can be proved by a simple induction on formula complexity). For any such f, let $f_\beta \subset V_\alpha$ be $f(\beta)$, so $\beta \vDash \phi(f)$ iff $\beta \vDash \phi(f'_\beta)$ and $\beta \vDash \phi^\Diamond(f)$ iff $\beta \vDash \phi^\Diamond(f'_\beta)$. Then we have the following:

$$\begin{aligned}
\beta \vDash (\forall F \phi)^\Diamond \quad &\Leftrightarrow \quad \text{for all } f : \alpha \to \mathcal{P}(V_\alpha), \beta \vDash \phi^\Diamond(f) \\
(*) \qquad &\Rightarrow \quad \text{for all } X \subseteq V_\alpha, \beta \vDash \phi^\Diamond(X') \\
&\Leftrightarrow \quad \text{for all } X \subseteq V_\alpha, V_\alpha \vDash \phi^{[\beta]}(X) \\
&\Leftrightarrow \quad V_\alpha \vDash (\forall F \phi)^{[\beta]}
\end{aligned}$$

The converse of $(*)$ can be proved as follows. Suppose that $f : \alpha \to \mathcal{P}(V_\alpha)$. Since $f_\beta \subseteq V_\alpha$, $\beta \vDash \phi^\Diamond(f'_\beta)$. But, by the above consideration, $\beta \vDash \phi^\Diamond(f)$ as required. □

For formulas taken from the usual language of set theory without abstraction operators, this is a more natural relationship, since in that case, $\phi^{[\beta]}$ just is ϕ.

Before stating and proving the relationship between (Refl) and (Refl_{ZF}), one additional lemma will be useful, relating $\phi^{[\beta]}$ and $\phi^{\downarrow}$:

Lemma 7.10. *For $\beta < \gamma$,*

$$V_\gamma \vDash \phi^{[\beta]} \textit{ iff } \langle \beta, \gamma \rangle \vDash \phi^{\downarrow}$$

Proof. By induction on formula complexity:

Atomic formulas: The only interesting case is where a term involved is an abstract term. In which case, we have, for example:

$$\begin{aligned} V_\gamma \vDash (a = \varepsilon X)^{[\beta]} &\Leftrightarrow V_\gamma \vDash a = X \cap V_\beta \\ &\Leftrightarrow a = X \cap V_\beta \\ &\Leftrightarrow a = ||\varepsilon X||^{\beta} \\ &\Leftrightarrow a = ||{\downarrow}\varepsilon X||^{\langle \beta, \gamma \rangle} \\ &\Leftrightarrow \langle \beta, \gamma \rangle \vDash (a = \varepsilon X)^{\downarrow} \end{aligned}$$

Connectives: This is trivial.
First-order quantification: We have that:

$$\begin{aligned} V_\gamma \vDash (\forall x \phi)^{[\beta]} &\Leftrightarrow \text{for all } a \in V_\gamma,\ V_\gamma \vDash \phi^{[\beta]}(a) \\ &\Leftrightarrow \text{for all } a \in V\gamma,\ \langle \beta, \gamma \rangle \vDash \phi^{\downarrow}(a) \\ &\Leftrightarrow \langle \beta, \gamma \rangle \vDash \phi^{\downarrow}(a) \end{aligned}$$

Second-order quantification: We have that:

$$\begin{aligned} V_\gamma \vDash (\forall F \phi)^{[\beta]} &\Leftrightarrow \text{for all } X \subseteq V_\gamma,\ V_\gamma \vDash \phi^{[\beta]}(X) \\ &\Leftrightarrow \text{for all } X \subseteq V_\gamma,\ \langle \beta, \gamma \rangle \vDash \phi^{\downarrow}(X) \\ &\Leftarrow \text{for all } f \in D_2,\ \langle \beta, \gamma \rangle \vDash \phi^{\downarrow}(f) \\ &\Leftrightarrow \langle \beta, \gamma \rangle \vDash \phi^{\downarrow} \end{aligned}$$

□

Finally, we can now prove the following:

Theorem 7.11. *Let α be any ordinal, and let $\mathcal{M}$ be a structure as described before. Then $\mathcal{M}$ satisfies the reflection rule (Refl) iff V_α satisfies (Refl_{ZF}).*

Proof. For the left to right direction: Let ϕ be a sentence of the language of standard set theory, with parameters, and suppose that $V_\alpha \vDash \phi$. Then, by lemma 7.9, $0 \vDash \phi^{\Diamond}$. So, by (Refl), $0 \vDash \Diamond\phi^{\downarrow}$. So, for some $0 \leq \gamma < \alpha$, $\gamma \vDash \phi$ (since $\phi^{\downarrow}$ is just ϕ). So, clearly, $V_\alpha \vDash \phi^{V_\gamma}$. Hence $V_\alpha \vDash \phi \rightarrow \exists\gamma\phi^{\phi_\gamma}$.

For the right to left direction: Let ϕ be a non-modal sentence of the language of EL (with parameters), and $\beta < \alpha$ with $\beta \vDash \phi^{\Diamond}$. By lemma 7.9, $V_\alpha \vDash \phi^{[\beta]}$. So, by ($\text{Refl}_{\text{ZF}}$), for some $\gamma < \alpha$, $V_\gamma \vDash \phi^{[\beta]}$. Then by lemma 7.10, $\langle \beta, \gamma \rangle \vDash \phi^{\downarrow}$. So $\beta \vDash \Diamond\phi^{\downarrow}$ as required. □

This result then tells us that we have good reason to accept the consistency of EL+BLV↓+Refl. The soundness theorem tells us that it will be consistent if it has a model, and proposition 7.11 tells us that it will have a model if a standard second-order reflection principle has a model. It also gives reason to be optimistic about being able to derive strong consequences. Second-order reflection principles are very strong, and what has been proved is an equivalence of sorts between these and (Refl). But this alone does not assure that such strong consequences will be forthcoming. Firstly, it may be that not all models of BLV↓+Refl are of the form V_α. Secondly, even if they are, the step from a sentence holding in all models of BLV↓+Refl to it being a deductive consequence would require completeness, and second-order logic is not complete. However, it turns out that much of standard set theory can be derived from BLV↓+Refl, as the next section will show.

7.7 Interpreting set theory, again

Having accepted (Refl), the task of proving $\Diamond Z$—where Z is the conjunction of suitably many axioms of set theory—appears more achievable. There are two ways it does this. Firstly, a simple consequence of (Refl) and BLV↓ is

$$(*) \qquad \Diamond\forall F\exists y(y = {\downarrow}\varepsilon F)$$

simply by applying (Refl) to $\forall F\Diamond\exists y{\downarrow}(y = \varepsilon F)$, which follows directly from BLV↓ and ($\Diamond E$!-I). That is, as well as each concept possibly forming a set, it is possible for them all to form a set at once. This is the alternative to the maximality idea which is expressed in section 7.4. This is crucial in proving, for example, the power set axiom, since it can be used to show that all the subsets of some set can lie in the domain of quantification at once.

The second reason that (Refl) makes proving $\Diamond Z$ more achievable is that, in many cases, where ϕ is an axiom, it is possible to prove $\phi^\Diamond$. The conjunction of these can then have the reflection principle applied to it (and since they will not involve the abstraction operator, $\phi^\downarrow$ in each case will just be ϕ). In fact, a fairly general comprehension principle of this sort is a simple consequence of proposition 7.7. It is:

$$\Diamond \mathrm{R}_0^x\text{-}\phi \wedge \Box \mathrm{R}_1^x\text{-}\phi \rightarrow \Diamond\exists y{\downarrow}(\mathrm{Set}\, y \wedge \Box\forall x{\downarrow}(x \in y \leftrightarrow \phi))$$

(Adding the ↓ operators to proposition 7.7 essentially has no effect due to the rigidity of ϕ and membership.)

Now, recall the following set existence axioms of ZF set theory with urelements:

Empty set $\exists x(\mathrm{Set}(x) \wedge \forall y(y \notin x))$

Pairing $\exists x\forall y(y \in x \leftrightarrow y = u \vee y = v)$

Union $\exists x\forall y(y \in x \leftrightarrow \exists z(z \in u \wedge y \in z))$

Power set $\exists x\forall y(y \in x \leftrightarrow y \subseteq u)$

Infinity $\exists x(\varnothing \in x \wedge \forall y(y \in x \rightarrow y \cup \{y\} \in x))$

Replacement $\forall R\big[\forall x\exists y\forall z(Rxy \leftrightarrow y = z) \rightarrow \exists x\forall y(y \in x \leftrightarrow \exists z(z \in u \wedge Rzy))\big]$

For many of these, the appropriate formulas can be proved to be rigid in the appropriate way.

Rigidity of the appropriate formulas for empty set and pairing are relatively simple to prove—they follow directly from the rigidity of identity. Union is also relatively simple to prove. Corollary 7.8 is needed to ensure that it is possible that the required formula is R_0.

The interesting cases are power set and replacement, which require an additional use of reflection to prove possible rigidity, and infinity, which requires reflection to prove an appropriate version of.

7.7.1 Power set

We need to prove that $\Diamond \mathrm{R}_0^x\text{-}(x \subseteq u)^\Diamond$, where $\subseteq$ is defined in the normal way in terms of $\in$ as:

$$x \subseteq y \stackrel{\text{def}}{=} \mathrm{Set}(x) \wedge \mathrm{Set}(y) \wedge \forall z(z \in x \to z \in y)$$

To do so, it will be useful to look at the properties of the relations $x \subseteq y$ and $(x \subseteq y)^\Diamond$. Firstly, it can be noted that $(x \subseteq y)^\Diamond$ behaves in similar ways to $\in_1$, in that it is rigid in the second sense:

Proposition 7.12. $\mathrm{BLV}{\downarrow} \vdash \mathrm{R}_1^x\text{-}(x \subseteq u)^\Diamond$

Proof. The proof is very similar to those of propositions 7.3 and 7.4. First, it can be shown that $x \subseteq u \to \Box(x \subseteq u)$. By the rigidity of membership, $(x \subseteq u)^\Diamond$ is equivalent to $\Diamond(x \subseteq u)$. So, we have $\Diamond(x \subseteq y) \to \Diamond\Box(x \subseteq u)$ and hence $\Box(x \subseteq y)^\Diamond$ by (G). This then allows rigidity to be proved.

As always , see appendix B for a full proof. □

It is then possible to prove:

Proposition 7.13. $\mathrm{BLV}{\downarrow} + \textit{(Refl)} \vdash \Diamond \mathrm{R}_0^x\text{-}x \subseteq u$

Proof. The basic idea of the proof is as follows. Firstly, it will be possible that $\mathrm{Set}_0(u)$ and also that $\Box\forall x((x \subseteq u)^\Diamond \to {\downarrow}(\mathrm{Set}_0(x)))$, by using propositions 7.6 and 7.12. Then, by using $(*)$, it is possible that $\Box\forall x(x \subseteq y \to {\downarrow}(E!x))$. See appendix B for full details. □

So, by propositions 7.13 and 7.7, (Power set)$^\Diamond$ will be provable.

7.7.2 Replacement and infinity

I will not give detailed proofs of replacement and infinity here. I shall simply note that it is a known result that, in a non-modal setting, it is known that both infinity and replacement follow from a second-order reflection principle (Bernays, 1976). Similar methods can be used to prove the modalised versions of infinity and replacement from the modal reflection principle.

7.8 Conclusion

In this chapter, I have shown how the technical part of a program of expansionist abstraction might be carried out. I doing so, I have also demonstrated a number of advantages that the approach may have over the static approach. For one thing, the bad company problem does not arise; even BLV is consistent. Secondly, the approach allows for a significant amount of set theory to be developed.[18]

So, the second of the major tasks that I highlighted in chapter 5 has at least been partially filled. The main obvious lacuna is in my motivation for the reflection principle. The motivation that I gave was brief. Moreover, it even seems to suggest a threat to the expansionist approach; a spectre of absolutely unrestricted quantification seems to lurk within the reflection principle. Confronting this spectre will be the task of my next chapter.

[18]Moreover, this set theory is developed in a way which is suggestive of the iterative conception of set. This may be advantageous if one thinks (like, e.g. Potter (2010)) that there is something wrong with the limitation of size approach to set theory which is shared by most of the approaches to set theory within the static framework.

Chapter 8

Conservativeness, diagonalisation and reflection

8.1 Introduction

The aim of this chapter is to motivate more fully the reflection principle which plays a key role in the development of the previous chapter. In doing so, however, a potential problem arises for the overall position which I have laid out concerning domain expansion. That problem is that, given the account of definition of quantifiers given in chapter 6, it seems that it may be possible to define an absolutely unrestricted quantifier.

In section 8.2 I introduce the main worry, that it seems possible to introduce an interpretation of the quantifiers which makes them, in effect, absolutely general. This problem is the one hinted at in 7.6, concerning the similarity of the reflection principle to a motivation for collapsing all interpretations into a single, absolute, interpretation. In 8.3, I consider two potential ways out of the problem, but ultimately reject them. Section 8.4 is somewhat of a digression. In it I discuss what restrictions there must be on an implicit definition of a quantifier—namely that, in some sense it must not conflict with the meanings already given to other vocabulary. The reason this is being discussed here is that it will play a role in resolving the aforementioned tension. In section 8.5 I then show how the initial worry can be avoided by reference to such restrictions. This will not solve the tension definitively in favour of the relativist, but rather provides a way out for either the relativist or the absolutist to appeal to extrinsic reasons to hold one or the other position.

Finally, in section 8.6, I turn back to the problem of motivating the reflection principle, and show how this can be done without falling into the problem.

8.2 Defining absolutely unrestricted quantification?

Consider the following candidate as a definition of a quantifier.[1]

$$(\mathrm{Def}_{AU}) \qquad \exists^{\mathrm{AU}} x\phi \leftrightarrow \Diamond \exists x{\downarrow}\phi$$

It can be seen that this satisfies the usual inference rules for quantifiers. Moreover, it can be proved that $\exists^{\mathrm{AU}}$ is absolutely unrestricted. That is, that $\Box\forall y{\downarrow}\exists^{\mathrm{AU}} x(y = x)$ (i.e., no matter how one interprets the quantifier, its domain will not exceed that of $\exists^{AU}$).

This raises a problem for the picture I have given of abstractionist generality relativism. For on the one hand, this definition appears to secure an absolutely unrestricted interpretation of the quantifiers (albeit one which depends essentially on the postulational modality for its elucidation). Moreover, given the view of quantification I gave in chapter 6, it is permissible to interpret the quantifier in such a way. But on the other hand, as has already been pointed out, BLV appears to allow for definitions of expanded domains no matter what domain is started from (by reasoning via Russell's paradox). One of these has to go.

What we have, in effect, is a tension between absolute generality and naïve comprehension. This time, however, the situation is in some ways more precise. Before, there were just competing intuitions between some form of naïve comprehension and absolute generality. Moreover, it was not clear what this form of naïve comprehension was, and what absolute generality amounted to. In this case (assuming the present view concerning the possibility of defining interpretations of quantifiers is accepted), we have two competing *definitions*, both of which seem in equally good standing.

One of these potential definitions must then be rejected. I will consider later how this might be done, by appealing to a principle of conservativeness for definitions of quantifiers.

8.3 Two possible ways out

8.3.1 Weakening comprehension

One option that could be taken is not to restrict which of the two definitions of quantifiers to accept, but instead to restrict which concepts there are, by weakening the comprehension principle. Recall that this is (in the present setting):

$$\exists F\Box\forall x(Fx \leftrightarrow \phi(x))$$

This principle embodies the idea that concepts are to be individuated *intensionally*. For example, given predicates which are extensionally equivalent, but perhaps not necessarily so, they can not be assigned the same concept according to this schema.[2] In particular,

[1] Williamson (forthcoming) considers something similar in the case where the modality is of a more usual metaphysical modality. His aim is not to *define* a new quantifier, but rather to provide a method by which a non-believer in possibilia can express some of the things that their opponent can.

[2] For example, consider a standard example of coextensive predicates which are not necessarily so: 'has a heart' and 'has a liver'. If these were assigned the same concept F, then it would be possible to reason using the above schema as follows. Firstly, we could see that, necessarily for all x, Fx iff x has a liver (by instantiating $\phi(x)$ by 'has a liver'). Secondly, necessarily for all x, Fx iff x has a heart (by instantiating $\phi(x)$ by 'has a heart'). Hence, necessarily everything which has heart has a liver and vice-versa, which is (or so it is supposed) false.

due to the appearance of the necessity operator in the schema, what concepts there are does not depend on the present domain of the first-order quantifiers.

The restriction which will be needed will, at the very least, need to rule out the existence of a concept R such that $\forall^{AU} x(Rx \leftrightarrow x \notin x)$ (i.e. one which would define a Russell set), and perhaps also a concept U such that $\forall^{AU} xUx$. But this raises a dilemma for a proponent of this strategy. Either the indifference of the comprehension principle to the first-order domain of quantification must be dropped, or it may be maintained. If the second of these cases, the concept R must be banned at every domain. But then, since it is such a concept which drives the domain expansion, the resulting set theory will be much weakened. I suspect that, for natural choices of restrictions, the resulting theory will be one of similar strength to BLV with a fixed domain and predicative comprehension (Burgess, 2005, pp.87–92), which is very weak—weaker even than first-order PA.

In the first case, it seems that we can no longer treat the second-order quantifier as ranging over concepts or properties (or perhaps substitutionally). For why should the existence or not of a property depend on what objects there are? There are two such dependences which may be plausible—denying the existence of non-instantiated properties for Aristotelian reasons, and having the existence of properties tied to the existence of objects to which their defining predicates refer—but neither of these would rule out the problematic properties R and U.

Alternatively, if the second-order quantifiers were interpreted *extensionally*, perhaps as plural quantifiers, then this may leave room for such a restriction,[3] although it is unclear what the restricted comprehension principle would be if it is to allow certain predicates to define a plurality in some cases but not others. But in any case, this option too looks very problematic. For, in that case, one would need to say that $\exists^{AU}$ is indeed a quantifier, but that it is not the case that there are some things which are all and only those things which it quantifies over. Moreover, the plural definite description 'the things being quantified over' or 'all the things' would have to be deemed to be empty (and not just in the sense that it refers to an 'empty' plurality).

8.3.2 Diagonalising the modality

A more radical suggestion would be to claim that the combination simply shows that the modality fails to deliver on the expressiveness that it promises. Just as, according to the generality relativist, BLV allows one to 'diagonalise out' any particular use of the quantifier (i.e. shows that it fails to be absolutely general, reasoning via Russell's paradox), the proponent of this solution (who, for the moment, I shall call a 'diagonalist') claims that BLV together with Def_{AU} allows one to diagonalise out the *modality*.

