Notre Dame Journal of Formal Logic
Volume 44, Number 1, 2003

Neo-Fregeanism: An Embarrassment of Riches

Alan Weir

Abstract Neo-Fregeans argue that substantial mathematics can be derived from a priori abstraction principles, Hume's Principle connecting numerical identities with one:one correspondences being a prominent example. The embarrassment of riches objection is that there is a plurality of consistent but pairwise inconsistent abstraction principles, thus not all consistent abstractions can be true. This paper considers and criticizes various further criteria on acceptable abstractions proposed by Wright settling on another one—stability—as the best bet for neo-Fregeans. However, an analogue of the embarrassment of riches objection resurfaces in the metatheory and I conclude by arguing that the neo-Fregean program, at least insofar as it includes a platonistic ontology, is fatally wounded by it.

1 Introduction

In the last decade or two there has been a revival of interest in logicism recast not as the doctrine that mathematics *is* logic but rather as the claim mathematical truths have something like the status assigned to them by the logicists. The *neologicist* contention is that mathematical truths are known, where they are, neither by some mysterious form of direct intuition nor by empirical confirmation, even of an indirect and holistic fashion via the scientific theories they contribute to. Rather mathematical knowledge arises on the basis solely of the understanding of the basic mathematical and logical concepts which anyone who grasps the mathematical truths has. This view might be interpreted as saying that mathematical truths are analytic, are true by virtue of meaning, similarly that fundamental mathematical inference rules are meaning-constitutive. Since the notion of analyticity is still under a cloud, in some quarters, a more broadly acceptable goal for the neologicist might be to try to establish that mathematical axioms are implicit definitions since, prima facie, anyway, this does not commit one to the notion of analyticity; this, indeed, is the direction which recent work has taken (see Hale and Wright [11]).

Received February 26, 2002; accepted March 3, 2003; printed April 16, 2004
2000 Mathematics Subject Classification: Primary, 03A05; Secondary, 03E65, 03E70
Keywords: neologicism, Frege, abstraction

Having "something like the status assigned them by the logicists" is a vague notion, one which the neologicists need to clarify if their view is to be assessed fully. But I take it to be clear enough to be going on with. In particular, if it could established that mathematical truths, or even just a substantial proportion of them, are known by a process different from that proposed by the neo-Kantian platonists or by Quinean empiricists, a process whose materials essentially involve only appeal to grasp of mathematical language, then this would constitute a major advance in the epistemology of mathematics. It would establish the "epistemic innocence" (cf. Shapiro and Weir [26], p. 160) of mathematics. So in what follows I will assume that the idea of "epistemic innocence" is clear enough for a fruitful debate with the neologicist to take place, while noting that the neologicist owes the wider philosophical community a fuller account of what it amounts to.

The main proponents of this program, philosophers such as Wright and Hale,[1] take it that they are continuing and developing a program initiated by Frege and so characterize the program as *neo-Fregean* as well as *neologicist*. Neo-Fregeans also want to uphold Frege's platonism at least to the extent of holding that truth in pure mathematics is as objective as truth in the empirical sciences, however exactly one wishes to analyze the notion of objectivity in the sciences. They reject, moreover, any form of relativism in mathematics (cf. Wright [36], p. 293). They also reject the idea that there is a plurality of mathematical domains—of different set theories, or different domains of sets, numbers, and categories, and so forth—which cannot all be accumulated into a single mega-universe.

Clearly, the neo-Fregean position is a highly attractive one for anyone sympathetic to the traditional view that mathematics is a system of objective truths knowable a priori but who is also sensitive to the usual epistemological problems raised against platonistic mathematics, most notably the puzzle as to how we could gain knowledge of a world of causally inert abstract objects. But how can grasp of mathematical language yield knowledge of the existence of a rich realm of abstract entities which exist independently of our language or conceptual system? To try to explain this, the neo-Fregean focuses on abstraction principles. Abstraction principles are principles of the form[2]

$$\alpha x(\varphi x) = \alpha x(\psi x) \leftrightarrow \Xi(\varphi, \psi)$$

where α is some term-forming variable-binding operator which forms singular terms from open sentences and Ξ is an equivalence relation over properties. One key example is Hume's Principle (HP):[3]

$$\forall X \forall Y((nxXx = nxYx) \leftrightarrow X \text{ 1-1 } Y)$$

with X 1-1 Y the second-order sentence which expresses the existence of a one-one correspondence between the Xs and the Ys. An important impetus to the revival of neo-Fregeanism has been the detailed sketch, by Wright [35], of what has become known as Frege's Theorem (cf. Boolos [2], p. 209; [36], p. 273)—the derivability of second-order arithmetic from second-order logic plus Hume's Principle. For example, from this principle one can derive in standard second-order logic[4] a theory even stronger than (though equiconsistent with) the usual Peano-Dedekind formulation of second-order arithmetic.[5]

The more general neo-Fregeanism goal, following on from this result, is to show that there is an abstraction principle, or set of principles, A such that A plus second-order logic yields all mathematical truths, or at any rate all those truths we need to

do empirical science and metamathematics. In addition to the formal goal, the neo-Fregean seeks to convince us that both the abstraction principle A and second-order logic are epistemically innocent. One way to cash this out would be to claim that anyone who grasped a proof of a mathematical result R from A and second-order logic is thereby in a position to know that R is true in something like an a priori fashion.[6] This means that our mathematician must be able to know a priori (or in an epistemically innocent way) the truth of A, and similarly know in some innocent fashion the truth of the axioms and soundness of the inference rules used in the derivation.

It is with the abstraction principles, and the problems which arise if one holds both that they are objectively true and that they are epistemically innocent, that this paper concerns itself. There are, of course, a number of other steep challenges which neo-Fregeanism faces, for example, the anti-Anselmian insistence that one cannot prove objective existence claims a priori, or the challenge of showing that the standard second-order logic used in the derivation of Frege's Theorem and presumably in any stronger mathematical results is a system of epistemically innocent truths. For this logic is a classical "nonfree" logic which includes the full impredicative axiom schema of comprehension whereby the second-order quantifiers are interpreted as including in their range every subset S of the domain of individuals (or include a property for each such extension).[7] However, in this paper I will focus solely on the innocence or guilt of the abstraction principles which the neo-Fregean appeals to.

The structure of the paper is as follows. Section 2 sets out the main objection I will be concerned with, the *Embarrassment of Riches* (ER) objection, while Section 3 looks at Wright's first main response, the appeal to conservativeness principles. Section 4 shows that this response on its own is inadequate while Section 5 looks at a couple of additional criteria that Wright sketches, arguing that they too will not do. In Section 6 I develop a third notion into a stronger criteria, *stability*, which seems to provide the best hope for the neo-Fregean of providing an answer to the embarrassment of riches objection. In Section 7, though, I argue that an analogue of the ER objection simply recurs at a metatheoretic level. In the final section, Section 8, I argue that this shows that the full neo-Fregean program is fatally flawed but that there may be less "Fregean" variants which can survive these objections.

2 Embarrassment of Riches

Since my focus is on the abstraction principles, not the logic, I shall assume for the purposes of the present argument that a priori existence proofs cannot be ruled out of court, that there are analytic or meaning-constitutive, or more broadly, perhaps, epistemically innocent, principles or rules and that the system of such principles and rules includes standard second-order logic. The embarrassment of riches objection is to the effect that more principles than can possibly all hold true together can be validated by the neo-Fregean methods.[8]

I will introduce the ER objections by starting with a related one which Wright calls the "Bad Company" objection, one raised by Boolos, Dummett, and Field: the objection is that Hume's Principle is formally very similar to the naïve rules for class, embodied, in one form, in Frege's notorious Axiom V:[9]

$$\forall X \forall Y (\{x : Xx\} = \{x : Yx\} \leftrightarrow \forall z (Xz \leftrightarrow Yz)).$$

If the former is analytic, so is the latter. But since the latter is inconsistent, it cannot, surely, be analytic, hence neither is Hume's Principle, nor any similar abstraction principle.

As it stands, this Bad Company objection is not all that strong. One possible response is to deny that Axiom V is inconsistent (or at least deny that the theory of Axiom V is trivial). This would involve fairly extensive restriction of classical logic, of course: Axiom V is inconsistent not only in intuitionistic logic but also in relevant logics such as T or RWX, these being weaker than the well-known E and R. However, one increasingly common strategy with nonclassical logics is to so set up the division between operational rules (e.g., in natural deduction systems introduction and elimination rules for the logical constants) and structural rules that the operational rules yield classical logic given classical structural rules but then block antinomy by allowing classical structural rules only in special cases (ideally cases involving all of standard mathematics). Wright, indeed, shows some sympathy with some fairly heterodox lines of thought by entertaining seriously the possibility of rejecting the applicability of Cantor's Powerset Theorem to the domains of interesting abstraction theories ([36], p. 294).[10] (Cantor himself thought the theorem did not apply to "inconsistent multiplicities").

Still, accepting as true unadulterated Axiom V is a very radical response to take. But one need not be so radical in order to respond effectively to Bad Company. For even if one accepts that Axiom V is trivially inconsistent, though formally speaking an abstraction principle with the same overall structure as Hume's Principle, this still does not tell very heavily against neo-Fregeanism. The neo-Fregean can deny that Axiom V is epistemically innocent simply by laying down consistency as a criterion on epistemic innocence so still affirming that Hume's Principle is innocent. Since hidden inconsistency could lurk in many other abstraction principles, the neo-Fregean will have to concede that analyticity or epistemic innocence is not a purely formal matter, nor a decidable one (cf. [36], p. 213, fn. 27). But this is a plausible position to adopt on independent grounds: analytic rules, in that sense, need not be *transparently* analytic to those who follow them. After all, the neo-Fregean will want to hold that indefinitely many, currently undecided, mathematical theses are a priori true, even though it may take a genius to come up with a proof of some of them.

There is, however, a related but far stronger point: there are indefinitely many consistent *but pairwise inconsistent* abstraction principles. If all consistent analytic principles are analytic, then both of two such principles are analytic and presumably true which is absurd.[11] This style of objection is what I mean by the Embarrassment of Riches or ER objection.

This point is made by Heck [14] utilizing abstraction principles of the form

$$\forall X \forall Y (\alpha x Xx = \alpha x Yx \leftrightarrow (P \vee \forall x (Xx \leftrightarrow Yx)))$$

where P contains no occurrences of the α abstraction operator. This principle is satisfiable if and only if P is.[12] Hence for incompatible values of P (e.g., P_0 = the universe is of size $\aleph_0$ versus P_1 = the universe is of size $\aleph_1$) we get satisfiable but incompatible principles, indeed, provably incompatible principles where, as in the two examples just given, we have two principles P_i, P_j such that $P_i, P_j \vdash \bot$.

A case which will be of particular interest in what follows occurs when P takes the form $\mathrm{Bad}(X)$ & $\mathrm{Bad}(Y)$ where Badness is a second-order property of properties for which equinumerosity is a congruence (a cardinality property). We then get

disjunctivized generalizations of Axiom V of the form

$$\forall X \forall Y (\{x : Xx\} = \{x : Yx\} \leftrightarrow ((\text{Bad}(X) \,\&\, \text{Bad}(Y)) \vee \forall x (Xx \leftrightarrow Yx))).$$

I will call disjunctivized Axiom V principles of this type *distraction principles*. For example, as instances of Bad we could choose finite, infinite, uncountably infinite, or Big, where a property is Big if and only if there is a function from it onto the universe. This latter version is in fact Boolos's *New V* [1] and the idea of distraction principles is simply a generalization of his notion. Further instantiations of the schematic 'Bad' include properties such as being of size at least $\aleph_n$, or at least $\beth_n$, or at least θ_n, where θ_n is the nth-inaccessible cardinal (here n is finite). We could also set exact cardinality limits on Bad, for example, countably infinite or exactly $\aleph_n$ / $\beth_n$ / θ_n, or weaken these clauses to at most $\aleph_n$ / $\beth_n$ / θ_n and so on. All these concepts and more are definable in second-order logic (Garland [9]).