What do I mean by this? The relativist pictures a scenario in which a user of a quantifier is shown that she can not be quantifying absolutely generally, by presenting her with an object (namely, the Russell set), which must not lie in her domain of quantification. The relativist then has a problem in expressing the general conclusion of this (namely, generality relativism), as previously discussed. It will be no good to try and express the thought by using quantification, by, for example, saying that *every* domain of quantification fails to be absolutely unrestricted. By the relativists own lights, the quantifier used is not absolutely unrestricted, and so can not preclude there being an absolutely unrestricted domain which simply did not fall under the quantifier used.

[3] Such an approach—of taking sets to be formed out of pluralities and to restrict the plural comprehension principle—is taken by Linnebo (2010a).

So, the relativist adopts a modal formulation instead, claiming that absolutely universal quantification is not *possible*. But, in order that this does not suffer the same fate as the quantified formulation, the modality involved must be *absolute*; if the notion of possibility is too narrow, then the formulation may allow for an absolutely unrestricted quantifier which is possible in a wider, more permissive sense than that which is used.

The diagonalist claims that it is this required absoluteness of the modality which fails. He claims that, just as the relativist can show that any quantifier is not absolutely unrestricted, he can show that any use of postulational modality fails to be appropriately absolute. This is done by presenting an interpretation of the quantifier which is evidently in good standing (defined by Def_{AU}), but which was not envisaged by the use of the modality (on pain of contradiction).

Now, the diagonalist proposal is a very radical position. It runs into severe difficulties when it comes to the question of how it is to be expressed, and it has drastic implications which go well beyond the present project. Whereas the relativist can respond to worries about expressing her position by making use of modality, the diagonalist has no such option—any attempt to do so (by, for example, quantifying over modalities) will surely be subject to the same diagonalisation process. The diagonalist must then resort to some sort of 'militant quietism'(cf. Button (2010)). According to this position, relativists (and, in this case, diagonalists), must refrain from adopting and stating a definite position. Instead, they should be content in waiting until somebody claims to have an absolutely unrestricted quantifier (or, in this case, an absolute postulational modality), and show that it in fact fails to be absolutely unrestricted (or absolute).

The diagonalist position (if arrived at from this perspective) will also have fairly radical consequences for modality in general. Since postulational modality is explained in terms of a more ordinary 'circumstantial' modality, properties of the former are likely to rub off on the latter. In the case of generality relativism—as expressed by the modality—a consequence is that the concept of a possible world (or similar notion) must be indefinitely extensible. Otherwise, the modality would simply reduce to quantification over (absolutely all) worlds, and it is a crucial part of the position that the modality must not be reducible to quantification. In particular, there will be an indefinitely extendible sequence of possible worlds in which the quantifier is interpreted in an increasingly expansive way.

But for the diagonalist position, the consequences are more extreme. For now it is not the concept of a possible world that is indefinitely extensible, but rather the strength of the modality itself. Consider the common picture of nested modalities, with say, physical possibility being more expansive than biological, metaphysical being more expansive than physical, and so on. According to the picture here, this sequence will be indefinitely extensible, with each element envisaging more possible interpretations. The relativist accepts a notion of absolute modality, but simply rejects that this is equivalent to quantification over any particular domain of worlds (or similar). The diagonalist, by contrast, must reject absolute modality altogether, instead postulating an indefinitely extendible sequence of ever stronger and stronger notions.

The aim was to survey some alternative ways out of the problem of competing definitions before coming on to my preferred solution in terms of conservativeness. In the previous case—restricting comprehension—we saw that there were some concrete proposals for how to modify the framework, but which ultimately had to be rejected. But in this case, it does not appear to be so much a way out of the problem, but instead

seems to require giving up on the project (since, according to it, there is no absolute modality which can be used). But perhaps something can be salvaged, whilst rejecting that the modalities involved are absolute.

How would such a salvaging go? Suppose that the diagonalist simply proceeds with the project of generating some set theory with Basic Law V together with some principles of domain expansion expressed in modal terms. To do this, they must of course not claim that the modality is absolute. Is such an approach viable? Some things inevitably will have to be given up. For one thing, the expression of generality relativism in terms of modality—that necessarily, it is possible to expand the domain, or something similar—no longer does what it aimed to do; it at best merely hints at a position. But this is no loss for the diagonalist, for they have already acknowledged that they can not express their view. Other conclusions stated in terms of the modality are also more-or-less unaffected. For example, the overall aim of the project will still be to claim that it is possible to reinterpret the quantifiers, making use of abstraction principles, so that they, in effect, range over a universe of sets. This is not significantly affected if it is conceded that the modality involved in expressing it is not absolute.

The final thing that must be asked is then what to make of Def_{AU}. By rejecting absolute modality, there is no particular problem. For what Def_{AU} does is show (according to the diagonalist) that the modality involved in its statement is not absolute. The genuine contradiction between Def_{AU} and BLV occurs when we are permitted to infer from some sentence ϕ, where each quantifier is $\exists^{AU}$ or $\forall^{AU}$, to the claim that ϕ is postulationally possible. This the diagonalist can concede, but under the proviso that the possibility involved in the conclusion may be different (and, in particular, weaker) than that involved in the definition of the quantifiers appearing in the premise.

There are still questions to be asked of diagonalist position, and in particular the militant quietism that it requires. What are we to make of the idea of a position which claims to be unstatable, for example? As such, I shall not consider the position further, but instead I will go on to argue that the impasse between the two competing definitions can be solved by rejecting one or the other by appealing to conservativeness requirements.

8.4 Conservativeness

Might there be some way to *reject* the proposed definition? I will claim that there is. But before doing so, more needs to be said about the framework in which such a solution will be given. In order to avoid the difficulty raised by conflicting definitions of quantifiers, it needs to be considered what there are in the way of necessary and sufficient conditions for an implicit definition of a quantifier to be acceptable. And there must indeed be restrictions. For example, it surely must be the case that it would be illegitimate to define a quantifier in such a way that, according to it, there is a natural number less that zero, or so that the existence of a Higgs boson becomes a matter of stipulation.

Now, as already discussed, one such requirement must be that the defined symbols have the appropriate inferential behaviour for quantifiers. Although this will of course be a *necessary* condition, it will not in general be sufficient. For example, Williamson (2003, pp.440-442) and Sider (2009, pp.391–392) give examples of ways to supposedly define a quantifier which satisfies the correct inference rules, but clearly should not be acceptable (so that, for example, according to the new quantifier, there are talking

donkeys).[4]

Given the view on quantifiers which I have previously advocated (namely, that the acceptability of a quantifier meaning is to be justified by reference to the context principle and syntactic priority thesis), there is one potential requirement for acceptability which must be rejected. That requirement is that a candidate quantifier meaning must correspond in some strong way with some absolute feature of the world, for example by 'latching on' in some way to a domain of quantification, or by 'carving nature at the joints'.[5] Any restrictions on what is to count as acceptably giving the meaning of a new quantifier must, like the condition of inferential adequacy, be of a more logical or linguistic character.

In particular, the kind of restriction that I have in mind will roughly be that a definition must not interfere (in some sense) with preexisting vocabulary and its meaning.[6] Since the general issues of meaning which are involved go quite some way beyond the scope of the present project, I will not be giving a single criterion which I claim to be the correct way of spelling out acceptability. Rather, I shall give the specification somewhat schematically, depending on notions which may be selected according to various commitments elsewhere. I do, however, claim that any reasonable such selection will be adequate for the use to which I put the restriction.

In general, a constraint will then be, roughly, that an implicit definition must not require a change in the meaning of any vocabulary which already has a given meaning.[7] What might such a constraint look like? My aim here will not be to argue for any one particular way of stating such a constraint. It will instead be to survey a range of possible constraints (which of these is most suitable will depend on issues concerning how the meaning of language is fixed in general, and is beyond the scope of this project). But each of the constraints considered will be suitable for the use to which I will put them—namely, to help in avoiding an apparent tension in the view that the meaning of quantifiers may be given by an implicit definition.

The aim is as follows. Given some 'starting' language $\mathcal{L}$, we want to know when a definition D—which will be a sentence in an extended language $\mathcal{L}^+$ consisting of $\mathcal{L}$ together with the symbol or word to be defined—does not conflict with, or does not require going beyond, the meanings given to the components of $\mathcal{L}$. The strategy to achieve this aim will roughly be the following. Given $\mathcal{L}$, we will identify a privileged

[4]Williamson's construction goes via giving a model-theoretic semantics for such a quantifier, Sider's by defining $\exists^+ x\phi$ to be true just in case some particular 'person who is logically perfect, maximally opinionated, and totally nuts' (p.391) believes '$\exists x\phi$'.

[5]This attitude is then similar in some ways to that taken by Hale and Wright (2009b) concerning when abstraction principles are successful. They reject the idea that part of what it takes for some abstraction principle to be acceptable is '*hitting off* reference to a range of entities qualified to play the role that the principle defines' (p.206). Also, see Sider (2009) on the topic of joint carving.

[6]Sider (2009) seems to suggest that some approach like this might succeed in ruling out his purported counterexample. In particular, he concedes that in his example '[i]ntuitively, the candidate meanings . . . assign names and predicates different meanings from their English ones' (p.392 fn.24).

[7]This is somewhat of a simplification. In many cases, it may be desirable when shifting to a larger domain to redefine existing vocabulary. For example, a shift from a domain of real numbers to one of complex may require that the exponential function has its definition widened to include complex arguments, and consequently the logarithm function would even need to be redefined on the reals, so as to be multi- rather than single-valued. See, for example, Buzaglo (2002).

But the possibility of such *concept expansion* should not affect the requirement of fixed meaning here. For the kind of conflict in mind here is that between the definition of the new quantifier and the old vocabulary *as already understood*.

set of (true) sentences, $T_{\mathcal{L}}$, of $\mathcal{L}$ which some way or other hold information about the meaning of components of $\mathcal{L}$. This set will consist of something like those sentences of $\mathcal{L}$ whose truth is settled by the meaning of $\mathcal{L}$. A candidate definition D will then be ruled out as conflicting with the meaning of $\mathcal{L}$ if it has consequences which conflict with $T_{\mathcal{L}}$. There are two ways of conceiving of this conflict—either as D having consequences which *contradict* some sentences of T, or as D having consequences, expressible in $\mathcal{L}$, which go beyond T in that they do not appear in T. Symbolically, let us say that some definition D is:

1. *Conservative* over $\mathcal{L}$ iff for all ϕ in the language $\mathcal{L}$, if $D, T_{\mathcal{L}} \vDash_{\mathcal{L}} \phi$, then $\phi \in T_{\mathcal{L}}$,
2. *Consistent* over $\mathcal{L}$ iff for all ϕ in the language $\mathcal{L}$, if $D, T_{\mathcal{L}} \vDash_{\mathcal{L}} \phi$, then $\neg\phi \notin T_{\mathcal{L}}$,

This proposal, as it currently stands, is obviously very unclear; there are two items in the above definitions—$T_{\mathcal{L}}$ and $\vDash_{\mathcal{L}}$—which have not been adequately explained. Moreover, it is not evident why such conditions might serve as sufficient conditions for a definition to be acceptable. My intention is that this unclarity can be resolved in one of many directions, depending on issues which go beyond the scope of the present project, and that, for many such ways of clarifying, the result is a plausible restriction on implicit definitions. There are two components of the foregoing sketch that require fleshing out: (a) What, in general, do we select for $T_{\mathcal{L}}$? That is, what is meant by a sentence being 'settled by the meaning of $\mathcal{L}$'?. (b) What is the notion of consequence involved in saying that some sentence is a consequence of D? This may not be as simple as considering only logical consequence,[8] since in many cases one may think that what counts as a consequence of D may depend on the meaning of $\mathcal{L}$ in a way which can not be captured by a judicious choice of $T_{\mathcal{L}}$. In order to make it clearer what may be chosen for these options (a), (b), I shall give some examples. I will then say why, for at least some combinations of these, one of conservativeness or consistency will be plausible as a restriction on definitions. Moreover, given a rejection of requirements of joint-carving, hitting-off, and the like, they will be plausible as forming necessary and sufficient conditions along with inferential adequacy.

8.4.1 (a) What do we take for $T_{\mathcal{L}}$?

The first gap that needs to be filled out in order to make the proposal clearer is that of what $T_{\mathcal{L}}$ should be. The question is not what particular set $T_{\mathcal{L}}$ should be for some particular language, such as the language of arithmetic. Before *that* may be answered, a more general question must be answered. That question is: what requirement should be made of $T_{\mathcal{L}}$ so that the resulting restriction on acceptable definitions is plausible? That is, for some particular language, how do we choose what to take for $T_{\mathcal{L}}$? As mentioned earlier, the aims will be to identify a set of sentences which, in some sense, are those settled purely by matters of meaning, though I shall also consider kinds of restrictions which are different from this, but that nonetheless lead to a plausible restriction.

[8] Of course, even if it is just logical consequence that is involved, it is contentious what even *this* amounts to.

(i) Analyticity

One option—and perhaps the most simple—would be to take some notion of analyticity as given, and then let $T_{\mathcal{L}}$ be the set of all those sentences of $\mathcal{L}$ which are analytic. Now—assuming that the notion of analyticity involved is in good standing, and making plausible assumptions about the consequence relation involved—this will result in conservativeness being a plausible restriction on definitions. The aim is that D should count as a definition and hence be analytic. So, if ϕ follows from T and D, it too should be analytic, assuming that the consequence relation involved preserves analyticity. But then, if ϕ only contains vocabulary from the old language $\mathcal{L}$, it must be in $T_{\mathcal{L}}$. So, if there is a ϕ from $\mathcal{L}$ which follows from D and $T_{\mathcal{L}}$ but is not in $T_{\mathcal{L}}$, some of the meaning of the $\mathcal{L}$ must be assumed to have changed—for something is now considered analytic which previously was not. Hence, a definition which is not conservative in this sense can be no good.

(ii) Theories of meaning

A related option, which perhaps simply amounts to a more sophisticated version of the previous one, would be to take $T_{\mathcal{L}}$ to be determined by a metatheoretic theory of meaning of $\mathcal{L}$ (in the sense of Davidson (1967)). If one has such a theory of meaning, then $T_{\mathcal{L}}$ can be the set of sentences in $\mathcal{L}$ whose truth is a theorem of the theory of meaning. There are a couple of remarks that may be made abut such an approach. Firstly, in the case where the language in question has some resources to talk about its own semantics (as, for example, it may be thought in the case of English, or other natural languages), much of the theory of meaning itself may be replicated in $T_{\mathcal{L}}$. As such, one may perhaps say that we simply let $T_{\mathcal{L}}$ be a theory of meaning for $\mathcal{L}$ (albeit one expressed in the object language). Secondly, it may be the case that what is determined to be analytic in this sense may be relative to some notion of analyticity in the metalanguage; thus, we do not have some absolute notion of analyticity (as in the previous option), but one which is relative to analyticity in another language.

Now, this option will result in conservativeness being a plausible option for much the same reason as the previous one. Since D should count as a definition, the theory of meaning for the extended language $\mathcal{L}^+$ had better have its truth as a theorem, either by adopting something like '$\mathrm{Tr}(D)$' as an axiom, or by adopting a metatheoretic version of D as an axiom from which its truth follows. For example, if one were to add the word 'bachelor' to a language for the first time, by explicit definition, one may expect the theory of meaning of the extended language to feature some axiom such as

$$\mathrm{Tr}(\ulcorner \mathrm{Bachelor}(x) \urcorner) \text{ iff } \left[\mathrm{Tr}(\ulcorner \mathrm{Unmarried}(x) \urcorner) \text{ and } \mathrm{Tr}(\ulcorner \mathrm{Man}(x) \urcorner)\right]$$

from which '$\mathrm{Tr}(\ulcorner \mathrm{Bachelor}(x) \leftrightarrow \mathrm{Unmarried}(x) \wedge \mathrm{Man}(x) \urcorner)$' will presumably follow as a theorem.

In addition, the consequence relation $\vDash_{\mathcal{L}}$ will be truth preserving in the sense that, if ϕ is determined to be true by the theory of meaning, and $\phi \vDash_{\mathcal{L}} \psi$, then ψ will be determined to be true according to the theory of meaning (this is essentially just the same as saying that consequence preserves analyticity). So, if D is non-conservative, the meaning of parts of $\mathcal{L}$ must change, since there will be theorems of the theory of meaning of $\mathcal{L}^+$ which only concern $\mathcal{L}$, but which are not theorems of the original

theory of meaning of $\mathcal{L}$. Hence, again, non-conservative purported definitions must be disallowed.

(iii) A holistic approach

The two preceding options seem to rely on having a notion of analyticity in some way—either directly as in the first case, or indirectly as in the second case. It may thus seem that the notions of conservativeness and consistency which have been suggested above are unavailable to those—like Quine (1951)—who eschew such notions, or at least who claim that no clear dividing line may be made between those sentences of a language which are analytic, and those which are not. But it may be possible for the criteria to be used by making use of a different choice of $T_{\mathcal{L}}$, and one which is more compatible with a holistic view of meaning according to which there is no clear demarcation between the analytic and the synthetic.

So, for example, one could take for $T_{\mathcal{L}}$, not those sentences which are *analytic*, but rather sentences from a language which are chosen, for pragmatic reasons, to hold come what may. Or, perhaps, since such a view may deny that there are any sentences which can be held absolutely fixed, $T_{\mathcal{L}}$ may instead consist of those sentences that are *sufficiently* fixed—those closer to the center of the web of belief, as it were. Such a choice of $T_{\mathcal{L}}$ would include logical truths, and presumably most (or all) of those sentences which are commonly taken to be paradigmatically analytically true. But there would be no claim that the membership of some sentence ϕ in $T_{\mathcal{L}}$ is a matter of absolute meaning, but that it is simply a matter of pragmatics.

Given such a conception of what $T_{\mathcal{L}}$ is, either of conservativeness or consistency may seem like good choices as a restriction on implicit definitions. Conservativeness might be considered for the same reasons as for when $T_{\mathcal{L}}$ is analyticity, since $T_{\mathcal{L}}$ under the current proposal is essentially just a substitute for analyticity. But the weaker requirement of consistency might also be an option instead of conservativeness. For perhaps, if a definition has particularly desirable uses or consequences, it might be permitted that it requires some sentences of the original language to be elevated in status. But what must be ruled out instead, is that the definition does not contradict one of the original 'fixed' sentences.

Now, this option is not without its difficulties. For one thing, it may result in membership of $T_{\mathcal{L}}$—and hence conservativeness and consistency—becoming a matter of degree. But perhaps modifications can be made to allow for this by, for example, letting $T_{\mathcal{L}}$ be a fuzzy set. For another thing, it is not clear to what extent a proposal similar to this would be acceptable to a Quinean (and, even if it were, which of conservativeness or consistency is preferable). In any case, I shall say no more about this particular option. I will simply refer to those elements of $T_{\mathcal{L}}$ as being those which are analytic, especially as, in the cases I will be interested in, the terms involved will have been introduced by laying down definitions, which can then presumably be ranked as being analytic. My intention in presenting this option has simply been to indicate that it may be possible to adapt such a restriction on definitions without relying on the concept of analyticity, if that were to be desired.

8.4.2 (b) Consequence

The second gap that must be filled, in order for conservativeness and completeness as given above to be properly defined, is the consequence relation $\models_{\mathcal{L}}$. The role that this played in establishing the plausibility of conservativeness as a restriction (at least under option (i)) was that it preserves analyticity. That is, $\models_{\mathcal{L}}$ is the relation that holds between two sentences if one entails the other solely in virtue of the meaning of components of $\mathcal{L}$. (Alternatively, slightly different pictures may be given along similar lines to options (ii) and (iii) above.)

There are a number of things that should be noted about $\models_{\mathcal{L}}$. Given the gloss just mentioned, we should expect $T_{\mathcal{L}}$ to be closed under $\models_{\mathcal{L}}$. This follows simply from the fact that $\models_{\mathcal{L}}$ is to preserve analyticity. But $\models_{\mathcal{L}}$ will not be a relation only on sentences or formulas of $\mathcal{L}$, nor can it be. For it is essential in giving the definitions of consistency and conservativeness that the relation is defined with D on the left hand side, which is itself not a sentence of $\mathcal{L}$.

Now, it may be thought at this point that there is no need to take $\models_{\mathcal{L}}$ as going beyond some notion of plain logical consequence. Logical consequence is defined not just for some language $\mathcal{L}$, but for any extension of such a language, as required. Moreover, any aspect of inference which is not purely logical but depends on aspects of the meanings of $\mathcal{L}$ seems to be accommodated by adding the appropriate conditionals to $T_{\mathcal{L}}$. For example, in a language which features the words 'bachelor' and 'unmarried', the inference from 'x is a bachelor' to 'x is unmarried' will hold as a matter of meaning (though not as a matter of pure logic). But, this does not require that some special consequence relation is required specifically for this language. Instead it will simply be the case that $T_{\mathcal{L}}$ includes the conditional 'x is a bachelor $\rightarrow$ x is unmarried'. The inference in question then will follow simply as a matter of logic (with $T_{\mathcal{L}}$ in the background), by an application of *modus ponens*.

But things may not be a simple as this. There may be inferences which hold as a matter of the meanings of components of $\mathcal{L}$, but which hold between sentences which themselves are not sentences of $\mathcal{L}$, but of some extension. In such a circumstance, the relevant conditional will not be a part of $T_{\mathcal{L}}$, since it too will not be a sentence of $\mathcal{L}$. Consider again the example of 'bachelor' and 'unmarried' as above. Now suppose that the language is extended to include a new constant or proper name, 'a'. Then an inference from 'a is a bachelor' to 'a is unmarried' should surely hold as a matter of the meanings of 'bachelor' and 'unmarried', as before. But in this case, the conditional 'a is a bachelor$\rightarrow$ a is unmarried' can not appear in $T_{\mathcal{L}}$, since it is not a sentence of $\mathcal{L}$. It seems then that a special consequence relation *is* required after all.

There are a number of ways that this could be resisted, especially for this example, but I think that ultimately, they do not work work.[9] One such way would be to claim that, although the instance 'a is a bachelor $\rightarrow$ a is unmarried' can not be a member of $T_{\mathcal{L}}$, its universal generalisation '$\forall x(x$ is a bachelor $\rightarrow$ x is unmarried)' will be, since it only features language from $\mathcal{L}$. And it is from *this* that the inference can be made

[9] In any case, recall that the aim here is to give these restrictions sufficiently schematically so that various versions can be considered. So, if somebody thinks that there is no need to go beyond logical consequence, that is fine. But if somebody this that it is required to go beyond purely logical consequence in this case or in others, that too should be accommodated. Thus, I do not need to show that one *must* go beyond logical consequence in the definitions, but merely that there are plausible reasons why one may wish to go beyond logical consequence.

purely logically over the background of $T_{\mathcal{L}}$. But still, making this inference will involve instantiating the quantifier appearing in it with a. But what will guarantee that the reference of a—if any—will lie in the domain of the quantifier in $\mathcal{L}$? If there is to be something which justifies this (perhaps the claim that the quantifier of $\mathcal{L}$ is absolutely unrestricted), then it will surely be the case that this inference is then justified by specific reference to features of the meaning of components of $\mathcal{L}$ (in this case, the domain of the quantifier), which brings us back to a need for a consequence relation which is specific to $\mathcal{L}$.

If this particular case is not convincing, there are other situations in which it seems more desirable that we should have a consequence relation which applies to extensions of $\mathcal{L}$ and which goes beyond purely logical consequence. For example, it has been claimed that some inference rule schemas should be *open ended*, in that the schema may be instantiated by formulas from extensions of a language. The inference rules for logical connectives are a clear case of this; the inference from, for example, '$\phi \wedge \psi$' to 'ψ' is to hold, not just for sentences from some particular language, but from any language, and any extension of any language. But there are also cases where such inference rules do not pertain to strictly logical vocabulary. For example, McGee (1997) argues that we should understand arithmetical induction as being open-ended, so that the schema can be instantiated by any formula from any extension of the language of arithmetic.[10] Or, it is plausible to think that the comprehension schema of second-order logic can be instantiated not just by formulas of some particular language, but by formulas of any possible language.

So, to summarise, for a given language $\mathcal{L}$, $\models_{\mathcal{L}}$ will be the relation which holds between sentences when one follows from the other in virtue of the meaning of $\mathcal{L}$. There is no requirement that the relation be one which only applies to sentences of $\mathcal{L}$. Nor is there a requirement that it be restricted to logical consequence.

8.4.3 An example

Now, what kind of candidates might there be for $T_{\mathcal{L}}$ and $\models_{\mathcal{L}}$ for some given language? Consider, for example, the language of arithmetic. As before, here too there will be multiple options, depending on how one sees the meaning of the language to be fixed. Perhaps the simplest option would be to take $T_{\mathcal{L}}$ to be some axiomatised theory such as PA (or its closure under provability), and $\models_{\mathcal{L}}$ to be the standard provability relation.[11] Such a view may arise from identifying meaning with use, together with the claim that the use of language must be such that assertability conditions are effectively computable in the way that provability is. But such a choice is likely to be unduly restrictive. Given the nature of both $T_{\mathcal{L}}$ and $\models_{\mathcal{L}}$, there will be a Gödel sentence G in the language of arithmetic which is not provable from $T_{\mathcal{L}}$, but which could perhaps be claimed to be

[10] McGee does not treat open-ended induction as an open-ended inference rule, as I am suggesting, but instead as a sentence schema. Open-ended schemata rather than inference rules could probably be accommodated into $T_{\mathcal{L}}$ by making some modifications, but, since there will already be some open-ended inferences involved in $\models_{\mathcal{L}}$ in the form of logical inferences, it is simpler to only consider open-endedness in $\models_{\mathcal{L}}$ rather than also in $T_{\mathcal{L}}$.

[11] This would of course require that the arithmetical truths derivable from PA be analytic. But one could instead take $T_{\mathcal{L}}$ to be a collection of sentences suitably conditionalised, on, say, the existence of numbers, in order to avoid this.

true by virtue of the meaning of arithmetical language. Such a conception will also rule out as non-conservative many extensions of the language which might be thought to be in good standing. For example, even a small amount of set theory will be enough to prove the Gödel sentence of $\mathcal{L}$. So some larger, non-recursively-enumerable theory for $T_{\mathcal{L}}$ may be desired (perhaps along with a weaker consequence relation).

One way to do so would be to consider a second-order language of arithmetic, and take the consequence relation to be full second-order consequence. Then the corresponding $T_{\mathcal{L}}$ would consist of all sentences of second-order arithmetic which follow from the second-order axioms. In this case, $T_{\mathcal{L}}$ will be negation complete, and so conservativeness and consistency come to the same thing. Alternatively, one could remain with first-order PA, but with the modification that, in inferences with extensions to the language of arithmetic, the induction scheme may be instantiated with formulas from the extended language. McGee (1997) shows that such a theory—like second-order arithmetic with the full consequence relation—uniquely determines the theory of the natural numbers. Hence (I suspect), it would amount to much the same thing as the second-order approach.

8.4.4 Piecemeal conservativeness and universal conservativeness

For either of conservativeness or consistency, there are two ways in which they may be taken. One way is for the requirement to be relative to some particular language $\mathcal{L}$ (together with $T_{\mathcal{L}}$ and $\vDash_{\mathcal{L}}$). So, the acceptability or not of a definition D will always be relative to some base language. In most cases, it is clear what the relevant language will be which D must be checked against—it is the language which the new vocabulary is being added to, and which D is expressed in.

Another way of taking conservativeness (and consistency) is in a universal form. Here, conservativeness is required, not just relative to some particular language, but relative to *any* language (with corresponding $T_{\mathcal{L}}$ and $\vDash_{\mathcal{L}}$). Again, I will not rely on either one of these being the *correct* option, but hope to make use of these in such a way that my arguments go through given any selection of these options.

I do not wish to claim that any of these options for $T_{\mathcal{L}}$ and $\vDash_{\mathcal{L}}$ is *correct*, nor that either of conservativeness or consistency is more appropriate for a criterion of acceptability. The aim is to provide a way in which some definition or definitions of a new quantifier may be ruled out. For this it will suffice to show that such a definition would be ruled out on any reasonable conception of acceptability along these lines. My aim will simply to be to target the weakest such conception—namely, consistency where $\vDash_{\mathcal{L}}$ is just taken to be provability. Unacceptability according to this measure will then entail unacceptability according to any other measure, since inconsistency entails non-conservativeness, and for any reasonable choice of $\vDash_{\mathcal{L}}$, $D, T \vdash \phi$ will imply $D, T \vDash_{\mathcal{L}} \phi$.

8.5 Rejecting definitions and conservativeness

So, how might conservativeness help to solve the problem? The idea will be to reject one or other of the definitions. So, for example, from the point of view of the relativist, Def_{AU} may be regarded as being non-conservative, and vice-versa. Thus, the solution will not tell in favour of relativism in particular. As I wrote in chapter 5, my aim is not

to provide expansionist abstraction—and, along with it generality relativism—-as an inevitable alternative to the static, absolutist approach; it is instead merely to provide an alternative.

In order to decide between the absolutist definition and the expansionist definitions then some extrinsic argument will be required (which I do not intend to give here). But that this must be resorted to is not a fault of the framework that has been set up (i.e. that of justifying implicit definitions of quantifiers by means of the context principle, and of making use of postulational modality). The framework is not intended as an argument for relativism by itself. The aim is rather to provide a framework within which both positions can be adequately expressed, and with which metaphysical worries about domain expansion can be resolved.

So then, may we reject one of the definitions as being non-conservative? At first, this may seem to be what Dummett (1994) calls 'to wield the big stick'. Dummett considers the case of someone who asks what the cardinality of all the cardinals is[12]and who is told that they must not do that, since it will lead to contradiction. In this case, somebody asks about 'all possibly introduced objects' or something to that effect, and is told that they can not do so since that would be to make a non-conservative definition.

The difference, however, is that in this case, the rejection of one of the definitions is justified by considerations concerning definitions in general. There is also room for an explanation of why either of the definitions may be problematic, in terms of indefinite extensibility. But how might one go about rejecting one or other of the two conflicting definitions? In either case, the aim will be to show that the other's principle is non-conservative when $\vDash_T$ is taken to be provability, which will then entail that the principle is non-conservative in any other reasonable sense. I shall go through both options (either rejecting Def_{AU} as non-conservative or rejecting BLV as non-conservative).

Consider a relativist who wants to claim that Def_{AU} amounts to a non-conservative addition to their language. In particular, they want to show that for some consistent T and ϕ in the language L of T, $T \cup \{\mathrm{Def}_{AU}\} \vdash \phi$, whilst $T \vdash \neg\phi$. That is, the aim is to show that $T \cup \{\mathrm{Def}_{AU}\}$ is inconsistent. In this case, the language L will be that of the modality together with the set operator, and the theory governing the language, T, will be EL+BLV↓. Then $T \vdash (\mathrm{GR})$. But, as noted above, Def_{AU} proves the negation of this.

The same can be said the other way around, with the absolutist rejecting BLV since it is non-conservative over Def_{AU}.

But something also needs to be said concerning the universalised forms of conservativeness, according to which a definition must be conservative over *any* theory which is in good standing. This is not a worry in the course of showing that the opponent's definition is non-conservative, since, by showing that it is non-conservative over one's preferred theory, one has shown that it is not universally conservative in the appropriate sense. But what *is* a worry is that, on the requirement of universal conservativeness, both definitions must be ruled out as non-conservative, since they are both non-conservative

[12]It seems to me that this is not the best example, since, as the consistency of HP—on the static view—shows, it is consistent that there be a cardinal of all cardinals. Since HP is consistent, and (unless modified somehow) entails that there is a cardinal for any concept (including self-identity, and being a cardinal), it is consistent for there to be a cardinal of all cardinals. Moreover, even in the context of a set theory such as ZFC, a contradiction only follows from the existence of a cardinal of all cardinals together with some additional assumptions—such as the claim that every cardinal is the cardinal of some *set*, that cardinals are themselves sets, and so on. A better example might be of somebody asking what the order type of the ordinals is.

over the theory resulting from the opposing definition.

In this case, however, it seems that such a worry can be avoided. For it must only be conservativeness over theories *in good standing* that must matter, and the opposing definition will *not* be seen as a theory in good standing. So, consider for example the relativist again. After showing that Def_{AU} is non-conservative, they must reject the claim that BLV is non-conservative, since it is non-conservative over Def_{AU}. But this is easily done. For, as they have already claimed, Def_{AU} is non-conservative, and so *not* a theory in good standing by which to judge the conservativeness of BLV. There is obviously a gap left: since the situation is symmetric, reasons must be given for going the relativist way, rather than the absolutist way. But these may be reasons extrinsic to issues concerning conservativeness and so on.[13]

8.6 Reflection

8.6.1 Reflection for the absolutist

Now, despite Def_{AU} being unacceptable to the relativist, it seems that it provides a way of increasing the amount of mathematics that can be derived from this approach. In particular, it seems to solve the problem raised earlier of expanding to an infinite domain without assuming that infinitely many separate expansions must take place. It would be desirable to achieve the same effect from the point of view of the relativist, by making use of weaker implicit definitions of quantifiers than Def_{AU}. In this section I shall describe the 'reflection' effect that Def_{AU} has, and describe a way that the relativist can achieve the benefits in a way which is acceptable from that point of view.

Recall that there is a particular problem concerning expanding a domain from one which is finite to one which is infinite. So far, all that has been motivated by way of domain expansions arising from abstraction are those from finite domains to larger, but still finite, domains.[14] The same goes for domain expansion according to Fine's view of procedural postulationism. Each procedural postulate (with the exception of those arising from his '$*$' operator, which I shall discuss later) will only result in a finite expansion if starting from a finite domain.

So, some additional means is required in order to transform such finite expansions into an infinite expansion. For example, Fine (2005, pp.92,94) suggests that the process of iterating a procedural postulate (which is indicated by the '*' operator) may continue into the transfinite.[15]

[13] For example, the ease with which set theorists talk about proper classes as objects may suggest that the kind of collapse of concepts to objects permitted by BLV may be desirable. But more argument is clearly needed.

[14] in particular, in the case of HP, a domain of cardinality n may be expanded to one of cardinality $n + 1$, and in the case of BLV, a domain on cardinality n may be expanded to one of cardinality 2^n.

[15] Fine glosses the * operator as follows:

> We may read β^* as: iterate β; and β^* is executed by executing β, then executing β again, and so on for any finite number of times. (p.92)

with an exception that it may need to be interpreted as going into the transfinite only for the case of set theory. But it seems to me that the operation must be transfinite even for the case of arithmetic. For, in that case, it is presumably necessary to expand one's domain to one containing infinitely many objects (namely, the natural numbers), whereas the * operator as stated here appears only to accommodate arbitrarily large finite domains.

But, given the epistemological aims of the enterprise, this will not be acceptable. For, if gaining an understanding of an infinite domain requires the transfinite iteration of an action—namely, coming to an understanding of increasingly larger domains through domain expansion of some kind—how are mere humans to achieve it? Instead some other means is required to gain such an understanding, and it must be by means which it is plausible for humans to achieve.

What is required instead is some way to achieve the same *effect*—i.e., an understanding of quantification over an infinite domain—but in a finite time. That is, what is required is a specification of truth conditions associated with a quantifier such that it, in effect, ranges over an infinite domain. Def_{AU} seems to provide such truth conditions (and, as I will argue shortly, so do some other principles which are acceptable to the relativist).

Def_{AU} motivates the following rule:

$$\frac{\phi^{\Diamond}}{\phi^{AU}}$$

where $\phi^{\Diamond}$ is as before, and where ϕ^{AU} results from superscripting each quantifier with 'AU'. Furthermore, ϕ^{AU} will then entail that $\Diamond\phi$, since the truth conditions associated with $\exists^{AU}/\forall^{AU}$ are assumed to describe a possible way of interpreting the quantifiers (since, recall, this is from the absolutist perspective).[16] Hence, we have the reflection principle.

As already noted, the reflection principle captures the thought that it is possible to 'close off' an operation of continually expanding one's domain. So, for example, if it is the case that, (necessarily) for any number it is possible to expand the domain to include its successor, the reflection principle entails that it is possible to expand the domain so that every number *has* a successor (in that same domain). Similarly, if it is the case that for any set, it is possible to expand the domain to include its powerset, then the reflection principle entails that it is possible to *close off* the iteration, so that every set *has* a powerset in that same domain. This is not to say that the reflection principle states that one can achieve the supertask of iterating infinitely. Rather, the possibility is witnessed by the interpretation of the quantifiers according to Def_{AU}, which, in order to be understood, does not even require an understanding of what it would be to continue a process into the transfinite, let alone the ability to do so.