The set theories which result are interesting in that they embody a limitation of size principle, widely seen as a non-ad hoc method of restricting naïve set theory and avoiding paradox. If two properties are the same size then either both are Bad, or both are Good (=∼Bad). Define a set to be the extension of a Good Property:

$$\text{Set } x \text{ iff } \exists X (\text{Good}(X) \,\&\, x = \{x : Xx\}).$$

Then it is easy, using the definition of $\in$ by

$$x \in y \text{ iff } \exists X (y = \{x : Xx\} \,\&\, Xx),$$

to prove from the abstraction principle in question a comprehension principle for sets (equivalently Good properties):

$$\forall X (\text{Good } X \rightarrow \forall x (x \in \{x : Xx\} \leftrightarrow Xx)),$$

and it will follow that if the extension of X is a set and X is equinumerous with Y then Y too determines a set as its extension.

However Heck's problem arises even in this particular case—there are incompatible distractions. In particular, let φ and ψ be cardinality properties which are

(a) provably incompatible in that $\vdash \sim \exists X (\varphi(X) \,\&\, \psi(X))$,[13] and
(b) such that both are provably properties for which equinumerosity is a congruence, that is, $\vdash \forall X, Y ((\varphi X \,\&\, X \text{ 1-1 } Y) \rightarrow \varphi Y)$, likewise for ψ.

Examples of a pair of properties which satisfy both (a) and (b) are the pair 'exactly of size $\aleph_0$' and 'at least of size $\aleph_1$'.

Consider then two distraction principles, D_1, in which the badness property is $\text{Bad}_1 = (\text{Big} \,\&\, \varphi)$, and D_2, in which $\text{Bad}_2 = (\text{Big} \,\&\, \psi)$, with φ and ψ as in (a) and (b).

Theorem 2.1 $D_1, D_2 \vdash \bot$.

Proof Note first that for any distraction principle D_i we have $\exists X (\text{Bad}_i(X))$ for the badness property featuring in that principle. The argument is a reductio: if not the principle collapses into Axiom V. So by existential instantiations (or by assumptions for existential elimination) from the two such existential generalizations derivable from D_1 and D_2, we have

$$\vdash \text{Big}(F) \,\&\, \varphi(F) \,\&\, \text{Big}(G) \,\&\, \psi(G). \tag{1}$$

From (1) we get $\vdash F$ 1-1 G by composition of functions, hence, from (1) again together with property (b) above we get, $\vdash \varphi(F)$ & $\psi(F)$ contradicting (a) above. □

So the neo-Fregean needs to discriminate further among the distraction principles. One further constraint might be that such principles must not only be proof-theoretically consistent (i.e., we do not have $D \vdash \bot$ for the distraction principle D) but satisfiable in a full standard second-order model. But here again it is easy to produce incompatible principles each of which is satisfiable, for example, D^{Fin} with Bad as (Dedekind) finite versus D^{Inf} with Bad as (Dedekind) infinite. Both of these principles have models. For models for distraction principles exist if and only if we can biject the class of good properties into a proper subset Sets of the domain of individuals D_0, the range being the good classes, or sets; all the bad properties we map into a *bad guy*, a dummy *proper class* object, $\spadesuit \in D_0-$Sets. (For simplicity and with no commitment to a nominalistic metaphysics, I will identify properties over a domain D of individuals with subsets of D.)

On various assumptions, we can find models for all the variants for Bad given above. It is straightforward to show that all and only the (nonempty) finite models satisfy Bad as finite, though the models are a degenerate case in which there are no sets, no good classes, and all properties are mapped to the one dummy object $\spadesuit$.[14] Using ZFC we can show that Bad as infinite has models in all and only the infinite cardinalities because the number of good, that is, finite, subsets of a universe of cardinality $\aleph_\alpha$ is just $\aleph_\alpha$. So we select an $\aleph_\alpha$-sized proper subset of D_0, biject all finite properties into it and the rest to $\spadesuit \in D_0-$Sets. But clearly no standard model can satisfy both these principles simultaneously.

And of course there are other abstraction principles which hold only in infinite models and so are incompatible with D^{Fin}, for instance, Boolos's New V [1] in which Bad is Big. This has models at $\aleph_0$ and also at all *nearly strong* inaccessible cardinalities,[15] if we add to ZFC the assumption that such cardinals exist.

Theorem 2.2 *The number of smaller subsets of a set S of nearly strong inaccessible cardinality θ is just θ.*

Proof Without loss of generality we can consider the cardinal θ itself rather than S. Since θ is regular, every small subset of θ is also a subset of some $\lambda < \theta$. There are at most $2^\lambda \leq \theta$ subsets of each such λ and there are exactly θ such cardinals λ. Thus there are at most $\theta \times \theta = \theta$ small subsets of θ (and obviously there are at least that many). □

So this time we select a θ-sized proper subset of the domain of individuals and biject the properties of cardinality $< \theta$ onto it and the rest to the dummy proper class. By similar techniques, ZFC plus the axiom of inaccessibles proves the existence of models for abstraction principles with Bad as 'Inaccessible', Bad as 'at least/exactly the size of the *n*th-inaccessible' and so on. These models are particularly interesting, of course, since, Sets being of inaccessible size, we get a ZF-ish theory (which can be derived from the abstraction principle given a standard proof theory.)[16] We do not in general get the Axiom of Choice but we do get it for some instantiations of Bad, for example, as 'exactly size α' or with Bad = Big as in the New V distraction principle (see [26], Section 3). Neither do we get foundation. To see this, (working in ZFC) partition Sets into two disjoint θ-sized sets S_1 and S_2 such that our bijection of the

properties into the individuals maps $\{\alpha\}$ to α, for all $\alpha \in S_1$ and maps the remaining small properties into S_2. Then there will exist a Big, universe-sized subdomain of Sets in which each member is equal to its unit set; that is, the term $\{x : x = y\}$, in an assignment to variables in which α is assigned to y, is itself assigned α since $\lambda x(x = y)$ is interpreted by $\{\alpha\}$. Nonetheless we can define the well-founded classes (relative to $\varnothing$ as base) by

$$\mathrm{WC}x \equiv_{\mathrm{def}} \forall X((X\varnothing \,\&\, \forall y(\forall z(z \in y \rightarrow Xz) \rightarrow Xy)) \rightarrow Xx)$$

(i.e., the inductive closure from the empty set under the membership operation) and prove from this the WC classes form a well-founded hierarchy. So in these theories with Bad as exactly the nth inaccessible we can get, by restricting to well-founded sets, ZFC (and a bit more, for $n > 1$).

Overall then, the class of Distraction principles yields a rich and interesting set of theories, an important subclass of abstraction principles. The problem for the neo-Fregean is precisely that it is too rich, that we have an Embarrassment of Riches. Even on fairly weak assumptions, there are incompatible distraction principles (e.g., Bad as finite versus Bad as infinite) such that both are satisfiable; if both can be known in epistemically innocent fashion both of them must be true, which is absurd.

Commenting on Heck's examples of consistent but pairwise inconsistent abstractions, Boolos asserts forthrightly:

> His article seems to me to do in, once and for all, the idea that "contextual definitions" like Hume's principle or Basic Law V, have, in general, any privileged logical status. (Boolos [3], p. 231)

3 Conservativism

Wright, however, does not accept that he has been "done in." He considers principles which, like D^{Fin}, are true in only finite models but he finds the incompatibility of Hume's Principle, for example, with such principles no more worrying than the formal resemblance all abstraction principles have to Axiom V ([36], pp. 295–97). But given the satisfiability of both principles, is this attitude justified? Have we not here pairs of principles each with an equal title to be classed as epistemically innocent but such that at least one must be false. Certainly neo-Fregeans cannot withhold the title of analytic or innocent from D^{Fin} on the grounds that they know via intuitive acquaintance with the world of mathematical objects that this world, hence the universe as a whole, is infinite. Nor can they rule out D^{Fin} on the grounds that as a basis for empirical science it appears to be somewhat unfruitful, to say the least. If appeal to intuition or pragmatic utility is allowed to determine which abstraction principles are legitimate and which are not then neo-Fregeans can have no principled objection to the mathematical epistemology of the Kantians or the empiricist epistemology of Quine or Putnam; and if that type of empiricist epistemology is acceptable then the use of abstraction principles, rather than, say, axiom systems such as ZFC, in the development of mathematics would be largely a matter of taste and convenience.

We must remember in this connection that mathematics is not neutral with respect to logic, certainly not with respect to second-order logic at any rate.[17] Thus $\mathrm{CH} \models\!\!\!\!\!\perp$ holds *for anyone who defines* $\models$ *set-theoretically* and who believes that the second-order formulation CH of the continuum hypothesis is false. Similarly the implication fails for those who hold CH to be true. For the former theorist, the continuum hypothesis is not a *logical* possibility in the semantic sense, where this somewhat

obscure notion is precisified model-theoretically. It is not a logical possibility since a set-theoretic universe in which the continuum hypothesis holds is mathematically unavailable on this view, hence presumably mathematically impossible, though the statement that the continuum hypothesis holds is consistent proof-theoretically, if ZF is (given standard, finitistic notions of proof). Similarly WO $\models \bot$ fails for a platonist who accepts WO, but holds for any platonist who accepts the axiom of determinacy, which entails the falsity of the well-ordering theorem WO, and so forth. For this latter theorist, mathematics rules out the existence of a structure which represents the logical possibility of WO. Since logical consequence is such a rich and structurally complex notion, it is inevitable that any position which moves beyond blind acceptance of some system of primitive rules will have to use mathematics in the investigation of the properties of logical systems, whether currently favored or disfavored. There can be no neutral, nonaligned mathematics, as far as logic is concerned.

Moreover, mathematics is no less fundamental than logic—for the neo-Fregean. It does not seem to make much sense to say that the principles of logic are *more* meaning-constitutive, more analytic, more epistemically innocent, than mathematical principles applicable to term-forming, rather than sentential, operators. How then can the neo-Fregeans rule out laying down that all structures, hence all empirical structures, must be finite? Why is that illegitimate, if it is legitimate to require that other mathematically impossible "structures" which can be specified without proof-theoretic inconsistency be ignored?

Wright, however, has a general and principled objection to principles such as D^{Fin} and the whole range of abstraction principles in which Bad takes the form 'exactly size α' or 'at most size β'. It is that such principles are *nonconservative* because they place upper bounds on the size of the total universe of individuals and hence on the range of acceptable empirical theories and this is something which no genuine mathematical theory can do, granted that mathematics is a priori. So the telling objection to D^{Fin} is not that it is useless for empirical science but that it has a property which no theory, not even a theory which happens in the actual world to be empirically fruitful, can have if it is a genuinely mathematical theory. This is Wright's primary response to the Embarrassment of Riches objection.