Another way of illustrating the effect of the reflection principle is in terms of possible worlds models. It motivates the claim that, given some modal sentence ψ, it is possible to reinterpret one's quantifiers so that they, in effect, range over the domain of a possible worlds model of ψ. Why? Suppose that $\phi^{\Diamond}$ holds in some world @ in a possible worlds model $\mathcal{M}$ and let D be the union of the domains of worlds accessible from @. Then, a simple induction on quantifier complexity shows that ϕ holds with its quantifiers interpreted as ranging over D, and any non-logical vocabulary interpreted as it is at @.[17]

[16] Some additional care will be needed when ϕ includes non-logical vocabulary, such as abstraction operators, so that these are not evaluated within the scope of the enclosing possibility operator.

[17] More precisely: Let $\mathcal{L}$ be a language (i.e. a collection of non-logical constants) and ϕ a sentence of that language. Then, if $\mathcal{M} = \langle W, I, \delta \rangle$ is a possible worlds model, with $w \in W$, then $\mathcal{M}, w \vDash \phi^{\Diamond}$ if and only if $\mathcal{M}' \vDash \phi$, where $\mathcal{M}$ is the (not possible worlds) model $\mathcal{M}' = \langle D', I' \rangle$, where $D' = \cup_{w' \text{ s.t. } wRw'} \delta(w)$ and for each non-logical constant ξ of $\mathcal{L}$, $I'(\xi) = I(w, \xi)$.

Indeed, from the absolutist point of view, it makes sense that this is the case. For from the absolutist point of view, a domain expansion is equivalent to the lifting of restrictions on the absolutely unrestricted domain. So, a domain expansion is possible just in case *there are* (unrestrictedly) some objects which can serve as members of that domain—that is, the domain of all 'possible' objects for domain expansions just is the domain of *all* objects.

8.6.2 Reflection for the relativist?

Now, with the desirable consequences that the reflection principle brings (namely, a way of 'completing' various processes of expansion), is there a way of getting these consequences without the full strength of Def_{AU}? That is, are there weaker definitions which (a) allow one to motivate the reflection principle and (b) are consistent/conservative over relativist commitments? I will argue that there are. In particular, I will not replace Def_{AU} with a *single* definition which motivates reflection in a way acceptable to the relativist. Instead, I shall provide a definition schema, which will provide a definition of a quantifier for each instance of the reflection principle. Together these will justify the reflection schema as a whole.

Before giving these definitions, it will be useful to set up some notation. The aim will be to define, relative to some sentence ϕ, a quantifier $\exists^{\phi}$ which will witness the relevant instance of reflection, i.e., the instance featuring ϕ. Where ψ is any sentence or formula, I will denote by ψ^{ϕ} the sentence or formula which results by replacing each instance of $\forall$ and $\exists$ in ψ by $\forall^{\phi}$ and $\exists^{\phi}$ respectively.[18] As before, I shall use $\psi^{\Diamond}$ to denote the replacement of each quantifier $\forall x$ or $\exists x$ in ψ by $\Box\forall x\downarrow$ or $\Diamond\exists x\downarrow$ respectively.

The aim is as follows. We want to give truth conditions associated with $\exists^{\phi}/\forall^{\phi}$ in such a way that, for example, from $\Box\forall x\Diamond\exists y Sxy$ we can derive $\forall^{\phi}x\exists^{\phi}ySxy$. In this case, since we wish to use this to derive the instance of the reflection principle featuring ϕ, the relevant ϕ will be $\forall x\exists ySxy$. Alternatively, the aim can be be put in terms of possible worlds models, as was discussed in the last section. Given a modal sentence (or formula) ϕ, we want to give truth conditions for a quantifier which, in effect, quantifies over the domain of a possible worlds model of ϕ.

Note how this aim differs from the consequence of Def_{AU}. That consequence is that it is possible to quantify over what is effectively the domain of a possible worlds model of *every* true modal sentence. The present aim is that, given any modal sentence, it is possible to quantify over what is in effect the domain of a possible worlds model of *that* sentence.

How is this aim to be achieved? The strategy will be the following. Instead of giving truth conditions for the new quantifiers *directly*, as was the case for Def_{AU}, they will be given indirectly. Truth conditions will be given directly for some sentences, in addition to a stipulation that the newly introduced quantifiers are to obey the appropriate free-logical inference rules (so, roughly, elimination rules will give necessary conditions, and

Proof. This can be verified by a simple induction on complexity of ϕ. The only non-trivial case is where ϕ is $\forall x\psi$ for some ψ. Then $\phi^{\Diamond}$ is $\Box\forall x\downarrow\psi$. Now, $w \vDash \Box\forall x\downarrow\psi$ iff for all w' such that wRw', for all $a \in \delta(w')$, $w \vDash \psi(a)$. Which (by the inductive hypothesis) is equivalent to $\mathcal{M}' \vDash \psi(a)$. etc. □

[18] It is unfortunate that this notation coincides with one often used for quantification *restricted* to ϕ (so that, for example $(\forall xFx)^{\phi}$ denotes the sentence $\forall x(\phi(x) \to Fx)$). For the present purposes then, it should be noted that this is *not* what ψ^{ϕ} signifies.

introduction rules sufficient conditions).

The implicit definition for the new quantifier $\forall^{\phi}$ will then be:

$$(*) \qquad \phi^{\phi} \leftrightarrow \phi^{\Diamond}$$

together with a stipulation of the inference rules. It is then trivial to infer the relevant instance of reflection from this. For example, suppose that ϕ is $\forall x \exists y S x y$. Then, ϕ^{ϕ}, which is the sentence for which we are directly giving the truth conditions, will be $\forall^{\phi} x \exists^{\phi} y S x y$. This will then be determined to be true just in case $\Box \forall x \downarrow \Diamond \exists y \downarrow S x y$, which is stated entirely in terms not including $\exists^{\phi}$. In addition, some necessary or sufficient conditions will be given for further sentences including $\exists^{\phi}$.

The definition $(*)$ above is unlike both Def_{AU} and the definition previously given for expansions resulting from abstraction principles.[19] These are both, to some extent, *explicit* definitions, directly giving the truth conditions for statements of the form $\exists^{+} x \phi$ (where $\exists^{+}$ is the quantifier being defined). $(*)$, by contrast, only directly gives the truth conditions of one sentence, which, although containing the new quantifier, might not simply be of the form $\exists^{+} x \phi$. For this reason, it is worth discussing whether it is legitimate to do so.

Recall that one requirement on an interpretation being legitimately a *quantifier* was that it should have the correct inferential behaviour. In the case of the previous definitions, it has been possible to prove that they satisfy the relevant rules. But is it legitimate—as is the case here—to simply *stipulate* that $\exists^{\phi}$ behaves like a quantifier? I do not see why not. The aim of a definition will simply be to give truth conditions for a sufficient range of sentences in a way which is conservative. There need not be any restrictions on *how* such truth conditions are given. In any case, giving the meaning of a new symbol in terms of inference rules is no different from the claim that the meanings of the logical connectives can be given by their inference rules (or the weaker claim that inference rules *partly* give the meanings of logical connectives). Or, for that matter, it is not very different from how abstraction principles must be seen as definitions of new singular terms (or rather, of new abstraction operator symbols). In this case, it must surely be part of the introduction of the abstraction operator that the terms it forms are indeed singular terms (that is, that they satisfy the appropriate inferential behaviour for singular terms). For being *grammatically* like a singular term (that is, in this case, flanking an identity symbol) is not enough to be a singular term in general (cf. Dummett, 1973; Hale, 2001b,c).

Another objection to the proposed stipulation might be that it fails to be sufficiently general. That is, that it fails to give the truth conditions for a sufficiently large rage of sentences featuring the new quantifier. The worry is then that this means that the definitions fail to confer the new vocabulary with a meaning. The objection is essentially just a variant of the Julius Caesar problem. It should be noted that this problem is not unique to $(*)$. For example, even though the definition of $\exists^{\S}$ (see fn. 19) is explicit, the truth conditions it gives depend on the truth conditions of formulas involving abstract terms, and, famously, the truth conditions of all such sentences in not settled by abstraction principles. The situation is similar for Def_{AU}.

[19] Recall, these were:

$$\exists^{\S} x \phi \leftrightarrow \exists F \phi(\S F)$$

where § is the abstraction operator in question, and $\phi(\S F)$ is to be evaluated via the abstraction principle in question.

There are two questions which then need to be asked. Firstly, to what extent is this lacuna in the truth conditions fixable—that is, can further principles be added as part of the implicit definitions so as to confer a complete range of sentences featuring the new symbols with truth conditions?[20] Secondly, even if this is not achievable, to what extent does this matter? Perhaps it could be claimed that, although the definitions do not succeed in completely conferring a definite meaning for the new quantifiers, they do succeed in conferring meaning partially. To fully answer these questions now would be a large undertaking. I believe, however, that adequate answers can be given to these questions. In particular, if truth conditions can not be given universally for the new symbols, I do not think that this affects the proposal to the extent that it is worthless. For sure, a gap in truth conditions would leave indisputable worries concerning the determinacy of reference of the proposed quantifiers.[21] But that would not affect the main conclusion, which is that it is possible to reinterpret quantifiers so that they quantify over a domain on which certain parts of mathematics are true.

8.7 Conclusion

I have here tackled two separate, but related problems for the way in which I have justified the introduction of new quantifiers. One is how to justify the introduction of quantifiers over infinite domains, without assuming that it is possible in principle to repeatedly expand ones domain infinitely many times. The second concerns the attitude that should be taken towards the apparent introduction of an absolutist quantifier in a similar way to how other quantifiers have been introduced.

The second problem can be avoided by appealing to what must be a general constraint on introducing quantifiers—namely that of conservativeness. Although this restraint is not sufficient to *rule out* the absolutist quantifier, it does give an account of why there is no compulsion to accept *both* the absolutist quantifiers and the relativist quantifiers. This still leaves a gap concerning how one is to justify that it should be the absolutist quantifier, rather than the relativist quantifiers based on BLV, that must be jettisoned. I do not intend to give a knock-down argument as to why this should be the case. However, the capability of the collection of relativist quantifiers to recover large amounts of set theory, to avoid the bad company problem, and—or so could be argued—to account for talk of proper classes and indefinite extensibility, is sufficient to make the relativist quantifiers a worthwhile topic of enquiry.

The adopter of the absolutist quantifiers has some method of justifying a reflection principle, which, along with some principles of expansion, motivates the expansion to infinite domains (although, since the quantifiers based on BLV must then be considered illegitimate, principles of expansion may no longer be available). I hope to have shown that similar justification is available to the relativist as well, by adopting a schema for introducing multiple quantifiers which together have the same effect.

[20] If so, this raises an additional question as to whether there is a problem with such a piecemeal approach.

[21] Although, even if truth conditions *could* be given for all sentences containing the new vocabulary (which would have to be given in a thoroughly non-recursive manner, due to Gödel's theorems), there might still be doubts about the determinacy of the domain of quantification for Skolomite reasons. For example, it could be claimed that if truth conditions were sufficient to determine that the domain of quantification is infinite, then no amount of further statement of truth conditions could determine the precise cardinality of the domain.

Chapter 9

Conclusion

9.1 Conclusion

I hope that I have done enough in this thesis to provide at least the beginnings of an alternative to the standard static way of looking at abstraction. That abstraction may best be conceived of in a broadly expansionist way has been suggested in a few places, such as Fine (2002). But there has not been much in the way of a detailed working out of how such a conception of abstraction may proceed. There have been programmes in the philosophy of mathematics which are clearly influenced by such a view (e.g. Fine, 2005; Linnebo, ms), but these depart from abstractionism in more ways than just allowing for domain expansion.

My aim was first to look at the static approach to abstraction for two main purposes: One was to determine what it is about the standard approach which is distinctively committed to a static domain. This, I argued in chapter 2, was the characteristic rule of negative free logic, ($E!$-I). Secondly, the aim was to provide some motivation for considering alternative approaches. The main such motivation was the bad company problem. I argued in chapters 3 and 4 that this problem includes an epistemological dimension which is not overcome by any of the restrictions commonly considered.

In the second part of my thesis, my aim was to make a start on providing the details of an expansionist account of abstraction. There were two main aims. The first, philosophical aim was to provide support for the conceptual underpinnings of expansionist abstraction. These are the claim that it is not possible to quantify over absolutely everything and, more importantly, that there is nothing mysterious in the thought that we can indefinitely expand our domain of quantification by means of reinterpretting our quantifiers. The second, more technical aim was to provide a formal theory of expansionist abstraction. As a crucial part of this, I showed that the expansionist approach does not suffer from the bad company problem (at least, it allows BLV to be used without threat of inconsistency), and that in this framework it is possible to develop a significant amount of set theory.

As I wrote in chapter 5, my aim was not to argue decisively that static abstraction must be rejected in favour of expansionist abstraction. But I hope that I have done enough to show that an expansionist approach is a viable alternative.

9.2 Outstanding issues

A number of issues have been left open concerning expansionist abstraction which I feel warrant further investigation, but which would either have constituted too much of a digression, or to which I could not do justice in the space available.

9.2.1 Other abstraction principles

My emphasis in chapter 7 has been BLV, since it is the abstraction principle for which there might be most concern about inconsistency, and since it is the abstraction principle which holds the most promise for a powerful foundation of mathematics. But it would also be desirable to examine the consequences of expansionist abstraction for different abstraction principles.

One such example would be HP. Since I gave a modal formulation of HP in section 7.2.1, it would be a simple task to make use of this modalised version of HP in place of the modalised version of BLV. I suspect that the result would be a theory that proves, for example, that for every number, it is possible to introduce a successor to that number. An application of the reflection principle would then allow one to derive that it is possible to expand the domain so that for every number there is a successor.

There is also the question of how abstraction principles in general might be modalised. One approach would be to provide a general method of constructing a modalised abstraction principle out of non-modal abstraction. An alternative would instead be to provide a general notion of a transworld equivalence relation, which could then serve as an abstraction relation.

Finally, there are questions to be asked about systems of expansionist abstraction which feature more than one abstraction principle at once. How would, for example, the system which consists of suitably modalised versions of HP and NP behave? My suspicion is that, for finite domains, HP will result in domain expansions, as it would by itself, and in infinite domains, NP would then result in domain expansions, in a similar way to how BLV does.

9.2.2 Bad company

As attested to in chapter 8, expansionist abstraction—or, at least, the framework which surrounds it—faces its own version of the bad company problem, in that there may be definitions of quantifiers which conflict with one another. Although I have sketched an outline of how this might be avoided, there is more that could be said.

I suggested that either consistency or a form of strong conservativeness could solve the problem. But there remains the question as to which of these is preferable, and what precise form it should take. Moreover, a version of the epistemological bad company problem arises, which must be dealt with.

Although this problem must of course be faced by the expansionist, there are reasons to think that the expansionist conception may fare better than the static conception. For one thing, the bad company problem only arises when we consider implicit definitions of quantifiers *in general*. If we restrict attention to those based on abstraction principles (as in chapter 6), there is no problem. Moreover, some of the suggestions that I made about suitable restrictions are plausibly decidable in the right kind of way. Consider,

for example, consistency relative to a single theory $T_{\mathcal{L}}$. Assuming that $T_{\mathcal{L}}$ and the consequence relation $\vDash_{\mathcal{L}}$ are both recursively enumerable, it can easily be checked for a purported definition D whether $D, T_{\mathcal{L}} \vDash_{\mathcal{L}} \bot$.

To motivate such a requirement would, however, constitute too much of a digression. Moreover, the resulting view would have to be fairly radical; it would permit a certain amount of relativism in how we can extend a language, for example.

9.3 Directions for development

As well as issues to be resolved concerning expansionist abstraction directly, there are a few directions in which I believe some of the aspects of this thesis can be developed.

9.3.1 Other forms of domain expansion

There might be questions, not about other abstraction principles, but about different forms of domain expansion altogether.

The kind of domain expansion which was discussed in chapter 7 was a kind of expansion of domains of sets 'upwards'. That is, BLV$\downarrow$ was seen as providing a method of expanding a domain so that what were considered proper classes under one interpretation of the quantifier correspond to sets under an expanded interpretation. The domain expansion adds sets at the 'top' of the set-theoretic universe.

But we might want also to consider domain expansion 'outwards', so that the domain expands to include, say, more subsets of the natural numbers. The reason is that the technique of *forcing* in set theory suggests just that. Forcing is a method used to prove independence results in set theory, and proceeds by showing how a model of set theory $\mathcal{M}$ may be expanded to a model $\mathcal{M}[G]$ which verifies some particular statement (e.g. the continuum hypothesis). This model may be such that, from the point of view of $\mathcal{M}$, it contains more subsets of the natural numbers.

Of course, forcing in this sense only concerns set sized models of set theory (moreover, it can only concern *countable* models without running into additional complications). But some writers (e.g. Hamkins, 2011) have claimed that a consequence is that there is no one true universe of sets, but a *multiverse* of different universes. Forcing extensions on this view are genuine extension of the (or *a*) universe. It might then be asked whether the framework for considering domain expansion by expansion may apply to domain expansion by forcing.

9.3.2 Quantifier variance

I mentioned briefly in chapter 6 the relationship between generality relativism and the metaontological view known as quantifier variance. This is the view that many ontological disputes (for example, that between mereological nihilists—who claim that there are no composite objects—and their opponents) can be seen as purely verbal disputes about what meaning the quantifiers have. But there is more to be said about this. In particular, it seems that the attitude to quantification which I suggested in that chapter can be utilised to defend quantifier variance against a number of issues.

In particular, it may be possible to levy the same objections to quantifier variance as have been levied against generality relativism. I believe that similar arguments could be given to defend the positions against these objections. In particular, it may be possible to advocate a modal formulation of quantifier variance, and to explain this modal formulation in much the same way as I did for generality relativism.

To do this would involve extending the account I give to extra-mathematical settings, and this brings with it a whole host of challenges. How can we, for example, give definitions of quantifiers that purport to range over a kind of physical object? What is the correct account of conservativeness for such definitions, and over what kinds of theories should conservativeness be considered? These challenges could—I believe—be overcome.

Appendix A

Modal logic with backtracking operators

In this appendix I present a proof system for propositional modal logic with the backtracking operator ↓. This is sound with respect to a possible worlds semantics which is essentially that of Hodes (1984b). I then present an extension of the system to one of second-order logic with abstraction operators and the ($\Diamond E!$-I) rule. This is EL (for *expansion logic*), which will form the background of abstraction with domain expansion.