To evaluate this response we have to look at conservativeness more closely. The formulation above is too loose of course: mathematics does place constraints on acceptable empirical theories: it rules out theories which say that the number of spacetime regions is both continuum-sized and of size $\aleph_0$, for example. Rather the idea is that mathematical theory should be compatible with any natural possibility; otherwise we would need to know, presumably a posteriori, that the physical world is not structured in one of the possible ways which are inconsistent with mathematics in order to know that mathematics is actually true. And that would conflict with the a priori status of mathematical truth.[18] Hence adding a true mathematical theory to an empirical theory T should not enable us to prove any more physical conjectures than those which already followed from T. If T does not entail C, if there is a possibility of T being true and C false, then mathematics should not conflict with that possibility. So there should likewise be a possibility of T holding together with any body of mathematical truths and yet C still being false. Perhaps, then, if we admit only conservative abstraction principles the set of such conservative principles

will be a consistent set and we will knock out at least one of each of the warring pairs of consistent but pairwise inconsistent abstraction principles.

There are a number of natural ways to characterize the above notion of conservativeness, all of them making use of relativizations of well-formed formulas (wffs) by formulas. Let P^A represent the result of restricting quantifiers in the wffs P by a formula in one free individual variable Ax. For example,

1. $P^A = P$ for atomic P;
2. the relativization transformation distributes over the sentential operators; and
3. (a) $(\forall y\varphi)^A = \forall y(Ay \rightarrow \varphi^A)$;
   (b) $(\exists y\varphi)^A = \exists y(Ay \,\&\, \varphi^A)$;
   (c) $(\forall X\varphi)^A = \forall X(\forall y(Xy \rightarrow Ay) \rightarrow \varphi^A)$;
   (d) $(\exists X\varphi)^A = \exists X(\forall y(Xy \rightarrow Ay) \,\&\, \varphi^A)$.

In the discussion of conservativeness which follows, I will assume we start from a language $\mathcal{L}$ which we expand to $\mathcal{L}^+$ by the addition of class abstracts—for example, $\mathcal{L}^+$ is closed under the operation of applying class brackets to wffs φx to form new singular terms $\{x : \varphi x\}$. Since I will be considering abstraction principles which yield a theory of classes sufficient to define ordered pairs subject to the law of ordered pairs,

$$(\langle x, y\rangle = \langle w, z\rangle \leftrightarrow (x = w \,\&\, y = z)),$$

we can consider only monadic second-order logic, with relations represented by properties of ordered pairs.

Probably the most natural form of conservativeness principle is the type utilized by Field (see [5], pp. 96–97, fn. 21; pp. 125–26). A syntactic version of his criterion is

> Let T be a theory in $\mathcal{L}$ and A an abstractionist theory in $\mathcal{L}^+$. T, A *need not be consistent*. Let $\sim Mx$ be $\sim\exists X(x = \{x : Xx\})$ so that the extension of Mx comprises the abstracts of theory A. Then if $T^{\sim M}$, $A \vdash C^{\sim M}$, then $T \vdash C$.

Thus we relativize our theory and consequence to the $\sim M$s, the nonmathematical, *concrete* subuniverse. Replace $\vdash$ by $\models$ and we get the semantic version of Field's criterion.

However, Wright had initially utilized the following conservativeness principle:

> Let θ be any theory with which Σ is consistent. Then Σ is conservative with respect to θ just in case, for any T expressible in the language of θ, $\theta \cup \{\Sigma\}$ entails the Σ-restriction of T only if θ entails T. ([36], p. 297, fn. 49)

(A Σ-restricted formula restricts the first-order quantifiers, in the intended interpretation, to the members of the domain of individuals which are not referents of the abstraction terms.) However, let θ be

> If there is an infinite property then Clinton is not an adulterer.

Hume's Principle plus θ deductively and hence also semantically entails that Clinton is not an adulterer but θ does not entail deductively or semantically this on its own (even if Clinton himself thinks that 'Clinton is not an adulterer' is true by virtue of meaning alone). And 'Clinton is not an adulterer' is the Σ-restriction of 'Clinton is not an adulterer'.[19] Thus even the finite version of Hume's Principle, in which the

initial second-order universal quantifiers are restricted to finite properties, is not conservative on Wright's criterion. For finite Hume, like HP, also entails that there are infinitely many individuals in the domain of the first-order quantifiers (Heck [16]). For this sort of reason, Wright abandons his initial notion of conservativeness in favor of a Field-type criterion.[20]

I will look at one further group of conservativeness principles which arise by letting the restriction predicate be a simple unary predicate Ex. The criteria are then the same as Field's except that we relativize to E rather than to $\sim\exists X(x = \{x : Xx\})$, so that the $\mathcal{L}^+$ abstractionist theory A is conservative if and only if T^E, $A \vdash C^E$ only if $T \vdash C$. The main idea here is that E picks out the *empirical* or physical items (and the restricted second-order quantifiers range over properties or sets of physical items) but we remain neutral as to whether mathematical items are part of the physical world or not. Thus, with T our empirical theory above, models of $T^{\sim M}$ embed the original physical structure in a substructure disjoint from the substructure satisfying the mathematical theory but T^E is compatible both with overlap and with disjointness. This type of principle I will call *Caesar-neutral* since such a principle applies equally well whether or not mathematical abstracts are necessarily disjoint from *empirical* items. But note also that $\mathcal{L}$ may already contain mathematical language, may contain some abstraction operators, (numerical operators, say) and in $\mathcal{L}^+$ we introduce a new one (a set-theoretic operator, for example). Here again the Caesar-neutral principle seems reasonable since it allows us to be neutral as to whether the new abstracts overlap the old ones or not; whether some numbers are also sets, it may be.

As introduced above, these two conservativeness principles, the Field and the Caesar-neutral, come themselves in two subbrands—syntactic and semantic.[21] Have we any grounds for preferring one to the other? Certainly there is something uncomfortable for the neo-Fregean in appealing to *semantic* consequence as part of a program designed to show that mathematics is analytic. For Frege's original idea was that mathematics should be *provable* from logic plus definitions, not that it should be a semantic consequence of it. Moreover, more recent attempts to legitimize the notion of analyticity appeal to such ideas as meaning-constitutive inference rules so there seems to be a close link between notions of analyticity and those of proof and derivation. On the other hand, unless one is prepared to accept (with Zermelo and a number of other prominent logicians of the earlier part of the last century—see Moore [18] and [19]) that infinitary proofs are as legitimate an idealization of actual inferential practice as proofs with $10^{10^{300}}$ steps, then proofs must satisfy the restrictions in the Gödelian theorems. In that case, a neo-Fregean who utilizes a proof-theoretic notion of entailment must give up on the completeness of analytically true mathematics.[22]

A further problem with a syntactic notion of conservativeness is that it is heavily dependent on proof architecture; on some proof systems even logic is not syntactically conservative. Thus in standard natural deduction systems, adding the negation rules to the $\rightarrow$ fragment of propositional logic yields new theorems in the old language, for example, Peirce's law—$(((P \rightarrow Q) \rightarrow P) \rightarrow P)$—while in others, for example, Gentzen's LK, the negative rules are conservative with respect to the negation-free fragment. But we surely do not want our notions of what is conservative, if questions of which mathematical principles are true or false are to hinge on

them, to depend on (arguably) aesthetic qualities such as the neatness, by this or that group's lights, of a particular proof architecture (cf. Weir [31]).

Moreover, there is an even more troublesome prospect for conservativeness defined syntactically in a variant of the Caesar-neutral form. Consider for the moment, for simplicity, only second-order abstractions. Suppose we add to the language not only a simple first-order predicate E which we envisage as picking out the empirical domain, but also a second-order predicate F which we use to relativize yet further the second-order quantifiers thus:

$$(\forall X\varphi)^{E,F} = \forall X((\forall y(Xy \rightarrow Ey)\ \&\ F(X)) \rightarrow \varphi^{E,F});$$

$$(\exists X\varphi)^{E,F} = \exists X((\forall y(Xy \rightarrow Ey)\ \&\ F(X))\ \&\ \varphi^{E,F}).$$

There seem little grounds for the neo-Fregean to object to second-order predicates which take first-order predicates as arguments. The abstraction operator, after all, is a second-order functional expression which takes first-order predicates as arguments.

One motivation for this modification of the Caesar-neutral criterion is to accommodate those who believe that not all predicates stand for genuine properties. Many scientific realists, for example, do not believe that all extensions determine a property; only some are the extension of genuine natural kinds which "cut reality at the joints." If we think of F then as picking out, in our intended interpretation, the genuine properties, then in relativizing a theory T—in a language extended by the introduction of an abstraction operator—to $T^{E,F}$ we are ensuring that the quantifiers in $T^{E,F}$ range over only the original empirical domain and the empirical properties of the items in that domain. Thus we should expect abstraction principles to be conservative here too. They should not enable us to prove anything about the original empirical domain or the empirical properties of items in that domain that we could not already prove before we added that principle.

If we construe this conservativeness principle syntactically, we get

If $T^{E,F}, A \vdash C^{E,F}$ then $T \vdash C$, E and F as above.

(Here T, C are wffs of $\mathcal{L}$.)

Theorem 3.1 *Any principle which, for some given infinite cardinality α, holds in all domains of size α[23] is syntactically conservative on the modified Caesar-neutral criterion.*[24]

Proof Suppose $\sim[T \vdash C]$; hence by Henkin completeness and Löwenheim-Skolem for Henkin models (see Shapiro [24], Section 4.3) there is a countable Henkin model H satisfying each instance of the axiom scheme of comprehension, with countable individual domain d, countable property domain D, D a set of subsets of d, such that $\models_H T, \sim C$. Suppose principle A is true in all domains of size α, for some infinite α. Add in enough individuals to d to get a size α domain d^* and expand the predicate domain to a full second-order domain $P(d^*)$. A is true in the new model H^* (in which we interpret all constants from the original language just as they are interpreted in H). Now just because there is a Henkin model satisfying $T, \sim C$ it does not follow that there is a full second-order model satisfying it. But we are concerned with $T^{E,F}, \sim C^{E,F}$. Let the extension of E (in H^*) be d and that of F be D so that in $T^{E,F}, \sim C^{E,F}$ first-order quantifiers are relativized so that they range over d, second-order relativized quantifiers range over D. Then an induction on wff complexity establishes that each wff in $T^{E,F}, \sim C^{E,F}$ has the same value in H^*

relative to any assignment to the first and second-order free variables of members of d and D, respectively, as the corresponding wff in T, $\sim C$ has in H. It follows that each wff in $T^{E,F}$, $\sim C^{E,F}$ is true in H^* if and only if the corresponding wff in T, $\sim C$ is true in H. Thus $\sim[T^{E,F}, P \models C^{E,F}]$. Hence by soundness $\sim[T^{E,F}, A \vdash C^{E,F}]$. □

Thus although the Caesar-neutral criterion is a very reasonable-looking requirement to place on a mathematical principle A in a language with second-order predicate constants, there is no problem in finding conservative, in this sense, but syntactically incompatible principles (e.g., distraction principles with Bad as exactly $\aleph_0$ versus Bad as at least $\aleph_1$).

4 Inconsistent Conservatives

So I will focus in this section on semantic conservativeness (omitting the qualification 'semantic' unless specifically wishing to contrast with syntactic conservativeness). I list now some apposite results.

Theorem 4.1 *Hume's Principle is conservative in the Field and Caesar-neutral senses (Pure* ZFC*).*

The meaning of the parenthetical reference to Pure ZFC is that, assuming ZFC in our metalanguage, we can prove that Hume's Principle is conservative in the Field and Caesar-neutral senses.[25] This of course is music to neo-Fregean ears.