A.1 Propositional logic

A.1.1 Language

Let $\mathcal{L}_{\text{prop}}$ be a typical language for propositional modal logic; it consists of countably many propositional variables $p, q, r, \ldots$, connectives $\wedge$ and $\neg$ and a necessity operator $\Box$. In addition, it shall have a backtracking sentential operator ↓, which will have the intended effect of exempting what follows it from the scope of the innermost modal operator from which it is not already exempt (so, for example, p, $\Box{\downarrow}p$ and $\Box\Box{\downarrow}{\downarrow}p$ should all be counted as equivalent). Other sentential connectives $\vee$, $\rightarrow$ and a possibility operator $\Diamond$ can be defined in the usual way.

A.1.2 Model theory

The model theory is essentially that of Hodes (1984b), with the main differences being: (a) Hodes puts forward a logic which is an extension of propositional S5, so that the accessibility relation is an equivalence, whereas the only restriction on the equivalence relation here is that it is serial. (b) Hodes only defines satisfaction for a certain class of sentences, whereas this model theory places no such restriction.

A model is a triple $\mathcal{M} = \langle W, R, a \rangle$, where W is a set (of possible worlds), $R \subseteq W \times W$ is the accessibility relation, and a is an assignment function which assigns to each propositional variable p at a world $w \in W$ a truth value $a(w, p) \in \{T, F\}$.

Only one restraint will be placed on the accessibility relation for now, and that is that it is *serial*. So, for any $w \in W$ there is a $w' \in W$ such that wRw'.

Then, a satisfaction relation is defined, not for each world, but for each finite sequence of worlds of the appropriate type (which I shall call *world sequences*). Define:

$$\text{(A.1)} \qquad \mathrm{WS}_{\mathcal{M}} = \{\langle w_1, \ldots, w_k\rangle : k \geq 1, \forall i \leq k, w_i \in W \text{ and } \forall i < k, w_i R w_{i+1}\}$$

As a result of seriality, for every world sequence there will be a world sequence extending it (and so there are world sequences of arbitrary length).

Some terminology for members of $\mathrm{WS}_{\mathcal{M}}$ will be useful. I shall write $\vec{w}$ for an arbitrary member of $\mathrm{WS}_{\mathcal{M}}$. Where $\vec{w} = \langle w_1, \ldots, w_k\rangle$, then:

$$\begin{aligned}
\vec{w}, w' &= \langle w_1, \ldots, w_k, w'\rangle \\
\vec{w}- &= \begin{cases} \langle w_1, \ldots, w_{k-1}\rangle, & k > 1 \\ \langle w_1\rangle, & k = 1 \end{cases} \\
t(\vec{w}) &= w_k \\
l(\vec{w}) &= k
\end{aligned}$$

Now, satisfaction of sentences at world sequences of $\mathrm{WS}_{\mathcal{M}}$ is defined as follows:

For propositional variables:

$$\mathcal{M}, \vec{w} \vDash p \quad \text{iff} \quad a(t(\vec{w}), p) = T$$

For propositional connectives:

$$\begin{aligned}
\mathcal{M}, \vec{w} \vDash \neg\phi \quad &\text{iff} \quad \mathcal{M}, \vec{w} \nvDash \phi \\
\mathcal{M}, \vec{w} \vDash \phi \wedge \psi \quad &\text{iff} \quad \mathcal{M}, \vec{w} \vDash \phi \text{ and } \mathcal{M}, \vec{w} \vDash \psi
\end{aligned}$$

For modal operators (including $\downarrow$):

$$\begin{aligned}
\mathcal{M}, \vec{w} \vDash \Box\phi \quad &\text{iff} \quad \text{for all } w' \text{ s.t. } t(\vec{w}) R w',\ \mathcal{M}, \vec{w}, w' \vDash \phi \\
\mathcal{M}, \vec{w} \vDash \downarrow\phi \quad &\text{iff} \quad \mathcal{M}, \vec{w}- \vDash \phi
\end{aligned}$$

A consequence relation can then be defined:

Definition A.1. Where Γ is a set of sentences in $\mathcal{L}$, and ϕ a sentence of $\mathcal{L}$, then $\Gamma \vDash \phi$ iff:

For all $\mathcal{M}$ and $w \in W_{\mathcal{M}}$, if $\mathcal{M}, \langle w\rangle \vDash \psi$ for each $\psi \in \Gamma$, then $\mathcal{M}, \langle w\rangle \vDash \phi$.

A.1.3 Proof theory

I this section I shall describe a natural deduction system for the logic. This will consist of an introduction and elimination rule for each connective and operator, and a definition of the notion of a deduction.

An important feature will be that deductions and inference rules will operate on *labelled* sentences. A labelled sentence is a pair $\phi;\vec{s}$, where ϕ is a sentence of $\mathcal{L}$, and $\vec{s}$ is a (possibly empty) finite sequence of natural numbers. Where the label is empty, I

shall write $\phi\,;-$. The same terminology as for world sequences will be used, with the exception that for labels, where $\vec{s} = \langle n_1, \ldots, n_k \rangle$ I shall write:

$$\vec{s}- = \begin{cases} \langle n_1, \ldots, n_{k-1} \rangle, & k > 0 \\ \langle\rangle, & k = 0 \end{cases}$$

The rules are as follows:

$$(\wedge\text{-I})\ \frac{\phi\,;\vec{s} \qquad \psi\,;\vec{s}}{\phi \wedge \psi\,;\vec{s}} \qquad (\wedge\text{-E}_1)\ \frac{\phi \wedge \psi\,;\vec{s}}{\phi\,;\vec{s}} \qquad (\wedge\text{-E}_2)\ \frac{\phi \wedge \psi\,;\vec{s}}{\psi\,;\vec{s}}$$

$$(\neg\text{-I})\ \frac{\begin{array}{c}[\phi\,;\vec{s}] \\ \vdots \\ q \wedge \neg q\,;\vec{t}\end{array}}{\neg\phi\,;\vec{s}} \qquad (\neg\text{-E})\ \frac{\begin{array}{c}[\neg\phi\,;\vec{s}] \\ \vdots \\ q \wedge \neg q\,;\vec{t}\end{array}}{\phi\,;\vec{s}}$$

$$(\Box\text{-I})\ \frac{\phi\,;\vec{s}}{\Box\phi\,;\vec{s}-} \qquad (\Box\text{-E})\ \frac{\Box\phi\,;\vec{s}}{\phi\,;\vec{s}, n}$$

(with a restriction on the (□-I) rule that $\phi\,;\vec{s}$ may only depend on assumptions with labels $\vec{t}$ such that $\vec{s}$ properly extends $\vec{t}$.)

$$(\downarrow\text{-I})\ \frac{\phi\,;\vec{s}}{\downarrow\phi\,;\vec{s}, n} \qquad (\downarrow\text{-E})\ \frac{\downarrow\phi\,;\vec{s}}{\phi\,;\vec{s}-}$$

With these rules, rules for the defined connectives and operators can be deduced. In particular, rules for ◊ will be:

$$(\Diamond\text{-E})\ \frac{\Diamond\phi\,;\vec{s} \qquad \begin{array}{c}[\phi\,;\vec{s}, n] \\ \vdots \\ \psi\,;\vec{t}\end{array}}{\psi\,;\vec{t}} \qquad (\Diamond\text{-I})\ \frac{\phi\,;\vec{s}, n}{\Diamond\phi\,;\vec{s}}$$

(with no restriction on the introduction rule, and a restriction on the elimination rule that the label $\vec{s}, n$ is not already in use).

Since deductions in the system will involve discharging assumptions and restrictions on which assumptions are allowed, the notion of a *deduction* rule corresponding to each inference rule—which will specify how assumptions are to be discharged—is needed (cf. Prawitz, 1965). Deduction rules are n-tuples of the form $\langle\langle \Gamma_1, \theta_1\,;\vec{s}_1\rangle, \ldots, \langle \Gamma_k, \theta_k\,;\vec{s}_k\rangle\rangle$, which say that when $\theta_1\,;\vec{s}_1, \ldots \theta_{k-1}\,;\vec{s}_{k-1}$ have been derived using undischarged assumptions $\Gamma_1 \ldots \Gamma_{k-1}$ respectively, then $\phi_k\,;\vec{s}_k$ can be derived with undischarged assumptions Γ_k. For example, the deduction rule corresponding to (¬-I) will be $\langle\langle \Gamma, p \wedge \neg p\,;\vec{t}\rangle, \langle \Gamma\backslash\{\phi\}, \neg\phi\,;\vec{s}\rangle\rangle$.

Two notions of proof-theoretic consequence can then be defined—one for labelled sentences, and one for unlabelled sentences (which is the one of primary importance):

Definition A.2. Let Δ be a set of labelled sentences of $\mathcal{L}_{\text{prop}}$, and $\phi\,;\vec{s}$ a labelled sentence of $\mathcal{L}$. Then $\Delta \vdash \phi\,;\vec{s}$ iff there is some sequence of ordered pairs:

$$\langle\langle\Delta_1, \theta_1\,;\vec{t}_1\rangle, \ldots, \langle\Delta_n, \theta\,;\vec{t}_n\rangle\rangle$$

such that:

- $\Delta_n = \Delta$ and $\theta_n\,;\vec{t}_n$ is $\phi\,;\vec{s}$
- For $i \leq n$, either:
  - $\theta_i\,;\vec{t}_i \in \Delta_i$, or
  - There are $j, k < i$ such that $\langle\langle\Delta_j, \theta_j\,;\vec{t}_j\rangle, \langle\Delta_k, \theta_k\,;\vec{t}_k\rangle, \langle\Delta_i, \theta_i\,;\vec{t}_i\rangle\rangle$ is an instance of one of the deduction rules corresponding to the inference rules.

Definition A.3. Where Γ is a set of (unlabelled) sentences of $\mathcal{L}$, and ϕ a sentence of $\mathcal{L}$, then $\Gamma \vdash \phi$ iff $\Gamma^* \vdash \phi\,;-$, where Γ^* is the set of labelled sentences resulting from replacing each $\psi \in \Gamma$ by $\psi\,;-$ (i.e. ψ together with an empty label).

A.1.4 Soundness

In this section I shall prove the soundness of $\vdash$ with respect to $\models$. That is, that for any set of sentences Γ and sentence ϕ, if $\Gamma \vdash \phi$ then $\Gamma \models \phi$.

This will go by way of defining a model-theoretic consequence relation for labelled sentences, analogous to the corresponding proof-theoretic consequence relation. Before defining this, I shall need to define a certain type of homomorphism from labels to (possibly empty) world sequences.

Definition A.4. A function $f : \mathbb{N}^{<\omega} \to \text{WS}_{\mathcal{M}} \cup \{\langle\rangle\}$ is a *homomorphism* iff, for all $\vec{s}, \vec{t} \in \mathbb{N}^{\omega}$:

- $\vec{s}$ properly extends $\vec{t} \Leftrightarrow f(\vec{s})$ properly extends $f(\vec{t})$
- $l(f(\vec{s})) = l(\vec{s})$.

A few important consequences of this definition are:

1. For any model, there will be at least one such homomorphism.
2. For any $f, \vec{s}$ and n, $f(\vec{s}, n) = f(\vec{s}), w'$ for some w' such that $t(f(\vec{s}))Rw'$.
3. If $l(\vec{s}) > 0$ then $(w, f(\vec{s}))- = (w, f(\vec{s}-))$.

Proof. 1. Since models are serial, there is an infinite sequence of worlds $\langle w_1, w_2, \ldots\rangle$ such that $w_i R w_{i+1}$ for all $i \in \mathbb{N}$. Now, simply define $f(\vec{s}) = \langle w_1, \ldots, w_{l(\vec{s})}\rangle$. It is simple to check that this satisfies the required properties.

2. $l(f(\vec{s}, n)) = l(f(\vec{s})) + 1$ and $f(\vec{s}, n)$ extends $f(\vec{s})$ since $\vec{s}, n$ extends $\vec{s}$. So, $f(\vec{s}, n) = f(\vec{s}), w$ for some w. The accessibility requirement follows since $f(\vec{s}, n) \in \text{WS}_{\mathcal{M}}$.

3. Suppose $l(\vec{s}) > 0$, so $\vec{s} = \langle n_1, \ldots, n_k\rangle$, $k \geq 1$. Since $l(f(\vec{s})) = l(\vec{s})$, $f(\vec{s}) = \langle w_1, \ldots, w_k\rangle$. So $w, f(\vec{s}) = \langle w, w_1, \ldots, w_k\rangle$. So $(w, f(\vec{s}))- = \langle w, w_1, \ldots, w_{k-1}\rangle = (w, f(\vec{s}-))$ □

We now define a model-theoretic consequence relation for labelled sentences.

Definition A.5. Let Δ be a set of labelled sentences, and $\phi\,;\vec{s}$ a labelled sentence. Then $\Delta \vDash \phi\,;\vec{s}$ iff for every $\mathcal{M}$, $w \in W_{\mathcal{M}}$ and homomorphism $f : \mathbb{N}^{<\omega} \to \mathrm{WS}_{\mathcal{M}}$,

$$\text{if } w, f(\vec{t}) \vDash \psi \text{ for each } \psi\,;\vec{t} \in \Delta \text{ then } w, f(\vec{s}) \vDash \phi$$

I shall write $w, f \vDash \Delta$ from now on to mean that, for all $(\psi\,;\vec{s}) \in \Delta$, $w, f(\vec{s}) \vDash \psi$. Now, a soundness theorem can be proved for labelled formulas:

Proposition A.1. *Let Δ be a set of labelled formulas and $\phi\,;\vec{s}$ a labelled formula. Then*

$$\textit{If } \Delta \vdash \phi\,;\vec{s} \textit{ then } \Delta \vDash \phi\,;\vec{s}$$

Proof. First, it needs to be checked that each deduction rule is sound. That is, when $\langle\langle \Gamma_1, \theta_1\,;\vec{s}_1\rangle, \ldots, \langle \Gamma_k, \theta_k\,;\vec{s}_k\rangle\rangle$ is an instance of a deduction rule, then if $\Gamma_i \vDash \theta_i\,;\vec{s}_i$ for each $i < k$, then $\Gamma_k \vDash \theta_k\,;\vec{s}_k$.

That this is so for the propositional connectives is standard (since the labels do not really play a role). It can be proved for the rules for operators as follows:

($\downarrow$-I): The deduction rule for ($\downarrow$-I) is $\langle\langle \Gamma, \phi\,;\vec{s}\rangle, \langle \Gamma, \downarrow\phi,;\vec{s}, n\rangle\rangle$. Suppose that $\Gamma \vDash \phi\,;\vec{s}$, so that for any w, f, if $w, f \vDash \Gamma$, then $w, f(\vec{s}) \vDash \phi$. Now, consider w, f such that $w, f \vDash \Gamma$, and we wish to show that $w, f(\vec{s}, n) \vDash \downarrow\phi$. Since f is a homomorphism, $f(\vec{s}, n) = f(\vec{s}), w'$ for some w' such that $t(f(\vec{s}))Rw'$. So, $w, f(\vec{s}, n) \vDash \downarrow\phi$ iff $w, f(\vec{s}), w' \vDash \downarrow\phi$ iff $w, f(\vec{s}) \vDash \phi$, which we already have.

($\downarrow$-E): The deduction rule is $\langle\langle \Gamma, \downarrow\phi\,;\vec{s}\rangle, \langle \Gamma, \phi,;\vec{s}-\rangle\rangle$. There are two cases to consider: Firstly, when $\vec{s}$ is empty, so that $\vec{s}-$ is also empty, and secondly when $\vec{s}$ is not empty.

Consider the first case. Suppose that $\Gamma \vDash \downarrow\phi\,;-$ so that for any w, f, if $w, f \vDash \Gamma$ then $\langle w\rangle \vDash \downarrow\phi$ (since $\vec{s}$ is empty, f can be safely ignored). Now consider some w, f such that $w, f \vDash \Gamma$ and so $\langle w\rangle \vDash \downarrow\phi$. So, by the definition of $\downarrow$, $\langle w\rangle \vDash \phi$, which is as required.

Suppose that $\vec{s}$ is not empty. Suppose $\Gamma \vDash \downarrow\phi\,;\vec{s}$ so that for all w, f, if $w, f \vDash \Gamma$ then $w, f(\vec{s}) \vDash \downarrow\phi$. Now consider w, f such that $w, f \vDash \Gamma$, so $w, f(\vec{s}) \vDash \downarrow\phi$. By definition of $\downarrow$, $(w, f(\vec{s}))- \vDash \phi$. But, since $l(\vec{s}) > 0$, $(w, f(\vec{s}))- = w, f(\vec{s}-)$. So $w, f(\vec{s}-) \vDash \phi$ which is what is required.

($\Box$-E): The deduction rule is $\langle\langle \Gamma, \Box\phi\,;\vec{s}\rangle, \langle \Gamma, \phi\,;\vec{s}, n\rangle\rangle$. Suppose that $\Gamma \vDash \Box\phi\,;\vec{s}$, so that for all w, f, if $w, f \vDash \Gamma$ then $w, f(\vec{s}) \vDash \Box\phi$. Now consider w, f such that $w, f \vDash \Gamma$, and we wish to show that $w, f(\vec{s}, n) \vDash \phi$. By the properties of f, $f(\vec{s}, n) = f(\vec{s}), w'$ for some w' such that wRw'. But, since $w, f(\vec{s}) \vDash \Box\phi$ and by the definition of $\Box$, for any w' such that wRw', $w, f(\vec{s}), w' \vDash \phi$. So, $w, f(\vec{s}, n) \vDash \phi$ as required.

($\Box$-I): The deduction rule is $\langle\langle \Gamma, \phi\,;\vec{s}\rangle, \langle \Gamma, \Box\phi\,;\vec{s}-\rangle\rangle$ with the restriction that if $\psi\,;\vec{t} \in \Gamma$ then $\vec{s}$ properly extends $\vec{t}$. Suppose that $\Gamma \vDash \Box\phi\,;\vec{s}$ for such a Γ. So, for any w, f, if $w, f \vDash \Gamma$ then $w, f(\vec{s}) \vDash \phi$.

Now, consider w, f such that $w, f \vDash \Gamma$. We wish to show that $w, f(\vec{s}-) \vDash \Box\phi$. Suppose not, then for some w' such that $t(f(\vec{s}-))Rw'$, $w, f(\vec{s}-), w' \nvDash \phi$.

Now, it is clear that a homomorphism f' can be defined such that:

- $f'(\vec{t}) = f(\vec{t})$ for all $\vec{t}$ which $\vec{s}$ properly extends

- $f'(\vec{s}) = f(\vec{t}-), w'$

and allowed to take any other permitted value elsewhere.