We can get a more general result using the notion of an *unbounded* abstraction principle, defining this as a principle such that for every cardinal κ, there is a larger cardinal λ such that the principle is satisfiable in all domains of cardinality λ.

Theorem 4.2 *All unbounded principles are conservative (in both Field and Caesar-neutral senses) (*ZFC*).*

Proof of Theorem 4.1 Suppose $\sim(T \models C)$. Then there is a full set model M, of cardinality κ, in which all of T are true and C is false. Expand M to M^* by adding an $\aleph_\alpha$-sized, $\aleph_\alpha \geq \kappa$, set $\mathbb{N}$ of new members—the numbers—to the individual domain D_0 of M and taking the full power set $P(D_0^*)$ of $D_0^* = D_0 \cup \mathbb{N}$, as the property domain (with all nonlogical constants assigned the same interpretation in M^* as in M). Each set X in $P(D_0^*)$ has a cardinal card X and the number of these cardinals is the number β of cardinals $\gamma, 0 \leq \gamma \leq \aleph_\alpha$ and $\beta \leq \aleph_\alpha$. We can thus map the cardinals of the sets in $P(D_0^*)$ into $\mathbb{N} \subseteq D_0^*$ by a function f, interpreting $nx\varphi x$ by $f(\text{card } X)$ where X is the extension of φx; this yields an interpretation in which Hume's Principle is true in M^*. Define $\mathbb{N}x$ by $\exists X(x = nx(Xx))$ so that the extension of $\sim\mathbb{N}x$ is the set of nonnumbers of the domain, hence a subset of D_0. Call an assignment to free variables an M-assignment if and only if σ assigns only members of D_0 to individual free variables and only members of $P(D_0)$ to predicate free variables. A proof by induction on wff complexity then establishes that for all $P \in \mathcal{L}$, $P^{\sim\mathbb{N}}$ is satisfied in model M^* by an M-assignment σ if and only if P is satisfied by that same assignment σ in M. It follows that $P^{\sim\mathbb{N}}$ is true in M^* if and only if P is true in M, hence M^* is a counterexample model for the entailment $[T^{\sim\mathbb{N}}, \text{HP} \models C^{\sim\mathbb{N}}]$. A variant of this argument establishes the result for the Caesar-neutral criterion. □

It will be useful here to generalize beyond second-order abstractions to the higher-order case. Rather than start from a base language of simple type theory, a cumulative type theory will suit our purposes better. Here we suppose that for every finite order we have predicates and variables of that order (countably many in the latter case) and that an atom $F(G)$ is well formed if and only if the order of F is greater than that of G. In the semantics, a (standard) model is a pair $\langle d, I\rangle$ where d is the individual domain, the range of the 0th-order variables. The cardinality of the model is the cardinality of d. The second component I of a model is an interpretation of all the constants in the appropriate domains. The nth-order quantifiers range over the nth-order domain (as remarked, we will get by with monadic predication since we can introduce ordered pairs once we have some set theory). The $(n+1)$th-order domain D_{n+1} is $D_n \cup P(D_n)$, the union of D_n with its power set. More generally, where $s \subseteq d$, define $s_0 = s$, $s_{n+1} = s_n \cup P(s_n)$ and define the cumulative hierarchy generated by s by $\bigcup_{i\in\omega} s_i$.

This language $\mathcal{L}_C$ of cumulative type theory is then expanded to a language $\mathcal{L}^+$ by adding an $(i+1)$th-order abstraction operator. We can think of $\mathcal{L}_C$ as the base language $\mathcal{L}_0$ of hierarchy. At $\mathcal{L}_{n+1}$ we apply the operator in question, for example, class brackets, to one-place open sentences $\varphi^{i+1}X^i$ of $\mathcal{L}_n$ to get the new singular terms $\{X^i : \varphi^{i+1}X^i\}_{i+1}$ of $\mathcal{L}_{n+1}$ and expand the set of atoms to include these. $\mathcal{L}^+$ is then $\bigcup_{i<\omega} \mathcal{L}_i$.[26] The interpretation of these class operators then will be that they represent functions from the nth-order properties into the individuals. Thus a fourth-order abstraction will take the form

$$\forall X^3 \forall Y^3 (ox X^3 x = ox Y^3 x \leftrightarrow E(X^3, Y^3))$$

where E is an equivalence relation over third-order predicates.

In the semantics we show by induction that interpretations are *stable* through the hierarchy, in that a wff has the same value (relative to an assignment) in all sublanguages in which it occurs; therefore it can be assigned a unique value in $\mathcal{L}^+$. A further useful extension is to add *abstractor quantifiers*.[27] Abstraction operators of order $n+1$ are formally functional terms which take nth-order open sentences as arguments and yield singular terms as outputs.[28] We can thus add, at each order of the language, quantifiers over such terms. In the language thus augmented, there can therefore occur sentences such as

$$\exists f^4 (\forall X^3 \forall Y^3 (f^4 X^3 = f^4 Y^3 \leftrightarrow E(X^3, Y^3))).$$

The range of an $(n+1)$th-order abstraction quantifier f is the set of all functions from D_n into D_0. Finally we can iterate this whole process by adding a further abstraction operator to generate a language $\mathcal{L}^{++}$ and so forth.

In order to prove Theorem 4.2 we need to generalize the notion of the relativization of a formula to our more complex languages. Define by recursion the metatheoretic terms $A^n[X]$, where X is an nth-order variable, by

$$\begin{aligned} A^1[X^1] &\equiv_{\text{def}} \forall y (X^1 y \rightarrow Ay), \\ A^{n+1}[X^{n+1}] &\equiv_{\text{def}} \forall Y^n (X^{n+1} Y^n \rightarrow A^n[Y^n]), \end{aligned}$$

and then generalize the predicate quantification clauses in the definition of φ^A to

$$\begin{aligned} (\forall X^n \varphi)^A &= \forall X^n (A^n[X^n] \rightarrow \varphi^A), \\ (\exists X^n \varphi)^A &= \exists X^n (A^n[X^n] \,\&\, \varphi^A), \end{aligned}$$

adding, for abstractor quantification,

$$(\forall f^n\varphi)^A = \forall f^n\forall X^n\forall y((f^nX = y \rightarrow (A^n[X^n] \rightarrow Ay)) \rightarrow \varphi^A),$$
$$(\exists f^n\varphi)^A = \exists f^n\forall X^n\forall y((f^nX = y \rightarrow (A^n[X^n] \rightarrow Ay)) \,\&\, \varphi^A).$$

Thus the relativized abstractor quantifiers generalize over functions whose range, for any properties in the cumulative hierarchy generated from the subset d_A of d which satisfies A, is also a member of d_A.

Now we can return to the proof of Theorem 4.2—All unbounded principles are conservative (in both Field and Caesar-neutral senses)—in which we are considering abstraction principles of arbitrary order in a language $\mathcal{L}^+$ which extends, by the addition of the operator, a language $\mathcal{L}$ which may itself contain other abstraction operators and other nonlogical names and predicates.

Proof Suppose then that nth-order abstraction principle A is unbounded and that $\sim(T \models C)$ where all wffs in T, C belong to $\mathcal{L}$. Let M be a counterexample model to the entailment with individual domain d of size α. Since A is unbounded, it is true in all models of some cardinality $\beta \geq \alpha$. Expand, if need be, the individual domain of M to create a size β domain d^* of a new standard model M^*. The interpretation function I^* of M^* agrees with I on all name and predicate constants. Furthermore, each nth-order abstraction operator in $\mathcal{L}$ is interpreted just as it is in $\mathcal{L}$, for inputs from D_n. For members of $D_n^* - D_n$ we let all the operators map to some dummy object in d. This means that the abstraction principles other than A may fail in $\mathcal{L}^+$. However, let $|D_n^*|$, where D_n^* is the range of the nth-order predicate variables in M^*, be the partition of D_n^* effected by the equivalence relation on the right-hand side of A. Since A holds in all models of cardinality β, there exists a function g from $|D_n^*|$ into d^*.[29] Interpreting the abstraction operator $\{x : \varphi x\}$ of A by g, A is true in M^*. Where $S \subseteq D_n^*$ is the interpretation of φx in $\mathcal{L}_n$ and s the member of $|D_n^*|$ to which S belongs, we assign $g(s)$ as the referent of $\{x : \varphi x\}$ in $\mathcal{L}_{n+1}$ and show that semantic values are stable as we go through the hierarchy. Finally we prove by induction on the complexity of arbitrary wff P that P has the same value in M relative to M-assignment σ as P^*, the Field or Caesar-neutral relativization of P, has in M^*. In the Caesar-neutral case, we assign d as the extension of Ex. An M-assignment assigns to each variable an item of the appropriate order from the cumulative hierarchy generated by d. The proof is a relatively straightforward generalization of the analogous stage in the proof of Theorem 4.1. For example, if we assume the theorem holds for φx and we are considering the inductive case for an individual universal quantification then for the Caesar-neutral criterion (the argument for the Field case is similar) we argue

> $\forall x\varphi x$ is true in M relative to σ iff
>
> for all x-variant (M) assignments $\sigma[x/\alpha]$, φx is true in M relative to $\sigma[x/\alpha]$ iff (Inductive Hypothesis)
>
> $\varphi^E x$ is true in M^* relative to $\sigma[x/\alpha]$, for all $\alpha \in d$ iff
>
> $\forall x(\sim Mx \rightarrow \varphi^E x)$ is true in M^* relative to σ (since any non-M-assignment x-variant $\sigma[x/\beta]$ satisfies $Ex \rightarrow \varphi^E x$ vacuously since β does not satisfy Ex).

Hence P is true in M if and only if P^* is true in M^* so that $\sim(T^*, A \models C^*)$.[30] □

We have seen that there are at least some unbounded abstraction principles—Hume's Principle is one—and it is easy to see that there are also unbounded distraction principles, for instance D^{Inf} where Bad = Dedekind-infinite. But are there too many unbounded abstraction principles? The answer is yes. For instance, suppose the general continuum hypothesis (GCH) is true. Then Bad as 'is the size of a successor cardinal' is satisfied at every successor cardinal $\aleph_{\alpha+1}$ since the number of small subsets is $\aleph_{\alpha+1}$—a fortiori the number of subsets not the size of a successor cardinal is $\leq \aleph_{\alpha+1}$ (cf. [26], p. 315). But by similar reasoning the distraction principle with Bad as 'the size of an "odd" successor', with an odd successor a cardinal of the form $\aleph_{\alpha+2n+1}$, is true at all odd successor cardinals; and similarly the distraction principle with Bad as 'the size of an even successor cardinal' is true at all even successor cardinals, granted GCH. These last two principles are unbounded, hence conservative, but clearly are not simultaneously satisfiable in a full standard model (cf. Fine [6], p. 514).

Or dropping GCH in the background metatheory but adding the strong axiom of inaccessibles—for every cardinal κ there is a larger (strong) inaccessible—we can show that taking Bad as 'has the size of a successor in the series of inaccessibles' yields an unbounded (hence conservative) distraction principle incompatible with taking Bad as 'has the size of a limit in the series of inaccessibles', though the latter is similarly unbounded ([26], p. 319).