So, $w, f'(\vec{s}) \not\models \phi$. So, $w, f' \not\models \Gamma$. So, there is $\psi\,;\vec{t} \in \Gamma$ such that $w, f'(\vec{t}) \models \neg\psi$. Now, by the restriction on the deduction rule, $\vec{s}$ properly extends $\vec{t}$. But then, $f'(\vec{t}) = f(\vec{t})$, but $w, f(\vec{t}) \models \psi$.

Now, since each deduction rule is sound, it is simple to show the result by induction on lengths of deductions. □

Finally, a soundness result can be given for the kind of entailments which are of principal importance—those between unlabelled sentences (since entailments between labelled sentences is supposed to be purely instrumental):

Proposition A.2. *Let Γ be a set of sentences of $\mathcal{L}_{prop}$ and ϕ a sentence of $\mathcal{L}_{prop}$. Then:*

$$\textit{If } \Gamma \vdash \phi \textit{ then } \Gamma \models \phi$$

Proof. All that is needed for the proof is that if $\Gamma^* \models \phi\,;-$, then $\Gamma \models \phi$ (where Γ^* is as in definition A.3). Suppose that $\Gamma^* \models \phi\,;-$. Consider some $w \in W_{\mathcal{M}}$ such that $\langle w\rangle \models \psi$ for each $\psi \in \Gamma$. Then clearly $w, f(\vec{s}) \models \psi$ for each $\psi, \vec{s} \in \Gamma^*$, since $\vec{s}$, and so $f(\vec{s})$ is empty in each case. So $w, f(\langle\rangle) \models \phi$. I.e. $\langle w\rangle \models \phi$ as required.

The soundness result then follows easily. □

A.1.5 Strengthening the logic

In developing a logic suitable for the background theory of abstraction without absolute generality, the logic will need to be strengthened to at least S4 (if not stronger). With the presence of the $\downarrow$ operator in the language, however, this is not simple. For example, consider the (4) axiom scheme:

(4) $$\Box\phi \rightarrow \Box\Box\phi$$

Where ϕ is $\downarrow\psi$ for some formula ψ, this has as an instance:

$$\Box\downarrow\phi \rightarrow \Box\Box\downarrow\phi$$

In the presence of such an axiom, it is simple to derive $\psi \rightarrow \Box\psi$, resulting in a modal collapse.

Similarly, the (4) axiom could be added by strengthening the ($\Box$-I) rule, so that the premise may depend on assumptions, as long as they are of the form $\Box\psi$. This has the same undesirable result.

The unrestricted (T) axiom (or a corresponding rule) could also cause problems. If the necessitation of the (T) axiom is also added (as would be expected), then it would have the following instance:

$$\Box(\Box\downarrow\phi \rightarrow \downarrow\phi)$$

which is equivalent to $\Box(\phi \rightarrow \downarrow\phi)$. This can be seen to entail $\Diamond\phi \rightarrow \phi$, again resulting in a modal collapse.

In order to add such axioms or strengthened rules, restrictions must be added on the kind of formulas which define instances.[1]

Define the *degree* of a sentence recursively as follows:[2]

- For atomic p, $\deg(p) = 0$
- $\deg(\phi \wedge \psi) = \max(\deg(\phi), \deg(\psi))$
- $\deg(\neg\phi) = \deg(\phi)$
- $\deg(\Box\phi) = \deg(\phi) \dot{-} 1$
- $\deg(\downarrow\phi) = \deg(\phi) + 1$

where $n \dot{-} m = \max(n - m, 0)$. Intuitively speaking, the degree of a formula is the depth of backtracking operators which are not cancelled out by a modal operator. So, for example, $\deg(\Box\downarrow\downarrow p) = 1$, since it features two scoping operators, only one of which is paired with a modal operator.

The following lemma will then be useful:

Lemma A.3. *Let ϕ be any sentence. Then, for any model $\mathcal{M}$ and world sequence $\langle w_m, \ldots, w_0\rangle$ (note the reversed order), with $m \geq \deg(\phi)$,*

$$\langle w_m, \ldots, w_0\rangle \vDash \phi \textit{ iff } \langle w_{\deg(\phi)}, \ldots, w_0\rangle \vDash \phi.$$

Proof. This can be proved by induction on the complexity of sentences.

The base case, where ϕ is an atomic sentence p, is simple. Then $\deg\phi = 0$, and $\langle w_m, \ldots, w_0\rangle \vDash p$ iff $\langle w_0\rangle \vDash p$ is immediate, as required.

Suppose ϕ is $\psi \wedge \theta$, and $\deg(\phi \wedge \psi) = n$. Without loss of generality, suppose that $\deg(\psi) = n$ and $\deg(\theta) = k \leq n$. Then,

	$\langle w_m, \ldots, w_0\rangle \vDash \psi \wedge \theta$	
iff	$\langle w_m, \ldots, w_0\rangle \vDash \psi$ and $\langle w_m, \ldots, w_0\rangle \vDash \theta$	
iff	$\langle w_n, \ldots, w_0\rangle \vDash \psi$ and $\langle w_k, \ldots, w_0\rangle \vDash \theta$	(by the inductive hypothesis)
iff	$\langle w_n, \ldots, w_0\rangle \vDash \psi$ and $\langle w_n, \ldots, w_0\rangle \vDash \theta$	(by the inductive hypothesis)
iff	$\langle w_n, \ldots, w_0\rangle \vDash \psi \wedge \theta$	

as required. The proof for negation is similar.

Suppose that ϕ is $\Box\psi$. There are two cases: where $\deg(\psi) = \deg(\Box\psi) = 0$ and where

[1] Another approach to restricting these axioms is given by Parsons (1983b, Appendix 1.2) for a slightly different language. Parsons' approach is not however readily applicable to the language featuring $\downarrow$. His language features scoping operators which, instead of *exempting* a subformula from some fixed number of modal operators, signify that some subformula falls under the scope of some fixed number of modal operators.

[2] This is the same definition as Hodes (1984b, p. 426) gives.

$\deg(\phi) > 0$. Suppose that $\deg(\psi) = \deg(\Box\psi) = 0$. Then,

$$\begin{array}{lll} & \langle w_m, \ldots, w_0\rangle \vDash \Box\psi & \\ \text{iff} & \text{for all } w' \text{ s.t. } w_0 R w', \langle w_m, \ldots, w_0, w'\rangle \vDash \psi & \\ \text{iff} & \text{for all } w' \text{ s.t. } w_0 R w', \langle w'\rangle \vDash \psi & \text{(by the inductive hypothesis)} \\ \text{iff} & \text{for all } w' \text{ s.t. } w_0 R w', \langle w_0, w'\rangle \vDash \psi & \text{(by the inductive hypothesis)} \\ \text{iff} & \langle w_0\rangle \vDash \Box\psi & \end{array}$$

as required.

Suppose that $\deg(\psi) = n > 0$. Then,

$$\begin{array}{lll} & \langle w_m, \ldots, w_0\rangle \vDash \Box\psi & \\ \text{iff} & \text{for all } w' \text{ s.t. } w_0 R w', \langle w_m, \ldots, w_0, w'\rangle \vDash \psi & \\ \text{iff} & \text{for all } w' \text{ s.t. } w_0 R w', \langle w_{n-1}, \ldots, w_0, w'\rangle \vDash \psi & \text{(by the inductive hypothesis)} \\ \text{iff} & \langle w_{n-1}, \ldots, w_0\rangle \vDash \Box\psi & \end{array}$$

as required.

Suppose that ϕ is $\downarrow\psi$ and $\deg(\psi) = n$. Then,

$$\begin{array}{lll} & \langle w_m, \ldots, w_0\rangle \vDash \downarrow\psi & \\ \text{iff} & \langle w_m, \ldots, w_1\rangle \vDash \psi & \\ \text{iff} & \langle w_{n+1}, \ldots, w_1\rangle \vDash \psi & \text{(by the inductive hypothesis)} \\ \text{iff} & \langle w_{n+1}, \ldots, w_0\rangle \vDash \downarrow\psi & \end{array}$$

as required. □

Now, the rules for $\Box$ can be strengthened so that restricted versions of (T), (4) and (G) are provable. We add the following rules for (4) and (T):

$$(\Box\text{-R})\ \frac{\Box\phi\,;\vec{s}}{\Box\phi\,;\vec{s}, n} \qquad (\Box\text{-E}')\ \frac{\Box\phi\,;\vec{s}}{\phi\,;\vec{s}}$$

both with the restriction that $\deg(\phi) = 0$. In the presence of both these rules, ($\Box$-E) becomes redundant. (($\Box$-R) stands for necessity *reiteration*, after a similar rule in Siemens (1977)).

There will be corresponding derived rules for possibility rather than necessity. These are:

$$(\Diamond\text{-R})\ \frac{\Diamond\phi\,;\vec{s}}{\Diamond\phi\,;\vec{s}-} \qquad (\Diamond\text{-I}')\ \frac{\phi\,;\vec{s}}{\Diamond\phi\,;\vec{s}}$$

It is simple to check that these are derivable in the presence of the corresponding rules for necessity.

That these are sound in transitive and reflexive frames can then be proved using the lemma:

Proposition A.4. *Let $\mathcal{M}$ be a model. If the accessibility relation R is transitive, then ($\Box$-R) is sound. If R is reflexive then ($\Box$-E′) is sound.*

Proof. **($\Box$-R):** Suppose that $w, f(\vec{s}) \vDash \Box\phi$, and we aim to show that $w, f(\vec{s}, n) \vDash \Box\phi$. I.e. that for any w' and w'' such that $t(f(\vec{s}))Rw'$ and $w'Rw''$, $w, f(\vec{s}), w', w'' \vDash \phi$. But, since $\deg(\phi) = 0$, and by lemma A.3, this is equivalent to $w'' \vDash \phi$, and again by the lemma, this is equivalent to $w, f(\vec{s}), w'' \vDash \phi$. By transitivity, $t(f(\vec{s}))Rw''$, so $w, f(\vec{s}), w'' \vDash \phi$ by the definition of $\Box$, as required.

The proof for (T) is similar. □

So get the logic s4.2, either the axiom:

(G) $$\Diamond\Box\phi \rightarrow \Box\Diamond\phi$$

or the corresponding rule:

$$\text{(G)}\ \frac{\Diamond\Box\phi\,;\vec{s}}{\Box\Diamond\phi\,;\vec{s}}$$

can be added, again with the restriction that $\deg(\phi) = 0$.[3] It is again simple to check that this is sound for directed frames.

A.2 Expansion Logic

The aim of this section will be to develop the propositional logic of the previous sections into a logic EL, for *expansion logic*. This will be the logic which will serve as a background logic for introducing modal abstraction principles.

Let $\mathcal{L}^2$ be a second-order language whose propositional part (i.e. the connectives and modal operators) is that of $\mathcal{L}^{\text{prop}}$. Add *object variables* $x, y, z, \ldots$, *concept variables* $F, G, H, \ldots$, quantifiers and an identity symbol. The language can also feature non-logical constants of various kinds. In particular, it may feature one or more *abstraction operators* § such that, where T is a monadic concept term—i.e. a monadic concept variable, a predicate or, where the language features them, a lambda term—§T is an object term.

An additional formation rule which the language will feature is that ↓ may attach to object terms as well as formulas. So, where t is an object term, ↓t is also an object term. The intended affect is for the (possibly non-rigid) term to be exempted from the scope of the innermost modal operator for the purposes of evaluating reference.

A.2.1 Model Theory

A model is a 4-tuple $\mathcal{M} = \langle W, R, D, \delta, I\rangle$ where W and R are as in the propositional case. D is a domain, and $\delta : W \rightarrow \mathcal{P}(D)$ such that $\bigcup_{w\in W} \delta(w) = D$. Let $D_2 = \{f : W \rightarrow \mathcal{P}(D) \ : \ f(w) \subseteq \delta(w)\}$. D_2 will be the domain of second-order variables.[4] Finally, I is

[3] The rule may be formulated in various other ways, avoiding the appearance of two nested modalities in the premise and conclusion. But nothing is lost in having such nested modalities, and the resulting rule is much simpler.

[4] This particular definition will mean that the semantics will validate that for any term t, $\forall F[Ft \rightarrow \exists y(y = t)]$. I.e. any atomic sentence containing t formed from a concept variable will ensure that the referent of t

a mapping from non-logical constants to objects of the appropriate type. For example, in the case of individual constants c, $I(c) \in D$, for monadic predicates P, $I(P) \in D_2$, and—of particular interest—in the case of abstraction operators, $I(\S)$ is a function $I(\S) : W \times D_2 \to D$.

For the purposes of EL, models will also be restricted to those in which the accessibility relation is reflexive, transitive and directed, and where the domain is increasing, in the sense that if wRw', then $\delta(w) \subseteq \delta(w')$. It will be convenient to treat the accessibility as an ordering relation, and so I will write $\leq$ in place of R from now on.

As in the propositional case, satisfaction will be defined for world sequences. First, the interpretation must be extended to all terms. Let a be an assignment of appropriate items to first- and second-order variables. Then, a valuation of each object term and concept term at a world sequence $[\![\cdot]\!]^{\vec{w},a}$ can be defined as follows:

- For object variables, $[\![x]\!]^{\vec{w},a} = a(x)$,
- For constants, $[\![c]\!]^{\vec{w},a} = I(c)$,
- For concept variables, $[\![F]\!]^{\vec{w},a} = a(F)$,
- For predicates, $[\![P]\!]^{\vec{w},a} = I(P)$,
- Where t is an object term, $[\![{\downarrow}t]\!]^{\vec{w},a} = [\![t]\!]^{\vec{w}-,a}$,
- Where $\S$ is an abstraction operator and T a concept term, $[\![\S T]\!]^{\vec{w},a} = I(\S)(t(\vec{w}), [\![T]\!]^{\vec{w},a})$.

  (Similar valuations can be given for other non-logical constants.)

Now, satisfaction at a world sequence relative to a variable assignment can be defined. The clauses for connectives and modal operators are as for the propositional case.

The additional clauses which are needed are:

$\vec{w}, a \vDash s = t$ iff $[\![s]\!]^{\vec{w},a} = [\![t]\!]^{\vec{w},a} \in \delta(w')$ for some w' s.t. $t(\vec{w})Rw'$

$\vec{w}, a \vDash Tt$ iff $[\![t]\!]^{\vec{w},a} \in [\![T]\!]^{\vec{w},a}(t(\vec{w}))$ and $[\![t]\!]^{\vec{w},a} \in \delta(w')$ for some w' s.t. $t(\vec{w})Rw'$

$\vec{w}, a \vDash \forall x\phi$ iff $\vec{w}, a^x_d \vDash \phi$ for every $d \in \delta(t(\vec{w}))$

$\vec{w}, a \vDash \forall F\phi$ iff $\vec{w}, a^F_f \vDash \phi$ for every $f \in D_2$

Where a^x_d is the assignment that differs from a only in assigning d to x (and similarly for second-order variables).

Semantic consequence is then defined in just the same way as before.

A.2.2 Proof theory

The proof theory is very similar to the propositional case, but with the addition of the following rules and axioms:

- Suitable rules for quantifiers (free-logical for the first order quantifiers, and non-free for the second-order quantifiers)

will fall within the current range of the quantifiers. This partly goes against the rejection of a negative free logic. However, such a result will not be derivable in the proof theory, and the semantics could be modified to remove the validity. However, this semantics is simpler, and will do for the purpose to which it is being put.

Alternatively, D_2 could be given as the set of all functions $f : W \to \mathcal{P}(D)$.

- An elimination rule for identity:

$$\frac{\phi(s);\vec{s} \qquad s = t;\vec{s}}{\phi(t);\vec{s}}$$

- A strong comprehension for concepts:

$$\exists F \Box \forall x (Fx \leftrightarrow \phi)$$

- The converse Barcan Formula:

$$\Box \forall x \phi \rightarrow \forall x \Box \phi$$

- The ($\Diamond E!$-I) rule:

$$(\Diamond E!\text{-I}) \ \frac{\phi(t) \quad \phi \text{ atomic}}{\Diamond \exists x (x = \downarrow t)}$$

- The following rules and axiom governing $\downarrow$ as applied to terms:

$$(\downarrow =\text{-E}) \ \frac{\downarrow s = \downarrow t;\vec{s}}{s = t;\vec{s}-} \qquad (\downarrow =\text{-I}) \ \frac{s = t;\vec{s}}{\downarrow s = \downarrow t;\vec{s}, n}$$

(RV) $\qquad x = \downarrow x$ (where x is a variable)

One derived axiom which is very useful in appendix B is the following:

(NNE) $\qquad \Box(E!x \rightarrow \Box E!x)$

which follows fairly immediately from CBF.

Soundness is then relatively simple to prove. The non-standard cases are for ($\Diamond E!$-I) and the rules for $\downarrow$ applied to terms. That ($\Diamond E!$-I) is sound is a simple consequence of the unusual clauses for atomic formulas. That the rules of $\downarrow$ applied to terms are sound is a simple consequence of the definition of $[\![\downarrow t]\!]$, and that $[\![x]\!]$ does not depend on the world at which it is evaluated (so variables are rigid).

Appendix B

Formal proofs

Proposition 7.1:
Assume $y = \varepsilon F$. We aim to prove $x \in y \leftrightarrow (Fx \wedge E!x)$. The right to left direction follows almost immediately:

1)	$Fx \wedge E!x$	; –	(Assumption)
2)	$y = \varepsilon F$	; –	(Assumption)
3)	$\exists G \exists z(y = \varepsilon G \wedge x = z \wedge Fz)$	; –	(From 1,2 and $\exists$-I)
4)	$\Diamond \exists G \exists z(y = \varepsilon G \wedge x = z \wedge Fz)$	; –	(T)
5)	$x \in y$	; –	(def.)