The neo-Fregean may well refuse to accept the truth of GCH and might not accept the axiom of inaccessibles (though the latter is very widely accepted among set-theorists). But embarrassment of riches arises on weaker assumptions. Take any predicate φ such that the φs and the non-φs are unbounded, that is, φx might be 'x is a successor cardinal'. There are infinitely many such predicates. Next take any 'at least κ' distraction principle D (i.e., in the principle D, Bad is 'at least of size κ') which holds at arbitrarily high φ cardinals and also at arbitrarily high non-φ cardinals. (Since it is a logical principle, D will hold in all models of cardinality κ if it holds in at least one.) 'At least countably infinite' will always satisfy these conditions. Now consider,

D_1 $\quad \text{Bad}_1(X) = X$ is size at least κ and there is some Y with $X \subseteq Y$ such that card(Y) is a φ cardinal.

D_2 $\quad \text{Bad}_2(X) = X$ is size at least κ and there is some Y with $X \subseteq Y$ such that card(Y) is a non-φ cardinal.

Theorem 4.3 (ZFC) *D_1 and D_2 are unbounded, pairwise unsatisfiable principles.*

Proof For any cardinal, we can find a larger φ cardinal $\aleph_\alpha$ such that D holds at $\aleph_\alpha$; it follows that the number of subsets of size $< \kappa$ of any $\aleph_\alpha$-sized domain must be $\leq \aleph_\alpha$. All subsets of size $\beta, \kappa \leq \beta \leq \aleph_\alpha$ are Bad_1, however, since they are of size at least κ and a subset of a property, $\lambda x(x = x)$, whose cardinality satisfies φ. Hence all these Bad_1 properties can be mapped onto the dummy proper class and the rest, the Good_1 properties, bijected into the domain of individuals. Thus D_1 is satisfiable in every model of size $\aleph_\alpha$. But D_2 is satisfiable in no φ cardinal-sized model. For any universe-sized subset of the domain is Good_2, since it is not a subset of a set whose size is a φ cardinal. But there are $2^{\aleph_\beta}$ such universe-sized Good_2 subsets, where $\aleph_\beta$ is the size of the domain, so D_2 is not satisfiable in such a model.

Similarly D_2 is satisfiable at arbitrarily high non-φ cardinals but D_1 is satisfiable at none of them. □

So once again we see that not only are there consistent but pairwise inconsistent duos of abstraction principles; there can also be pairwise incompatible but semantically conservative abstraction principles. Wright's first criterion for winnowing out good from bad abstractions—conservativeness—cannot do the filtering job on its own.

Perhaps the neo-Fregean will reject even this metatheoretic argument establishing the existence of jointly incompatible but conservative theories, though it would seem that to do so the neo-Fregean would need to have no truck with ZFC set theory and all its works and pomps. Certainly, there would be a pragmatic inconsistency or self-refutation if the neo-Fregean relied, in a metatheoretic validation of his or her position, on results which could not be derived from abstraction principles which the neo-Fregean found acceptable and it is possible that the neo-Fregean will settle on abstraction principles incompatible with ZFC. On the other hand, ZFC is a very fruitful mathematical theory which is accepted by most set-theorists. This does not preclude the possibility of root and branch criticism of the theory from the philosophers but, unless the theory can be shown to be inconsistent, the more of this standard mathematical theory the neo-Fregeans reject, the less plausible their position becomes. Hence the neo-Fregeans, though they ought to aim at eventually throwing away the ladder of ZFC and similar set theories, will want to land on a spot from which a large body of that theory (certainly enough to do contemporary physics and to yield conservativeness results for proper parts of the theory, such as Hume's Principle) can be recaptured.

5 Modest Conservatives

Can we get around the problem raised in Section 4 if we tighten further the conditions on the acceptability of abstraction principles? Wright proposed in [36] a second conservativeness criterion which he later characterizes thus:

> Distractions entail conditionals of the form:
>
> $$-(\exists F)(\phi F) \rightarrow (\forall F(\forall G)(\Sigma F = \Sigma G \leftrightarrow (\forall x)(Fx \leftrightarrow Gx))$$
>
> The immediate intent of the proposed constraint is that anything derivable by the *reductio* of the antecedent of such a conditional afforded by its paradoxical consequent should be in independent good standing. . . . So, an abstraction is good only if any entailed conditional whose consequent is Basic Law V (or, therefore, any other inconsistency) is such that all further consequences which can be obtained by discharging the antecedent are in independent good standing, as may be attested by their derivation in pure higher-order logic (like the case of New V) or their independent derivability from the abstraction in question (like the case of Hume's Principle). (Wright [38], p. 326)

So let A be any abstraction and C any consequence of A. Classically, C is equivalent to $\sim C \rightarrow \bot$, so Wright requires that $\sim\sim C$, which can be obtained by discharging the antecedent, is of "independent good standing," hence requires (granted the classical equivalence of C and $\sim\sim C$) that anything derivable from an abstraction be derivable "independently."

Wright acknowledges the need for clarification here:

> But what does that mean? In particular, how might it be characterized so as not to outlaw any proof by reductio ad absurdum? ([38], p. 327)

Wright suggests that an "independent derivation" must not be "paradox-exploitative" and gives the following account of the latter notion:

> a derivation from a conservative abstraction is paradox-exploitative just if there is a representation of its form of which any instance is valid and of which some instance amounts to a proof of the nonconservativeness of another abstraction. For instance, the derivation of the successor-inaccessibility of the universe from the Distraction canvassed above is paradox-exploitative because it may be schematized under a valid form of which another instance is a derivation, from the appropriately corresponding Distraction, that the universe contains 144 objects. ([38])

The corresponding distraction is presumably the distraction with Bad = 'has exactly 144 instances' but that distraction is unsatisfiable. But perhaps Bad = 'is of size $\aleph_0$' will do the job just as well for Wright, since this puts a cap on the physical universe, contrary to conservativeness.

All this is surely rather odd. It cannot be that an axiom or principle P is suspect because there is a proof π of C from P which shares a form with a proof π^* of D from Q, D and Q instantiating the relevant schematic forms of C and P, and where D is something we would reject: this kind of "sharing form" is not criminal. Wright requires that Q is not just any old formula: it must pass the first criterion of conservativeness. And it is true that the most obvious proof that the universe is at least successor inaccessible from the distraction with Bad = 'at least successor inaccessible' shares form with a similarly obvious proof (utilizing the collapse of the distractions into Axiom V if there is no Bad property) that the universe is of size exactly $\aleph_0$, a proof whose premise is the distraction with Bad = 'is of size $\aleph_0$'. But then there is a similar proof that the universe is infinite from the distraction D^{Inf} in which Bad = 'Dedekind infinite'. Is this distraction to be rejected because of structural similarities between proofs of results from D^{Inf} as premise and proofs of dodgy results, such as that the universe is of size exactly $\aleph_0$, from other abstractions? For D^{Inf} is satisfiable at all infinite cardinalities, just like Hume's Principle.

Wright may say that there are "independent proofs" of the infinity of the universe from D^{Inf}, ones which in a clear sense appeal only to properties of the abstracts themselves, for instance by proving that there are infinitely many natural numbers, that is, set-theoretic surrogates for natural numbers defined in the usual Zermelo or von Neumann ways. But 'paradox-exploitation' was supposed to give sense to the notion of 'independent derivability'; we cannot then require the latter notion to make sense of the former. Moreover, it is not true that, as Wright says,

> the only resources they ["roguish distractions"] have to show . . . that the universe is limit-inaccessible or successor inaccessible, or whatever, are those furnished by the inconsistency of Basic Law V and the consequent modus tollens on the relevant conditional. ([38], p. 326)

For any proof of a result C from premise P there are infinitely many other proofs of that result from that premise. Consider the following proof schema, applicable to any distraction, that there are a Bad number of abstracts, specifically sets:

> Take $r = \{x : \text{Set } x \,\&\, x \notin x\}$. If r is a Set, if the property $\lambda x(\text{Set } x \,\&\, x \notin x)$ is Good, then comprehension holds of r. That is, $\forall y(y \in r \leftrightarrow (\text{Set } x \,\&\, x \notin x))$, so in particular,
>
> $$r \in r \leftrightarrow \text{Set } r \,\&\, r \notin r,$$
>
> from which it follows that $\sim$Set r. So Set $r \rightarrow \sim$Set r hence $\sim$Set r, that is, from our definition of Set, the property of being a non-self-membered Set is Bad. So, if Bad is some cardinality concept, we can prove in the above fashion that there are a Bad number of sets, namely, the sets which do not belong to themselves.

This proof seems as set-theoretic as any. Yet we can use it to show that the universe must be *at least* of the cardinality given by the Badness concept, since the subuniverse of sets has that cardinality, without appealing to any result about the cardinality of the whole universe. (To be sure, since the Bad cardinal is infinite so that all singleton properties determine unit sets, we can conclude further that the universe is *exactly* of the cardinality given by Bad, but this way of proving the result "originates in a requirement that the distraction imposes on its own abstracts" to paraphrase Wright [38], p. 329.)

Is the above proof *paradox-exploitative*? If so, what of the standard proofs that there is no Russell set, that there is no universal set (else by Subsets there would be a Russell set), or that the powerset of x is larger than x—why are the standard proofs of these results not also paradox-exploitative? If so, is this exploitation such a bad thing?

Another constraint which Wright suggests adding to conservativeness is "modesty":

> an abstraction is Modest if its addition to any theory with which it is consistent results in no consequences—whether proof- or model-theoretically established—for the ontology of the combined theory which cannot be justified by reference to its consequences for its own abstracts. And again, *justification* is the crucial point: an abstraction may fail this constraint even though every consequence it has for the ontology of a combined theory may be seen to *follow from* things it entails about its proper abstracts; in particular, it will not count if, as in the case of the Limit-inaccessible Distraction, a consequence for the combined ontology is needed as a lemma in the proof that the abstracts have a property from which that very consequence follows. ([38], p. 330, Wright's emphasis)

Wright's emphasis on justification is indeed essential here. For suppose we drop all reference to issues of justification. What is left seems to be a reflection principle which I will call Modest Reflection. Let $\mathcal{L}$ be an abstraction-operator free language, $\mathcal{L}^+$ the extension of $\mathcal{L}$ resulting from adding an abstraction operator governed by a logical abstraction A, and P a sentence of $\mathcal{L}$.

Modest Reflection — If $A \models P$ then $A \models P^M$ (and of course there is a syntactic version in which $\models$ is replaced by $\vdash$).

That is, if a thesis holds in all A universes, the abstract subuniverse *reflects* that thesis—the consequence P for the combined ontology holds only when the restricted version of P holds for the abstracts. In such a case let us say that *A reflects modestly*; the principle is a sort of negative converse of conservativeness:

If $A, T^{\sim M} \models P^{\sim M}$, then $A \models P$,

where $^{\sim M}$ restricts to the *non*abstracts.

Wright's text, in particular the reference to "no consequences" for the combined theory which "cannot be justified by referent to *its* consequences for *its own* abstracts (emphasis mine) suggests the stronger principle,

If $A, T \models P$ then $A \models P^M$ (for any T consistent with A).

But this constraint seems too strong. Suppose we take HP and add it to ZF (but the point will also hold for *empirical theories*); so far as we know HP is consistent with ZF. Or equivalently add HP to (HP $\rightarrow$ ZF), with ZF a second-order finite axiomatization. The pair HP, (HP $\rightarrow$ ZF) entail that there is an uncountable infinity of individuals. But HP on its own does not entail that there is an uncountable infinity of numbers, so HP comes out as immodest on this reading. Perhaps the constraint is rather

If $A, T \models P$ then $A, T \models P^M$.