For the left to right direction:

6)	$x \in y$	; –	(Assumption)
7)	$\Diamond \exists G \exists z(y = \varepsilon G \wedge x = z \wedge Fz)$	; –	(Def.)
8)	$y = \varepsilon G \wedge Gx \wedge E!x$	; 0	(Assumption for $\Diamond$-E and $\exists$-E)
9)	$\varepsilon G = \downarrow \varepsilon F$	; 0	(Both are y, from 8, main assumption)
10)	$\forall z(Gz \leftrightarrow \downarrow(Fz \wedge E!x))$	; 0	(BLV$\downarrow$)
11)	$\downarrow(Fx \wedge E!x)$	; 0	(From 8, 10 by $\forall$-E and $\leftrightarrow$-E)
12)	$Fx \wedge E!x$	; –	(11, ($\downarrow$-E))

□

Proposition 7.3:
We need to show $\Box \forall x(x \in y \leftrightarrow \downarrow x \in y)$. For the left to right direction:

1)	$E!x$	; 0	(Assumption)
2)	$x \in y$	; 0	(Assumption)
3)	$\Diamond \exists F \exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	; 0	(2, Definition)
4)	$\Diamond \exists F \exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	; –	($\Diamond$-R)
5)	$x \in y$	; –	(4, Definition)
6)	$\downarrow(x \in y)$	; 0	(5, $\downarrow$-I.)
7)	$x \in y \rightarrow \downarrow(x \in y)$	; 0	(2,6,$\rightarrow$-I, discharging 2)

Then, for the right to left direction:

8)	$\downarrow(x \in y)$	;0	(Assumption)
9)	$x \in y$	;–	(8, $\downarrow$-E)
10)	$\Diamond\exists F\exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	;–	(9, definition)
11)	$y = \varepsilon F \wedge Fx \wedge E!x$	;1	(Assumption for $\Diamond$-E and $\exists$-E)
12)	$\Box\forall x(Gx \leftrightarrow \downarrow(E!x \wedge Fx))$	;1	(Assumption, from comprehension)
13)	$\forall x(Gx \leftrightarrow \downarrow(E!x \wedge Fx))$	;10	(12, $\Box$-E)
14)	$\varepsilon G = \downarrow\varepsilon F$	;10	(13, BLV$\downarrow$)
15)	$y = \downarrow\varepsilon F$	;10	(11, $\downarrow$-I, RV)
16)	$y = \varepsilon G$	;10	(14, 15, =-E)
17)	$E!x$	;10	(11, NNE)
18)	$\downarrow(E!x \wedge Fx)$	;10	(11, $\downarrow$-I)
19)	Gx	;10	(13,17,18, $\forall$-E and $\leftrightarrow$-E)
20)	$\exists F\exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	;10	(16,17,19, $\exists$-I)
21)	$\Box\exists F\exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	;1	(20, $\Box$-I)
22)	$\Diamond\Box\exists F\exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	;–	(21, $\Diamond$-I. We can now discharge 11)
23)	$\Box\Diamond\exists F\exists z(y = \varepsilon F \wedge Fx \wedge x = z)$	;–	(22, G)
24)	$x \in y$	;0	(23, $\Box$-E, definition)
25)	$\downarrow(x \in y) \rightarrow x \in y$	;0	(7, 24 $\rightarrow$-I, discharging 7)

Finally, we can bring the whole thing together:

26)	$\Box\forall x(x \in y \leftrightarrow \downarrow x \in y)$	;–	(discharge 1)

□

Proposition 7.5:
Assume $x \in_1 y$. Then we want to show that $x \in_0 y \leftrightarrow (E!x \wedge \mathrm{Set}_0(y))$. The left to right direction is trivial and follows directly from the definition. For the right to left direction, we assume $E!x \wedge \mathrm{Set}_0(y)$ and show $x \in_0 y$:

1)	$x \in_1 y$	;–	(Assumption)
2)	$E!x$	;–	(Assumption)
3)	$\mathrm{Set}_0(y)$	;–	(Assumption)
4)	$\exists F(y = \varepsilon F)$	;–	(Def. of Set_0)
5)	$y = \varepsilon F$	;–	(Assumption for $\exists$-E)
6)	$\Diamond\exists G(y = \varepsilon G \wedge Gx \wedge E!x)$	;	(Def., from 1)
7)	$y = \varepsilon G \wedge Gx \wedge E!x$	;0	(Assumption, for $\Diamond$-E and $\exists$-E)
8)	$\varepsilon G = \downarrow\varepsilon F$	;0	(Both are y, from 5,7)
9)	$\forall z(Gz \leftrightarrow \downarrow(Fz \wedge E!z))$	;0	(BLV$\downarrow$)
10)	$\downarrow Fx$	;0	(From 7, 9 by $\forall$-E and $\leftrightarrow$-E)
11)	$Fx \wedge y = \varepsilon F \wedge E!x$	;–	(From 2,5,10 by $\downarrow$-E and $\wedge$-I)
12)	$\exists F\exists z(y = \varepsilon F \wedge z = x \wedge Fx)$	;–	(From 11 by $\exists$-I)
13)	$x \in_0 y$		

Assumptions 5 and 7 can then be discharged by $\exists$-E and $\Diamond$-E, and then the final result is easily assembled by conditional proof, discharging the other assumptions. □

Proposition 7.6:
Suppose $\mathrm{Set}_1(x)$. Then, for the left to right direction, we have the following formal proof:

1)	$\exists F(x = \varepsilon F)$	; –	(Assumption. Is definition of $\mathrm{Set}_0(x)$)
2)	$x = \varepsilon F$	; –	(Assumption for $\exists$-E)
3)	$E!y$	; 0	(Assumption)
4)	$y \in x$	; 0	(Assumption)
5)	$y \in x$	; –	(By R_1-$\in$)
6)	$\Diamond\exists G\exists z(z = y \wedge x = \varepsilon G \wedge Gy)$	; –	(Definition of $\in$)
7)	$x = \varepsilon G \wedge E!y \wedge Gy$	; 1	(Assumption for $\Diamond$-E and $\exists$-E)
8)	$\varepsilon G = \downarrow\varepsilon F$	; 1	(2,7, =-E)
9)	$\forall y(Gy \leftrightarrow \downarrow(Fy \wedge E!y))$	; 1	(BLV$\downarrow$)
10)	$\downarrow(Fy \wedge E!y)$	; 1	(7,9, $\forall$-E, $\rightarrow$-E)
11)	$E!y$	; –	(10, $\downarrow$-E)
12)	$\downarrow E!y$	; 0	($\downarrow$-I)
13)	$y \in x \rightarrow \downarrow E!y$	; 0	($\rightarrow$-I, discharge 4)
14)	$\Box\forall y(y \in x \rightarrow \downarrow E!y)$	; –	($\forall$-I, $\Box$-I, discharge 3)
15)	$\Box\forall y(y \in x \rightarrow \downarrow E!y)$	; –	(Discharge 2,7 by $\Diamond$-E and $\exists$-E)

For the right to left direction, we have the following formal proof:

1)	$\Box\forall y(y \in x \rightarrow \downarrow E!y)$	; –	(Assumption)
2)	$\Diamond\exists G(x = \varepsilon G)$	; –	(Definition of $\mathrm{Set}_1(x)$)
3)	$x = \varepsilon G$	; 0	(Assumption for $\exists$-E and $\Diamond$-E)
4)	$\forall y(y \in x \rightarrow \downarrow E!y)$	; 0	(1, $\Box$-E)
5)	$\Box\forall y(Fy \leftrightarrow y \in x)$	; –	(Assumption, from comprehension)
6)	$E!y$	; 0	(Assumption)
7)	Gy	; 0	(Assumption)
8)	$y \in x$	; 0	(3, 7 and corollary 7.2)
9)	$\downarrow E!y$	; 0	(From 4,6,8)
10)	$E!y$	; –	($\downarrow$-E)
11)	$y \in x$	; –	(8, R_1-$\in$)
12)	Fy	; 0	(From 5,10,11)
13)	$Gy \rightarrow \downarrow(E!y \wedge Fy)$	; 0	(From 7,10,12, discharge 7)
14)	$\downarrow(E!y \wedge Fy)$	; 0	(Assumption)
15)	$E!y \wedge Fy$	; –	($\downarrow$-E)
16)	$y \in x$	; –	(From 5, 15)
17)	$y \in x$	; 0	(16, R_1-$\in$)
18)	Gy	; 0	(corollary 7.2)
19)	$\forall y(Gy \leftrightarrow \downarrow(E!y \wedge Fy))$	; 0	(13, 14, 18, Discharge 6)
20)	$\varepsilon F = \varepsilon G$	; 0	(BLV$\downarrow$)
21)	$x = \varepsilon G$	; –	
22)	$\exists F(x = \varepsilon F)$	; –	
23)	now discharge everything		

□

Proposition 7.7:
Assume that ϕ is such that $\Diamond \mathrm{R}_0^x\text{-}\phi$ and $\Box \mathrm{R}_1^x\text{-}\phi$. Now, let F be such that $\Box\forall x(Fx \leftrightarrow \phi)$ by comprehension, and we have $\Diamond\exists y(y = \varepsilon F)$ by ($\Diamond E!$-I). Start by assuming $y = \varepsilon F\,;0$ and we show $\Box(y = \varepsilon F)\,;0$.

First, we set up some basics:

1)	$\Diamond\Box\forall x(\phi(x) \leftrightarrow \downarrow(E!x \wedge \phi(x)))$	; –	(by $\Diamond \mathrm{R}_0^x$-)
2)	$\Box\forall x(\phi(x) \leftrightarrow \downarrow(E!x \wedge \phi(x)))$	;0	(Assumption for $\Diamond$-E)
3)	$\Box\forall x(Fx \leftrightarrow \phi(x))$	; –	(Assumption, by comprehension)
4)	$\Box\forall x(Fx \leftrightarrow \phi(x))$	;0	($\Box$-R)
5)	$y = \varepsilon F$	;0	(Assumption)

Now, we aim to prove $\Box(y = \varepsilon F)\,;0$ by means of BLV:

6)	$E!z$	;00	(Assumption)
7)	Fz	;00	(Assumption)
8)	$\phi(z)$	;00	(By 4,6,7)
9)	$\downarrow(\phi(z) \wedge E!z)$	;00	(By 2,6,8)
10)	$\phi \wedge E!z$	;0	($\downarrow$-E)
11)	$Fz \wedge E!z$	;0	(By 3,10)
12)	$Fz \rightarrow \downarrow(Fz \wedge E!z)$	;00	(7,11, discharge 7)
13)	$\downarrow(Fz \wedge E!z)$	;00	(Assumption)
14)	$Fz \wedge E!z$	;0	($\downarrow$-E)
15)	$\phi(z) \wedge E!z$	;0	(By 3,14)
16)	$E!z$	;00	(NNE)
17)	$\downarrow(\phi \wedge E!z)$	;00	(15, $\downarrow$-I)
18)	$\phi(z)$	;00	(By 2,16,17)
19)	Fz	;00	(By 4,16,18)
20)	$\downarrow(Fz \wedge E!z) \rightarrow Fz$	;00	(13,19, discharge 13)
21)	$\forall z(Fz \leftrightarrow \downarrow(Fz \wedge E!z))$	;00	(6,12,20, discharge 6)
22)	$\varepsilon F = \downarrow\varepsilon F$	;00	(BLV$\downarrow$)
23)	$y = \varepsilon F$	;00	(5,22,$\downarrow$-I)
24)	$\Box(y = \varepsilon F)$	;0	($\Box$-I)

Now, we continue the proof to arrive at $\Box\forall x\downarrow(x \in y \leftrightarrow \phi)\,;-$:

25)	$E!x$	;1	(Assumption)
26)	$\Box\Diamond E!x$	; –	(From 25, by NNE and G)
27)	$x \in y$	; –	(Assumption)
28)	$x \in y$	;0	(By R_1^x- $\in$)
29)	$\Diamond E!x$	;0	(By 26, $\Box$-E)
30)	$E!x$	;01	(Assumption, for $\Diamond$-E)
31)	$x \in y$	;01	(By R_1^x- $\in$)

32) $y = \varepsilon F$; 01 (By 24)
33) Fx ; 01 (By 30,31,32 and corollary 7.2)
34) $\phi(x)$; 01 (By 4,30,33)
35) $\phi(x) \wedge E!x$; 0 (By 2,30,34. We can now discharge 30)
36) $\phi(x) \wedge E!x$; – (By 2)
37) $x \in y \rightarrow \phi(x)$; – (27, 36, discharge 27)

38) $\phi(x)$; – (Assumption)
39) $E!x$; – (By R_0^x-ϕ)
40) $\phi(x)$; 0 (By R_0^x-ϕ)
41) $E!x$; 0 (By R_0^x-ϕ)
42) Fx ; 0 (By 3,40,41)
43) $\exists F(y = \varepsilon F \wedge E!x \wedge Fx)$; 0 (By 5,41,42)
44) $\Diamond\exists F(y = \varepsilon F \wedge Fx)$; – (By 43, $\Diamond$-I)
45) $\phi \rightarrow x \in y$; – (By 38,43, def., discharging 38)

The rest of the proof is then simple. □

Proposition 7.12:
The aim is to prove that $\Box\forall x((x \subseteq u)^{\Diamond} \leftrightarrow \downarrow(x \subseteq u)^{\Diamond})$. For the left to right direction:

1) $E!x$; 0 (Assumption.)
2) $x \subseteq^{\Diamond} y$; 0 (Assumption.)

(we now aim to prove $x \subseteq^{\Diamond} y$; –, i.e. $\Box\forall z\downarrow(z \in x \rightarrow z \in y)$; –)

3) $E!z$; 1 (Assumption.)
4) $z \in x$; – (Assumption.)
5) $z \in x$; 1 (By R_1- $\in$)
6) $\Box(z \in x \wedge E!z)$; 1 (NNE and R_1- $\in$)
7) $\Diamond\Box(z \in x \wedge E!z)$; – ($\Diamond$-I)
8) $\Box\Diamond(z \in x \wedge E!z)$; – (G)
9) $\Diamond(z \in x \wedge E!z)$; 0 ($\Box$-E)
10) $z \in x \wedge E!z$; 00 (Assumption for $\Diamond$-E)
11) $z \in x$; 0 (By R_1- $\in$)
12) $z \in y$; 0 (By 2)
13) $z \in y$; – (By R_1- $\in$)

Now, assumptions can be discharged via $\Box$-I and $\forall$-I to get:

14) $\Box\forall z\downarrow(z \in x \rightarrow z \in y)$; – (Discharging 3,4)
15) $\Box\forall x(x \subseteq^{\Diamond} y \rightarrow \downarrow x \subseteq^{\Diamond} y)$; – (Discharging 1,2)

Finally, 10 can be safely discharged by ($\Diamond$-E).

For the right to left direction:

1) $E!x$; 0 (Assumption.)
2) $x \subseteq^{\Diamond} y$; – (Assumption.)

(we now aim to prove $x \subseteq^{\Diamond} y$; 0, i.e. $\Box\forall z\downarrow(z \in x \rightarrow z \in y)$; 0)

3) $E!z$; 00 (Assumption.)
4) $z \in x$; 0 (Assumption.)

(we can't use R_1- $\in$ directly here to get $z \in x$; –, so the next 5 lines do this by going via the $\Diamond$-R rule)

5) $z \in x$; 00 (By R_1- $\in$)
6) $\Diamond(z \in x \wedge E!z)$; 0 (3,5, $\Diamond$-I)
7) $\Diamond(z \in x \wedge E!z)$; – ($\Diamond$-R)
8) $z \in x \wedge E!z$; 1 (Assumption for $\Diamond$-E)
9) $z \in x$; – (By R_1- $\in$)
10) $z \in y$; – (By 2)
11) $z \in y$; 0 (By R_1- $\in$)

Now, assumptions can be discharged via $\Box$-I and $\forall$-I to get:

12) $\Box\forall z\downarrow(z \in x \to z \in y)$; 0 (Discharging 3,4)
13) $\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow x \subseteq^{\Diamond} y)$; 0 (Discharging 1,2)

Finally, 8 can be safely discharged by ($\Diamond$-E). □

Proposition 7.13:
First, we show as a lemma, $\Box(\exists F(y = \varepsilon F) \to \Box\forall x((x \subseteq y)^{\Diamond} \to \downarrow\exists F(x = \varepsilon F))$. This can be done as follows:

1) $\exists F(y = \varepsilon F)$; 0 (Assumption.)
2) $E!x$; 00 (Assumption.)
3) $x \subseteq^{\Diamond} y$; 00 (Assumption.)
4) $x \subseteq^{\Diamond} y$; 0 (By R_1^x-$x \subseteq^{\Diamond} y$)
5) $\Box\forall z\downarrow(z \in x \to z \in y)$; 0 (Def.)

we now aim to prove $\exists F(x = \varepsilon F)$ by proving $\Box\forall z(z \in x \to E!z)$ and using proposition 7.6

6) $z \in x$; 01 (Assumption.)
7) $E!z$; 01 (Assumption.)
8) $z \in x$; 0 (by R_1^x- $\in$)
9) $z \in y$; 0 (by 5)
10) $z \in y$; 01 (by R_1^x- $\in$)
11) $\Box\forall z(z \in y \to \downarrow E!z)$; 0 (by proposition 7.6 for y, and 1)
12) $\downarrow E!z$; 01 (by 10, 11)
13) $E!z$; 0 ($\downarrow$-E)
14) $\Box\forall z(z \in x \to \downarrow E!z)$; 0 (Discharge 6, 7.)
15) $\exists F(x = \varepsilon F)$; 0 (By proposition 7.6 for x)

finally, we discharge assumptions using ($\to$-I) and ($\Box$-I)

16) $\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow\exists F(x = \varepsilon F))$; 0 (Discharge 2,3)
17) $\Box(\exists F(y = \varepsilon F) \to \Box\forall x(x \subseteq^{\Diamond} y \to \downarrow\exists F(x = \varepsilon F))$; – (Discharge 1)

Now, using the lemma, proposition 7.12 and $(*)$, we can prove the result. There are two cases. One where $\neg \mathrm{Set}(y)$, and one where $\mathrm{Set}(y)$. A proof for the first case is omitted; in this case the result is vacuous. The second case can be proved as follows:

1)	$\Diamond\exists F(y = \varepsilon F)$	; –	(Definition of $\mathrm{Set}(y)$)
2)	$\exists F(y = \varepsilon F)$	; 0	(Assumption for $\Diamond$-E and $\exists$-E)
3)	$\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow E!F(x = \varepsilon F))$	; 0	(From above lemma)
4)	$\Diamond\forall FE!\downarrow\varepsilon F$	; 0	$(*)$
5)	$\forall FE!\downarrow\varepsilon F$	; 00	(Assumption for $\Diamond$-E)

we now prove $\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow E!x)$; 00

6)	$E!x$	; 000	(Assumption)
7)	$x \subseteq^{\Diamond} y$	; 000	(Assumption)

(the following few lines are to make use of the (4) axiom to shorten the label)

8)	$\Diamond(E!x \wedge x \subseteq^{\Diamond} y)$	; 00	($(\Diamond$-I))
9)	$\Diamond(E!x \wedge x \subseteq^{\Diamond} y)$	; 0	($(\Diamond$-R))
10)	$E!x \wedge x \subseteq^{\Diamond} y$	; 01	(Assumption for $\Diamond$-E)
11)	$\forall x(x \subseteq^{\Diamond} y \to \downarrow\exists F(x = \varepsilon F)$	; 01	(From 3 and $\Box$-E)
12)	$\downarrow\exists F(x = \varepsilon F)$	; 01	($\forall$-E and $\to$-E)
13)	$\exists F(x = \varepsilon F)$	; 0	($\downarrow$-E)
14)	$x = \varepsilon G$	; 0	(Assumption for $\exists$-E)
15)	$x = \downarrow\varepsilon G$	; 00	($\downarrow$-I)
16)	$E!x$	; 00	(By 5)

Now, assumptions can be discharged by $\Box$-I and $\to$-I:

17)	$\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow E!x)$	; 00	(Discharge 6,7)
18)	$\Diamond\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow E!x)$	; 0	($\Diamond$-I)
19)	$\Diamond\Box\forall x(x \subseteq^{\Diamond} y \to \downarrow E!x)$	; –	($\Diamond$-R)

Finally, assumptions 2, 5, 10 and 14 can be discharged by $\Diamond$-E and $\exists$-E

□

Bibliography

Antonelli, G. Aldo (2010a). "Notions of Invariance for Abstraction Principles". In: *Philosophia Mathematica* 18.3, pp. 276–292.

— (2010b). "Numerical Abstraction Via the Frege Quantifier". In: *Notre Dame Journal of Formal Logic* 51.2, pp. 161–179.

Benacerraf, Paul (1973). "Mathematical Truth". In: *The Journal of Philosophy* 70.19, pp. 661–679.

— (1981). "Frege: The Last Logicist". In: *Midwest Studies in Philosophy* 6.1, pp. 17–36.

Bernays, P (1976). "On the problem of schemata of infinity in axiomatic set theory". In: *Sets and Classes: On the Work by Paul Bernays*. Ed. by G. H Müller, pp. 121–172.

Blanchette, Patricia A. (1994). "Frege's Reduction". In: *History and Philosophy of Logic* 15.1, pp. 85–103.

Bolzano, B. (1817). "Rein analytischer Beweis des Lehrsatzes, dass zwischen zwei Werten, die ein entgegengesetztes Resultat gewaehren, wenigstens eine reelle Wurzel der Gleichung liege, Translation into English by S. Russ". In: *Historia Mathematica* 7, pp. 156–185.

Boolos, George (1987). "The Consistency of Frege's *Foundations of Arithmetic*". In: *On being and saying: essays in honour of Richard Cartwright*. Ed. by Judith Jarvis Thomson. MIT Press, pp. 3–20.

— (1989). "Iteration again". In: *Philosophical Topics* 17, pp. 5–21.

— (1990). "The Standard of Equality of Numbers". In: *Meaning and Method: Essays in Honor of Hilary Putnam*. Reprinted in Demopoulos, 1995. Cambridge: Cambridge University Press, pp. 261–277.

— (1999). "Is Hume's principle analytic?" In: *Logic, logic, and logic*. Ed. by Richard Jeffrey. Harvard Univ Pr, pp. 301–314.

Burge, Tyler (1975). "Truth and singular terms". In: *Noûs* 8, pp. 309–25.

— (1993). "Content Preservation". In: *Philosophical Review* 102.4, pp. 457–488.

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

Burgess, John P. and Gideon Rosen (2005). "Nominalism Reconsidered". In: *The Oxford Handbook of Philosophy of Mathematics and Logic*. Ed. by Stewart Shapiro. Oxford: Oxford University Press, pp. 515–535.

Button, Tim (2010). "Dadaism: Restrictivism as Militant Quietism". In: *Proceedings of the Aristotelian Society (Hardback)*. Vol. 110. 3pt3. Wiley Online Library, pp. 387–398.

Buzaglo, M. (2002). *The logic of concept expansion*. Cambridge Univ Press.

Carnap, Rudolf (1947). *Meaning and Necessity*. University of Chicago Press.

Cartwright, Richard L. (1994). "Speaking of Everything". In: *Noûs* 28.1, pp. 1–20.

Chalmers, David J., David Manley, and Ryan Wasserman, eds. (2009). *Metametaphysics: New essays on the foundations of ontology*. Oxford University Press.

Cook, Roy T. (2009a). "Hume's Big Brother: counting concepts and the bad company objection". In: *Synthese* 170.3, pp. 349–369.
— (2009b). "New waves on an old beach: Fregean philosophy of mathematics today". In: *New Waves in philosophy of mathematics*. Ed. by Otàvio Bueno and Øystein Linnebo. Palgrave Macmillan, pp. 13–34.
Davidson, Donald (1967). "Truth and Meaning". In: *Synthese* 17, pp. 304–323.
Demopoulos, William, ed. (1995). *Frege's Philosophy of Mathematics*. Harvard University Press.
Deutsch, Harry (2010). "Diagonalization and truth functional operators". In: *Analysis*.
Dretske, Fred (2000). "Entitlement: Epistemic Rights Without Epistemic Duties?" In: *Philosophy and Phenomenological Research* 60.3, pp. 591–606.
Dummett, Michael (1973). *Frege: Philosophy of language*. Duckworth.
— (1991). *Frege: Philosophy of mathematics*. Harvard Univ Pr.
— (1994). "What is mathematics about?" In: *Mathematics and Mind*. Ed. by Alexander George, pp. 11–26.
Ebert, Philip and Marcus Rossberg, eds. (Forthcoming). *Status Belli: Neo-Fregeans and Their Critics*.
Ebert, Philip and Stewart Shapiro (2009). "The good, the bad and the ugly". In: *Synthese* 170.3, pp. 415–441.
Field, Hartry (1980). *Science without numbers: a defence of nominalism*. Princeton University Press.
— (1989). "Realism, Mathematics and Modality". In: *Realism, Mathematics and Modality*. Oxford: Basil Blackwell.
Fine, Kit (1981). "First-Order Modal Theories I–Sets". In: *Noûs* 15.2, pp. 177–205.
— (2002). *The limits of abstraction*. Oxford University Press.
— (2005). "Our knowledge of mathematical objects". In: *Oxford studies in epistemology*. Ed. by T. Z. Gendler and J. Hawthorne. Vol. 1. Oxford: Clarendon Press, pp. 89–109.
— (2006). "Relatively Unrestricted Quantification". In: *Absolute Generality*. Ed. by Agustin Rayo and Gabriel Uzquiano. Oxford: Oxford University Press, pp. 20–44.
— (2007). "Response to Alan Weir". In: *Dialectica* 61.1, pp. 117–125.
Fitch, F. B. (1952). *Symbolic Logic*. NewYork: Ronald.
Frege, Gottlob (1879). *Begriffsschrift, a formula language, modelled upon that of arithmetic, for pure thought*.
— (1884). *The Foundations of Arithmetic*. translated J.L. Austin. Northwestern University Press.
— (1893). *The Basic Laws of Arithmetic*. Translated Montgomery Furth. University of California Press.
— (1948). "Sense and Reference". In: *The Philosophical Review* 57.3, pp. 209–230.
George, Benjamin (2006). "Second-Order Characterizable Cardinals and Ordinals". In: *Studia Logica* 84 (3). 10.1007/s11225-006-9016-7, pp. 425–449.
Glanzberg, Michael (2001). "The liar in context". In: *Philosophical Studies* 103, pp. 217–251.
— (2004). "Quantification and Realism". In: *Philosophy and Phenomenological Research* LXIX.3.
— (2006). "Context and unrestricted quantification". In: *Absolute Generality*. Ed. by Agustin Rayo and Gabriel Uzquiano. Oxford: Oxford University Press, pp. 45–74.
Goldman, Alvin (1979). "What is Justified Belief?" In: *Justification and Knowledge*. Ed. by George Pappas. Boston: D. Reidel, pp. 1–25.
Hale, Bob (1987). *Abstract Objects*. Blackwell.
— (1999). "Intuition and Reflection in Arithmetic II". In: *Proceedings of the Aristotelian Society, Supplementary Volumes* 73, pp. 74–98.

— (2000a). "Abstraction and set theory". In: *Notre Dame Journal of Formal Logic* 41.4, pp. 379–398.
— (2000b). "Reals by abstraction". In: *Philosophia Mathematica* 8.3, pp. 100–123.
— (2001a). "A Response to Potter and Smiley: Abstraction by Recarving". In: *Proceedings of the Aristotelian Society*. New Series 101, pp. 339–358.
— (2001b). "Singular Terms (1)". In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 31–47.
— (2001c). "Singular Terms (2)". In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 48–71.
— (2005). "Real Numbers and Set theory–Extending the Neo-Fregean Programme Beyond Arithmetic". In: *Synthese* 147.1, pp. 21–41.
— (2006). "Kit Fine on the Limits of Abstraction". In: *Travaux de Logique* 18. Ed. by P. Joray.
Hale, Bob and Crispin Wright (2000). "Implicit definition and the a priori". In: *New essays on the a priori*. Reprinted in Hale and Wright, 2001a, pp. 286–319.
— (2001a). *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press.
— (2001b). "To Bury Caesar…" In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 335–98.
— (2003). "Responses to commentators". In: *Philosophical Books* 44.3, pp. 245–263.
— (2009a). "Focus restored: Comments on John MacFarlane". In: *Synthese* 170.3, pp. 457–482.
— (2009b). "The metaontology of abstraction". In: *Metametaphysics: New essays on the foundations of ontology*. Ed. by David J. Chalmers, David Manley, and Ryan Wasserman. Oxford University Press, pp. 178–212.
Hamkins, Joel David (Aug. 2011). "The set-theoretic multiverse". In: *ArXiv e-prints*. arXiv:1108.4223 [math.LO].
Hazen, Allen (1985). "Review of Wright, 1983". In: *Australasian Journal of Philosophy* 63.2, pp. 250–254.
— (1990). "Actuality and Quantification". In: *Notre Dame Journal of Formal Logic* 31.4.
Heck Jr., Richard G. (1992). "On the consistency of second-order contextual definitions". In: *Noûs* 26, pp. 491–4.
— (1993). "The Development of Arithmetic in Frege's *Grudgesetze der Arithmetik*". In: *The journal of symbolic logic* 58. Reprinted in Demopoulos, 1995, pp. 579–601.
Hellman, Geoffrey (2006). "Against 'Absolutely Everything'!" In: *Absolute Generality*. Ed. by Agustin Rayo and Gabriel Uzquiano. Oxford: Oxford University Press, pp. 75–97.
Hirsch, Eli (2002). "Quantifier variance and realism". In: *Philosophical Issues* 12, pp. 51–73.
— (2009). "Ontology and alternative languages". In: *Metametaphysics: New essays on the foundations of ontology*. Ed. by David J. Chalmers, David Manley, and Ryan Wasserman. Oxford University Press, pp. 231–259.
Hodes, Harold (1984a). "Logicism and the Ontological Commitments of Arithmetic". In: *Journal of Philosophy* 81.3, pp. 123–149.
— (1984b). "On Modal Logics Which Enrich First-Order S5". In: *Journal of Philosophical Logic* 13, pp. 423–454.
Lavine, Shaughan (2006). "Something about everything: Universal quantification in the universal sense of universal quantification". In: *Absolute Generality*. Ed. by Agustin Rayo and Gabriel Uzquiano. Oxford: Oxford University Press.
Lear, Jonathan (1977). "Sets and Semantics". In: *The Journal of Philosophy* 74.2, pp. 86–102.

Leitgeb, Hannes (Forthcoming). "Abstraction Grounded: A Note on Abstraction and Truth". In: *Status Belli: Neo-Fregeans and Their Critics*. Ed. by Philip Ebert and Marcus Rossberg.

Linnebo, Øystein (2010a). "Pluralities and sets". In: *The Journal of philosophy* 107.3, pp. 144–164.

— (2010b). "Some Criteria for Acceptable Abstraction". In: *Notre Dame Journal of Formal Logic* 52.3, pp. 331–338.

— (ms). "The Potential Hierarchy of Sets". In:

Linnebo, Øystein and Gabriel Uzquiano (2009). "Which Abstraction Principles are Acceptable? Some Limitative Results". In: *The British journal for the philosophy of science* 60.2, pp. 239–252.

MacFarlane, John (2002). "Frege, Kant, and the Logic in Logicism". In: *The Philosophical Review* 111.1, pp. 25–65.

McGee, Vann (1997). "How we learn mathematical language". In: *The Philosophical Review* 106.1, pp. 35–68.

Milne, Peter (1986). "Frege's Context Principle". In: *Mind* 95.380, pp. 491–495.

Montague, R (1963). "Syntactic treatments of modality, with corollaries on reflexion principles and finite axiomatizability". In: *Acta Philosophica Fennica* 16, pp. 153–67.

Parsons, Charles (1974). "The Liar Paradox". In: *Journal of Philosophical Logic* 3. Reprinted in Parsons, 1983a, pp. 221–267, pp. 381–412.

— (1983a). *Mathematics in Philosophy*. Cornell University Press.

— (1983b). "Sets and modality". In: *Mathematics in Philosophy*. Cornell University Press, pp. 298–341.

Peacocke, Christopher (2004). *The Realm of Reason*. Oxford University Press.

Pederson, Nikolaj (2011). *Hume's Principle and entitlement: On the epistemology of the neo-Fregean programme*. to appear in Ebert and Rossberg (Forthcoming).

Potter, Michael (2010). "Abstractionist Class Theory : Is There Any Such Thing?" In: *The Force of Argument: Essays in Honor of Timothy Smiley*. Ed. by Jonathan Lear and Alex Oliver. Routledge.

Potter, Michael and Timothy Smiley (2001). "Abstraction by Recarving". In: *Proceedings of the Aristotelian Society*. New Series 101, pp. 327–338.

Potter, Michael and Peter Sullivan (2005). "What Is Wrong with Abstraction?" In: *Philosophia Mathematica* 13.2, pp. 187–193.

Prawitz, Dag (1965). *Natural Deduction: A proof theoretic study*. Mineola, New York: Dover.

Prior, Arthur N. (1960). "The Runabout Inference-Ticket". In: *Analysis* 21.2, pp. 38–39.

Quine, W. V. (1950). "Identity, Ostension, and Hypostasis". In: *Journal of Philosophy* 47.22, pp. 621–633.

— (1951). "Two Dogmas of Empiricism". In: *The Philosophical Review* 60.1, pp. 20–43.

Rayo, Agustin and Gabriel Uzquiano, eds. (2006). *Absolute Generality*. Oxford: Oxford University Press.

Rumfitt, Ian (2003). "Singular terms and arithmetical logicism". In: *Philosophical Books* 44.3, pp. 193–219.

Russell, Bertrand (1901). "Letter to Frege". In: *From Frege to Gödel: a source book in mathematical logic, 1879-1931*. Ed. by J. Van Heijenoort. Harvard Univ Pr.

Shapiro, Stewart (1991). *Foundations without foundationalism: a case for second order logic*. Oxford: Clarendon Press.

— (2000). "Frege meets dedekind: a neologicist treatment of real analysis". In: *Notre Dame Journal of Formal Logic* 41.4, pp. 335–364.

— (2003). "Prolegomenon to any future neo-logicist set theory: abstraction and indefinite extensibility". In: *The British Journal for the Philosophy of Science* 54.1, pp. 59–91.
Shapiro, Stewart and Alan Weir (1999). "New V, ZF and abstraction". In: *Philosophia Mathematica* 7.3, pp. 293–321.
— (2000). "'Neo-Logicist' Logic is not Epistemically Innocent". In: *Philosophia Mathematica* 8.2, pp. 160–189.
Shapiro, Stewart and Crispin Wright (2006). "All Things Indefinitely Extensible". In: *Absolute Generality*. Ed. by Agustin Rayo and Gabriel Uzquiano. Oxford: Oxford University Press, pp. 255–304.
Sider, Theodore (2007). "NeoFregeanism and Quantifier Variance". In: *Aristotelian Society, Supplementary Volume* 81, pp. 201–32.
— (2009). "Ontological Realism". In: *Metametaphysics: New essays on the foundations of ontology*. Ed. by David J. Chalmers, David Manley, and Ryan Wasserman. Oxford University Press, pp. 384–423.
Siemens, David F. (1977). "Fitch-style rules for many modal logics." In: *Notre Dame Journal of Formal Logic* 18.4, pp. 631–636.
Studd, James (2012). "The iterative conception of set: A (bi-)modal axiomatisation". In: *Journal of Philosophical Logic*.
Tait, W.W. (1981). "Finitism". In: *The Journal of Philosophy* 78.9, pp. 524–546.
Turner, Jason (2010). "Ontological Pluralism". In: *Journal of Philosophy* 107.1, pp. 5–34.
Uzquiano, G. (2009). "Bad company generalized". In: *Synthese* 170.3, pp. 331–347.
Weir, Alan (2003). "Neo-Fregeanism: An Embarrassment of Riches". In: *Notre Dame Journal of Formal Logic* 44.1, pp. 13–48.
— (2007). "Honest Toil or Sheer Magic?" In: *dialectica* 61.1, pp. 89–115.
Williamson, Timothy (2003). "Everything". In: *Philosophical Perspectives* 17.1, pp. 415–465.
— (forthcoming). "Barcan Formulas in Second-Order Modal Logic". In: *Themes from Barcan Marcus*. Lauener Library of Analytical Philosophy, vol. 3.
Wright, Crispin (1983). *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.
— (1997). "On the Philosophical Significance of Frege's Theorem". In: *Language, Thought, and Logic: Essays in Honour of Michael Dummett*. Ed. by R. Heck. Oxford University Press, pp. 201–44.
— (2001a). "Is Hume's principle analytic?" In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 307–332.
— (2001b). "On the Harmless Impredicativity of $N^{=}$ (Hume's Principle)". In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 229–256.
— (2001c). "Response to Dummett". In: *The Reason's Proper Study: Essays toward a Neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon Press, pp. 256–271.
— (2004a). "Intuition, Entitlement and the Epistemology of Logical Laws". In: *Dialectica* 58.1, pp. 155–175.
— (2004b). "Warrant for nothing (and foundations for free)?" In: *Aristotelian Society Supplementary Volume* 78.1, pp. 167–212.
Zalta, Edward N. (2010). "Frege's Logic, Theorem, and Foundations for Arithmetic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2010. URL: `http://plato.stanford.edu/archives/fall2010/entries/frege-logic/`.