But this just is Modest Reflection where T has a finite axiomatization or where we allow infinitary wffs, for then we just consider $T \rightarrow P$.

Theorem 5.1 *Every logical distraction in which unit properties are Good reflects modestly.*

Proof Suppose there is a counterexample model M to A entails P^M, one with domain d. Let $n \subseteq d$ be the set of all referents of individual constants in P and $a \subseteq d$ be the set of all abstracts in M. Since unit properties are Good, A holds only in infinite domains and a, d, and $n \cup a$ all have the same cardinality. Construct the model M^* by letting its domain be $n \cup a$, so its variables range over the cumulative hierarchy $\mathrm{CH}_{n,d}$ generated by $n \cup a$; interpret its individual constants as in M and its predicate constants by the restriction of the M-interpretation to $\mathrm{CH}_{n,d}$. This is a counterexample model to P—proof by induction over wff complexity. Since the distraction is logical and since M^* is the same size as M, A holds in M^* too. Hence A does not entail P. □

So Wright needs the clause about all consequences being "justified" and not merely "following" from "things it entails about its proper abstracts." But what on earth does this mean? It suggests some tight proof-theoretic notion, as when a classicist might hold to classical semantic consequence but pay special attention to consequences derivable in relevant logic or some such. Even if something could be made of this, what on earth does it have to do with "meaning-constitutive" or "a priori" or "epistemically innocent" principles? One can see how simple rules such as $\&E$ or $\vee I$ are meaning-constitutive (if, at any rate, one is not rabidly Quinean to an extent that the later Quine himself shied away from). But it is very hard to see what proof-theoretic modesty or the complex definition of paradox-exploitation has to do with this. The whole approach exudes a strong whiff of ad hocery; the epicycles which are being generated give out strong signals that we are in the presence of a degenerating research strategy, if not program,[31] as Wright himself seems to acknowledge:

> That is apt to seem uneasily complex and less clearly motivated than one would wish. ([38], p. 327)

6 Stability

However, just as the neo-Fregean program seems to be in deep trouble, Wright comes up with a much more powerful, simple, and intuitive idea: any epistemically kosher abstraction must not only be conservative, it must be compatible with all other conservative abstractions:

> it is not clear that any purpose is served by the continuing insistence on derivations of a given valid form. Why not just say that pairwise incompatible but individually conservative abstractions are ruled out—however the incompatibility is demonstrated—and have done with it? ([38], p. 328)

Are there any abstractions which are both conservative and compatible with any other conservative abstraction (i.e., there is a model in which both are true)? Call any such abstraction *irenic*; and say that an abstraction is *stable*, if for some cardinal κ, it is true at all and only models of cardinalities $\geq \kappa$ (cf. [6], p. 511).

Theorem 6.1 *The stable abstractions are the irenic ones.*

Proof (Left to right) Suppose A is stable; by Theorem 4.2 it is conservative, being unbounded. Since A is stable it holds in all models $\geq \kappa$, for some κ (remember *conservative* simpliciter means semantically conservative). Consider now a "Ramsified" version of A in which we replace each constant term (name, predicate, abstraction operator) by a variable of appropriate type and preface the result $A[x_1, \ldots, x_n]$ (where the variables need not all be individual variables) by the corresponding string of existential quantifiers to get $\exists(x_1, \ldots, x_n)A[x_1, \ldots, x_n]$, a purely logical formula I will represent by $\exists[A]$. This formula cannot be true in a model M of size less than κ else by interpreting each constant c by the object, property, or operator function assigned to the corresponding variable x_i in the assignment which satisfies $A[x_1, \ldots, x_n]$ we would generate a model of size $< \kappa$ which satisfies A.

Now let B be another conservative principle introduced by a new abstraction operator and take the language of principle A to be the base language $\mathcal{L}$ for the new principle, so that by adding the abstraction operator of B to $\mathcal{L}$ we get our new language $\mathcal{L}^+$. We cannot have

$$B, (\exists(x_1, \ldots, x_n)A[x_1, \ldots, x_n])^{\sim B} \models \bot,$$

else by conservativeness we would have $\exists(x_1, \ldots, x_n)A[x_1, \ldots, x_n]) \models \bot$ and hence $A \models \bot$, contrary to the stability of A. So there is a model N of B, $(\exists(x_1, \ldots, x_n)A[x_1, \ldots, x_n])^{\sim B}$. Moreover, if we reduce this to a model $N^{\sim B}$ with individual domain the non-Bs, the result will be a model of $(\exists(x_1, \ldots, x_n)$ $A[x_1, \ldots, x_n])$ since this is a purely logical sentence. By interpreting each constant c—name, predicate, or operator—in A by the item assigned to the variable which instantiates c in $A[x_1, \ldots, x_n]$ by an assignment σ which verifies $(\exists(x_1, \ldots, x_n)A[x_1, \ldots, x_n])$ we get a model $N^{*\sim B}$ in which A is true. Hence $N^{*\sim B}$, and thus $N^{\sim B}$ and so N must be of size $\lambda \geq \kappa$. But N is a model of B. By the definition of stability, A is true in N together with B.

(Right to left) Every irenic abstraction is stable. Suppose nth-order abstraction A:

$$\forall X \forall Y(\alpha x Xx = \alpha x Yx \leftrightarrow E(X, Y))$$

is unstable so that for each cardinal κ, there is a higher cardinal λ such that A fails at some model of size λ. In such a model the $(n+1)$th-order formula

$$\exists f \forall X \forall Y(fX = fY \leftrightarrow E(X, Y))$$

fails. Consider now the abstraction B:

$$\forall W\forall Z(\beta x\,Wx = \beta x\,Zx \leftrightarrow [\sim\exists f\forall X\forall Y(fX = fY \leftrightarrow E(X,Y)) \vee \forall x(Wx \leftrightarrow Zx)]).$$

The right-hand side is an equivalence relation since whenever the left disjunct is true (and so abstraction A false) every property bears the relation to every other while when the left disjunct is false the whole formula is coextensive with the equivalence relation $\forall x(Wx \leftrightarrow Zx)$. But when the left disjunct is true, principle B is trivially satisfied by letting $\beta x\,Wx = \beta x\,Zx$ for any assignment to W and Z, that is, by having a single abstract, while principle A is unsatisfied. On the other hand, when the left disjunct is false so is B, because it is equivalent in those contexts to Axiom V, though abstraction A is true. Since $[\sim\exists f\forall X\forall Y(fX = fY \leftrightarrow E(X,Y))$ holds at models of arbitrarily high cardinalities, B is unbounded and so conservative.[32] But as we have seen, A is semantically incompatible with B so A is not irenic. □

What the neo-Fregean needs then are (nontrivial) stable principles, best of all stable principles which do not hold below the continuum but "kick in" a few beths further up. For in that case, stable abstraction principles will suffice for the derivation of the mathematics needed for modern science; they will provide abstract ontologies of sufficient size to construct the reals, complex numbers, functions over reals and so forth.[33] Now Shapiro and Weir ([26], p. 319) showed that "at least κ" distraction principles, $\kappa > \omega$, are unstable (there stability is called "the strong unbounded condition," cf. p. 318), every such distraction failing at each of an unbounded series of singular limit cardinals. But in the context of our cumulative type theory, we can find fairly natural distraction principles which are stable.

For example, start either from Hume's Principle or the comparable but in some respects more useful distraction D^{Inf}:

$$\forall X\forall Y(\alpha x\,Xx = \alpha x\,Yx\} \leftrightarrow ((\text{Infinite}(X)\ \&\ \text{Infinite}(Y)) \vee \forall x(Xx \leftrightarrow Yx))).$$

(where 'Infinite' is 'Dedekind Infinite', for example, there is a bijection from the property into a proper subproperty). Using AC we can prove D^{Inf} is true in all infinite cardinalities (at $\aleph_\kappa$ there are $\aleph_\kappa$-many finite sets; map the others to the dummy proper class). Moreover, from this, from the fact that all finite properties determine sets, it is clear that semantically it is at least as strong as SF (ZF minus the axiom of infinity) restricted to pure sets (to exclude the ill-founded ones).

Classing our initial principle as D_1 we now add a further second-order Distraction principle in which Bad, or rather Bad^2, is $\sim\text{Num}^2(X^1)$ where Num x is our definition of the finite numbers or their set-theoretic surrogates and

$$\text{Num}^2X \equiv_{\text{def}} \forall x(Xx \rightarrow \text{Num}\ x).$$

D^2 is

$$\forall F^1\forall G^1(\{x : F^1x\}_1 = \{x : G^1x\}_1 \leftrightarrow ((\sim\text{Num}^2(F^1)\ \&\ \sim\text{Num}^2(G^1)) \vee \forall x(F^1x \leftrightarrow G^1x))).$$

By dint of the occurrences of the numerical or zero-order class operator on the right-hand side (when we unpack Num^2), this is a nonlogical abstraction. The Bad first-order properties, as specified by this distraction, are those which are nonnumerical^2, that is, are not subsets of the set of finite numbers of the domain.[34]

Next add the third-order Distraction principle in which Bad^3 is $\sim \text{Num}^3(F^2)$ where

$$\text{Num}^3X^2 \equiv_{\text{def}} \forall Y(X^2Y \rightarrow \text{Num}^2Y).$$

This third-order distraction D^3 is then

$$\forall F^2 \forall G^2(\{X : F^2X\}_2 = \{X : G^2X\}_2 \leftrightarrow \\ ((\sim \text{Num}^3(F^2) \;\&\; \sim \text{Num}^3(G^2)) \vee \forall X(F^2X \leftrightarrow G^2X))).$$

The Bad second-order properties, as specified by this distraction, are those which are nonnumerical^3, that is, not all of the first-order properties which instantiate them are numerical^2 properties.

Continue further by adding a fourth-order distraction D^4 with Bad^4 defined in terms of Num^4—having only Num^3 instances—

$$\text{Num}^4X^3 \equiv_{\text{def}} \forall Y(X^3Y \rightarrow \text{Num}^3Y)$$

and so on through all the finite types.[35]

Theorem 6.2 *The set of all these principles is satisfied in all and only models of size* $\geq \beth_\omega$. *It is stable and irenic.*

Proof Take any standard model M with individual domain d of cardinality $\geq \beth_\omega$. This will satisfy D^{Inf} (or HP) by assigning some countable subset as the extension $|\text{Num}|$ of Num. Again in every standard model, the continuum-sized powerset of $|\text{Num}|$ is the extension $|\text{Num}^2|$ of Num^2, the $\beth_2$-sized powerset of $|\text{Num}^2|$ is the extension of Num^3 and so forth. Since there are continuum-many Good^2 (i.e., Num^2) first-order properties, D^2 is satisfiable by mapping these into a continuum-sized subset of d and all other properties into a dummy class and using that map to interpret the operator $\{x : F^1x\}_1$. Note that D_2 could not be satisfied in any domain smaller than the continuum. Similarly we interpret D^3 by means of a map from the $\beth_2$ many Good^3 properties into the domain, and likewise through all the principles D^i for $i \in \omega$. Hence $\bigcup_{i\in\omega} D^i$ is satisfied by M, though in any domain smaller than $\beth_\omega$, for some k, all principles D^j for $j \geq k$ will fail to be satisfied. Moreover, we can show that $\bigcup_{i\in\omega} D^i$ is irenic by essentially the same argument as used in Theorem 6.1. We Ramsify each D^i to yield a purely logical sentence,

$$(\exists(x_1, \ldots, x_n)D^i[x_1, \ldots, x_n]).$$

Where B is any conservative abstraction, the set

$$\{B, (\exists(x_1, \ldots, x_n)D^i[x_1, \ldots, x_n])^{\sim B}(I \in \omega)\}$$

is satisfiable in a model N, else

$$(\exists(x_1, \ldots, x_n)D^i[x_1, \ldots, x_n])(i \in \omega) \models \bot,$$

contrary to the satisfiability of $\bigcup_{i\in\omega} D^i$. By shrinking N down to the submodel $N^{\sim B}$ with individual domain the non-Bs we get a model which satisfies all of the $(\exists(x_1, \ldots, x_n)D^i[x_1, \ldots, x_n])$ and so a variant model of the same size which satisfies $\bigcup_{i\in\omega} D^i$. This shows as before that N must be of size $\geq \beth_\omega$ hence, by the stability of $\bigcup_{i\in\omega} D^i$, the set of sentences $\bigcup_{i\in\omega} D^i$, B is satisfied by N. □

This theory $\bigcup_{i\in\omega} D^i$—call it BETH_ω—is thus immune from the embarrassment of riches problem *and* gives us a slice of the cumulative hierarchy up to $V_{\beth_\omega}$ albeit in a rather restrictive form. We have the natural numbers, all sets of natural numbers, all subsets of the powerset of the set of natural numbers and so on. Ontologically, then, we have all the pure structures we need for the applied mathematics for contemporary science, numbers, reals, functions over reals and so on. However quite simple set-theoretic principles fail. Thus if $\{x : \varphi x\}$ is a set of order $n + 1$ then there is no guarantee that its unit set exists (as a set) because there is no guarantee that $\{x : \varphi x\}$ is also a set of order n. Nor is it clear how the neo-Fregean could actually apply this ontology in science since there are no sets of urelements, just sets of numbers, sets of sets of numbers and so forth. Perhaps she could introduce a further "impure" set operator, for instance, one governed by a distraction principle with bad as 'at least $\beth_\omega$'. This principle is not stable and neither is the result of augmenting BETH_ω with it. But perhaps the neo-Fregean could accept this: there is no a priori applied set theory but there is, she might claim, an a priori pure mathematical theory, BETH_ω. And if we need more things in our heaven and earth than provided for by BETH_ω we can extend the type theory into the transfinite and thereby force the size of the universe up even higher.

This prospect raises a worry. If there is the possibility of adding stronger and stronger such principles, how big is the universe? Might there not be a proper class of stable principles, in which case, if the lower limits which each principle forces the universe to have are unbounded, there will be no (set-theoretic) model of the whole set of principles (cf. [6], p. 514). But this situation is not so different from that which faces the ZF theorist who cannot prove that a set-theoretical model for her intended interpretation of the theory exists. It is consistent with ZF that there are no inaccessible cardinals, in which case the set of ZF axioms holds in no set-sized standard model. Moreover the "intended model" has a domain—the universe of sets—which is provably, in the theory itself, not a set. This shows that stability cannot be a necessary condition on acceptability of a theory. One might, though, try for a more disjunctive criterion: a principle A is acceptable if and only if it is either stable or true (or necessarily true) in the intended interpretation. Or, to avoid adding in a primitive truth predicate or ascending up a further order in the type theory in order to define truth, we could define acceptability, relative to an abstractionist theory A, by [P is stable or P is provable from A]. If the abstractionist theory A suffices for sufficient proof theory to let us represent the relation 'provable in second-order logic from A' then the abstractionist theory will be able to prove its own acceptability.

7 ER II

Has the neo-Fregean hit the jackpot then? One cause for concern surfaced earlier in connection with the criteria of paradox-exploitation and justificatory modesty. It is not enough for the neo-Fregean to find a criterion which characterizes a consistent set of abstraction principles which together yields as much mathematics as we think we need (for application in science, for example). The neo-Fregean also needs an argument which shows that *all* the principles satisfying the criterion are analytic or meaning-constitutive or implicit definitions which in some interesting sense are epistemically innocent. We could come to know their truth without resort to mysterious intuition or appeal to pragmatic criteria of usefulness of science. But what has the acceptability, in the sense of the previous section, of an abstraction got to

do with it being meaning-constitutive or an implicit definition? This objection can be given more force by considering the following worry, analogous to the original embarrassment of riches objection.

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. Angus is a finitist who accepts as his sole second-order abstraction principle the distraction D^{CInf} with Bad = Countably Infinite. He holds that the only properties one can generalize over in abstraction principles are numerically definite ones and maintains that only finite properties are numerically definite. Indeed he might hold that only such properties exist. Bronagh, however, takes as her principle the distraction with Bad = $\beth_\omega$-sized,[36] while for Calum, Bad = the size of the first inaccessible, θ_0. Dervla defines Bad(F) by

$$\mathrm{Big}(F)\ \&\ F \text{ is the size of a Mahlo cardinal } \&\ \sim \mathrm{GCH}$$

so that, since Dervla can prove the universe is Bad, Dervla can prove the General Continuum Hypothesis (GCH) is false. Finally Ewan, who thinks that all the others are wimps, defines Bad(F) by

$$\mathrm{Big}(F)\ \&\ F \text{ is the size of a measurable cardinal } \&\ \mathrm{GCH}$$

so that Ewan can prove the GCH.

Suppose now we agree with Calum. Then we can rule out Angus's theory, since it places a cap on the universe at $\aleph_0$ and we know that the universe is bigger than that; indeed we might believe the empirical universe has more individuals than that, has continuum-many spacetime points, perhaps. Angus's theory is unstable and nonconservative. Where P is the claim that there are least $\aleph_1$ things and where Ax picks out the abstracts of D^{CInf}, we have $P^{\sim A}$, $A \models \bot$ but not $P \models \bot$ (we believe). Indeed Angus's theory is provably false, from our perspective, since the universe is provably not countable; his theory is unacceptable. Similarly Bronagh's theory is nonconservative since it caps the universe at $\beth_\omega$. Both Dervla and Ewan have massively nonconservative theories: there are no (standard) models of either, since there are no Mahlo-sized or measurable sets. Again both theories are disprovable. Calum's theory, however, is trivially provable and so acceptable.

The obvious difficulty here is that Angus, Bronagh, Dervla, and Ewan can all tell similar stories. 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 finitist Angus, to be sure, might have problems accommodating contemporary science since it seems, to most, to be steeped in commitment to continuum-sized and larger universes. But of course if he insists that intellectual integrity requires us to write off standard physics as an intellectual incoherence which, inexplicably for the moment, works well (compare Berkeley on infinitesimals), the neo-Fregean is in no position to reject this argument on pragmatic grounds of utility for empirical theory lest the Quinean seize on the admission as acceptance of a Quinean epistemology of mathematics. Note, moreover, that though Dervla and Ewan will think that Angus, Bronagh, and Calum place a nonconservative cap on the size of the universe, that is

not how that trio will see things. Assuming that cardinals are sets, in all three of those theories, it is provably the case for every cardinal size there is a larger one. All three theorists can deny that the universe as a whole has a size: for Angus, the notion of $\aleph_0$ as a legitimate number is a myth; it represents rather the absolute infinite; Bronagh holds the same view of $\beth_\omega$, Calum of θ_0.

Do we, then, have an analogue of Embarrassment of Riches returning to haunt us at the metatheoretic level? It might seem not. Even the notions of consistency and consequence are essentially contested. We might find that logics L_1 and L_2 both have proponents; each claims their own logic as a legitimate formalization of the notion of entailment but denies that the other logic is. We could also find that a widely accepted mathematical theory T entails existential consequence E in logic L_1 but not in logic L_2. If a theorist duly deduces E from T using L_1 can she not be said to know E unless she can further prove that there is a distinction between correct and incorrect conceptions of entailment and that L_1 is an explication of the correct notion? Clearly not, this sets an impossibly high standard for justification and knowledge. It cannot, therefore, be held that Calum can only know, innocently, the mathematical consequences he derives from his distraction principle unless he can somehow refute, to everyone's satisfaction, the claims of Angus, Bronagh, Dervla, and Ewan to be providing rival, legitimate positions. To be justified in one's claims regarding some topic one does not have to be able to knock out all other contenders to knowledge in a contest held in some Archimedean arena.

Nonetheless, even in the case of consistency and logical consequence, there is a legitimate concern the neo-Fregean has to answer. If L_2 is not a correct logic then from the neo-Fregean perspective there must be something in the practices of those who use it, or attempt to use it, which prevents its rules from being analytic, meaning-constitutive, or otherwise epistemically innocent. Similarly the rules and principles of L_1 must have this favored epistemically innocent status. The users of L_1 need not be able to demonstrate this is the case. Nor indeed is it necessary that we, the metatheorists, be able to do so either. But if we cannot offer *some* account of what is for one theory to be correct, the other not, then the idea that the existential consequences of T in L_1 limn the true structure of mathematical reality, but the rival ontology extracted from T by L_2 does not, has no plausibility at all.

Only radical Quineans are likely to hold to the thesis that no logical practices can be said to be analytic or meaning-constitutive and that none can be ruled out as devoid of a coherent meaning. However the claim that the full second-order logic invoked by neo-Fregeans is a body of analytic rules or axioms is, as remarked in Section 1, much more contentious. The move from second-order logic to abstraction principles is yet more contentious still. The neo-Fregean who cashes out 'a priori' as something like analytic or meaning-constitutive has to persuade us that it is reasonable to think that among rival abstractionist theorists such as those found in the Angus to Ewan group, at most one principle is analytic or meaning-constitutive. Supposing Ewan does limn the true structure of reality; it must be the case that his inferential practices—in inferring instances of the right-hand side of his distraction principle from the left-hand side and vice versa, for example—are analytic while those of the others are not. The neo-Fregean has to reject the notion that the inferential practices of Angus, Bronagh, Calum, and Dervla are every bit as analytic of *their* notions of class as Ewan's is of his. This, I would argue, is hugely implausible.

The neo-Fregean might, then, construe the a priori nature of mathematical knowledge not in terms of analyticity but in terms of implicit definition. In empirical science we can have two perfectly consistent but pairwise inconsistent theories both satisfiable by (nonisomorphic) abstract structures. Yet only one of them might implicitly define a system of physical magnitudes (and perhaps explicitly define it, if the conditions for Beth's theorem are met) because of the brute empirical fact that a real structure answering to the one exists but not to the other. Can the neo-Fregean hold that, for example, Calum might know, in brute external fashion, that his sets exist and Angus et al. fail to know the same of theirs for no other reason than that Calum's universe is the actual universe of sets, none of the other theorists' universes is?

The danger here, is obvious. How does the neo-Fregean position differ from Quinean holistic empiricism in which mathematical theories are posits which, like the rest of theoretical science, are to be confirmed or disconfirmed only indirectly to the extent that they contribute to a well-confirmed overall theory of the world? In what sense is Calum's knowledge a priori? Had the mathematical universe been different, his mathematical beliefs would have been false, though they would have arisen in exactly the same way.

From a traditional platonistic perspective, of course, this counterfactual is an empty one with an impossible antecedent: the same mathematical universe exists in all possible circumstances. This suggests a possible response by the neo-Fregean. The neo-Fregean might respond by rejecting the claim that acceptability, because it depends on notions of provability and model-theoretic consequence, depends on mathematical notions which stand in need of further justification. The neo-Fregean, might, for example, interpret these notions modally. In so doing, one could argue against Angus, Bronagh, and Calum and so on, on the grounds that they all limit mathematical reality—there *could* be more than a finite, or $\beth_n$, or accessible number of things, and any theory which says otherwise cannot be conservative.[37]

But there are evident problems with this response: if one appeals to a principle of *modal maximality*, 'whatever size could exist, does actually exist in mathematical reality', how on earth is one to represent this mathematically? What abstraction principle will one use? One might demur from providing a single principle and appeal instead to an infinite set of principles: add as many abstractions as one can till one reaches a maximal acceptable set. But why think there will be a unique such set? Even if one eschews uncountable languages and supposes we have a neutral notion of what size a *set* (or perhaps *proper class*) of abstraction principles could be, it is not the case that there is a neutral linear ordering of abstraction principles in terms of the size of the universe they permit as the cases of Dervla and Ewan show.

Most fundamentally of all, though, this modal response owes us an explanation of our knowledge of modality and more generally an account of the nature of modality. How do we know that there *could* be infinite sets? If we do not know this, how can we rule that Angus's finitary theory places illegitimate bounds on the size of the mathematical ontology? Clearly the neo-Fregean making this modal reply cannot analyze possibility as the existence of set-theoretic models since then our supposed knowledge that there could be infinite sets becomes knowledge of the actual existence of sets containing infinite sets and we are back with the problem we started with. Perhaps the neo-Fregean will take modality as primitive. But if she adopts a realist account of modality, we are owed an explanation of how we acquire our

knowledge of what is possible and what is not. A Lewisian type of modal realism would once again bring us back to the very same problems: how do we know there exist causally and spatially isolated possible worlds containing infinite sets—not by intuition surely? Nor is it obvious that rival accounts of modal realism, possibilities as properties of actual reality and so forth, have any better answer to these epistemological problems than Lewis has.

Or the neo-Fregean might analyze necessity and possibility (at least of the type in question in mathematics) in terms of analyticity or kindred notions. A proposition is necessary if it can be derived using only analytic or meaning-constitutive inference rules, or some such. But once again we move from frying pan to fire. I think it is plausible that abstraction principles such as HP, when formulated in rule form, yield rules which are meaning-constitutive of the operators they introduce just as certain types of introduction and elimination rules are arguably analytic for logical operators. But there is nothing to discriminate among abstraction principles in this regard (at least where they are all consistent); they can all be regarded as analytic, in this sense, of the operators they introduce. And to say that some are not genuinely possible, in the *analytic* sense of possibility, because they conflict with the *real* analytic abstraction principles which partially determine what is possible and what is not, once again involves us in a vicious regress.

The neo-Fregean may say that all intellectual argument and discussion must start from some framework of assumptions, even when revising, after the fashion of Neurath in his boat, those assumptions. In our case, the starting point of most philosophers of mathematics is that of a ZFC-like theory, so we are justified in interpreting *stability* and *acceptability* using that theory, even if the theory is a ladder which we kick away when moving to acceptance of an abstraction principle.[38] But if, as the foregoing considerations suggest, any reasonable abstractionist theory we arrive at will itself provide a vantage point from which we can see that many different theories will validate themselves as stable and acceptable and others as unacceptable, how can we justify hewing to the one we have arrived at? Not, surely, because of its closeness to ZFC. How could it be that a theory is a priori true because it fits well with a historically dominant theory which was developed by theorists who almost all rejected neo-Fregeanism and its account of a priori truth?

These considerations then, while they cannot in the nature of the case amount to a conclusive proof that no satisfactory criterion for winnowing out acceptable from unacceptable abstraction principles will emerge, strongly indicate that there is no such criterion which can do the job the neo-Fregean needs it to do: roughly, single out as a priori or epistemically innocent a consistent set of principles which can be interpreted in a semantically homogenous fashion with respect to the empirical part of the physical theories they form part of and which yield classical analysis and the mathematics needed for science.

8 Final Remarks

Even if this is so, however, it does not follow that the neo-Fregean program has accomplished nothing. There may, for example, be significant partial successes. For there may be ways to blunt the above difficulties which capture much of what the neo-Fregean set out to achieve—some less ambitious but recognizably similar program may be one which can be carried through. Among the possible revisions of the neo-Fregean program, the most radical move is to stand one's ground right at the

outset of the sequence of difficulties sketched above and refuse to concede that some abstraction principles are unacceptable. For example, one embraces all abstraction principles, including Axiom V, as meaning-constitutive truths. As remarked at the beginning of Section 2, one must then blame the triviality of the classical naïve set theory not on Axiom V but on the logic which generates triviality and since triviality ensues in fairly weak logics, this option involves quite a radical breach with Frege's thoroughly classical approach to logic. But that in itself is not a refutation. The most developed form of the naïve approach is that to be found in the dialetheism of Priest ([22] and [23]). Priest accepts that Axiom V yields contradiction but concludes that since it is analytically true, so are some contradictions and adopts a paraconsistent logic in order to avoid triviality. But it is not necessary (or at least not obviously necessary) that one embrace true contradictions if one embraces Axiom V: radical enough revisions to the logic will block the derivation of contradiction (cf. Weir [33] and [34]). In both cases, however, one has to show that the revisions are not so radical as to block the derivation of standard mathematics from Axiom V. If either of these naïve approaches could be made to work, they would help toward validating at least one major aspect of the neo-Fregean program, namely, the idea that mathematics follows from meaning-constitutive truths.

There is, however, a less radical way to circumvent the embarrassment of riches objection by embracing equally and without discrimination all abstraction principles and that is to abandon any claim that second-order formal calculi, at least with the full impredicative axiom scheme of comprehension, are logics. Rather one restricts logic to *classical* first-order logic, or perhaps predicative second-order systems and combines logic thus circumscribed with abstraction schemata such as first-order Axiom V.

$$\{x : \varphi x\} = \{x : \psi x\} \leftrightarrow \forall x(\varphi x \leftrightarrow \psi x).$$

This, as Parsons has shown, is consistent and Heck has extended the result to predicative Axiom V in a setting of *predicative* second-order logic (Parsons [21], Heck [15]). The strategy can be extended to show that the set of all first-order abstraction principles is consistent.[39] The drawback here is that the resulting system is rather weak: certainly much weaker than second-order Peano Arithmetic, far less analysis or even the lower reaches of set theory. Nonetheless a theorem of infinity is provable in the system; indeed, as Heck shows, the predicative theory is stronger than the arithmetic theory Q.

A neo-Fregean amending her views in this way could no longer claim that all mathematical truths are analytic or epistemically innocent. She would have to adopt a two-tiered approach. There exists an a priori proof that there are infinitely many (presumably abstract) objects with the properties described in a theory around the strength of Q. As to their further properties, as to whether there are continuum-sized domains of abstract objects, for example, with the structural properties characterized in analysis—here one can only put forward conjectures to be tested by the "fruitfulness of their consequences." As against this one may say that if pragmatic justification is permitted for parts of mathematics why not everywhere? But perhaps conjectures regarding a realm of abstract objects are on a better footing when one has an independent (in this case a priori) justification that the realm of objects itself exists. Nonetheless this type of revision undoubtedly also takes us far from the usual neo-Fregean conception.

A different response is to maintain, as in the radical case above, that all abstraction principles are true but to avoid incoherence not by radical change of logic but by relativizing truth. If one can interpret the abstraction principles, and the existential claims following from them, as true in some sort of mind-dependent fashion, then one can accept each of the principles as analytic and as generating a notional universe, different such universes for different principles. One cannot, classically, amalgamate these universes; but then many antirealists hold that there can be a plurality of mind-dependent domains, domains which are incompatible or incommensurable in some way and so cannot be accumulated or subsumed into a single all-encompassing domain. If one was a realist in general but an antirealist about mathematics in particular then this would yield exactly the right metaphysical position for a classicist who wishes to maintain that all (consistent) abstraction principles (of whatever order) are analytic. No mathematical domains exist in reality but a plurality of often incompatible such domains exist *virtually* (whatever exactly that could mean; clearly there are enormous problems for the view being mooted in explicating this).

Here then we divorce the two strands of neo-Fregeanism—the epistemological and the ontological—distinguished by Hale and Wright (cf. the introduction to [12]). The resulting *antiplatonist neo-Fregeanism* is less vulnerable to any *ontological argument* jibe since there is no commitment to the derivability of *objective* existence claims from concepts alone. The real universe is not at all the same as the notional universes which humans construct; on this view, the question as to the cardinality of the real universe is an absolute one to be answered not by mathematical theory but rather by empirical, nonanalytic theories.

Wright himself toys with something like this nonrealist line of thought:

> we shall have to say that how many objects there are, and hence which objects of which kinds there are, is something which is relative to the scheme of concepts we happen to employ; so that in the abstract realm, our adoption of a particular conceptual scheme affects not merely which objects we shall *recognize* to exist, as in the concrete case, but which objects *actually* exist. That is not perhaps an incoherent view. ([36], p. 293)

He goes on to say, though, that this position "is utterly foreign to the Fregean spirit which the new logicism was supposed to safeguard." Certainly it is foreign to the platonistic strands in Fregean thought; but it may be the only way to safeguard the idea that our justification for our mathematical theories rests not with intuition nor with any indirect, and somewhat precarious, assessment of its utility in science but flows rather from the meaning of the mathematical operators which figure in our theories. The resulting view would perhaps be close to Dummett's in his *Frege: Philosophy of Mathematics* [4]: reference for mathematical terms is a "softer" notion than for nonmathematical terms. Whether this is a reasonable move for a neo-Fregean to make will depend on how dearly held the ontological aspect of Fregeanism is, compared to the epistemological.

However that may be, the conclusion I draw over all is that, in the form in which it is presented by its leading exponents—as vindicating in nonempiricist, non-Kantian fashion, mathematics platonistically construed—neo-Fregeanism is critically wounded by the embarrassment of riches objection; however, the neo-Fregean program has yielded rich insights into mathematical truth and epistemology and less platonistic variants of the program may yet bear fruit.

Notes

1. See the papers in Hale and Wright [12], particularly Hale [13], Wright [36], [37], and [38] (page references are to [12] not to the original articles). See also Hale [10] and Wright [35]. For a different, more constructivist neologicism, see Tennant [28], [29], and [30].

2. The neo-Fregeans concern themselves mostly with *second-order* abstraction principles, in which the right-hand side specifies an equivalence relation over the domain of properties, rather than first-order abstraction principles specified by an equivalence relation over individuals of which Frege's abstraction of identity for directions from parallelism for lines—*Grundlagen*, §§64–65, Frege [8], pp. 74–77—is a well-known example.

3. The term is Boolos's in [1], p. 171 following Frege's rather honorific reference to the Treatise Book I, III.i in *Grundlagen*, §63, [8], p. 73.

4. See Shapiro [24] for an account of standard second-order logic which I take to include Axiom Schemata of Comprehension for predicate formulas of any adicity.

5. Heck [16] shows that Hume's Principle generates a stronger theory than the usual formulation of second-order Peano Arithmetic (with 0 and successor or predecessor) relative to standard bridge principles defining the notions of the one theory in terms of the other. Burgess, Hazen, and Hodes also noted the consistency of the system ([16], fn. 12). Wright notes ([36], p. 273, fn. 4) that Parsons first pointed out in 1964 what Wright [35] later showed in some detail, namely, that Hume's Principle yields second-order arithmetic.

6. Important questions arise concerning the relationship between the knowledge of the mathematical logician who derives R from A in this way and the "ordinary" mathematician (in cases of simple arithmetic, this can be any individual with a basic competence in counting and so forth) who knows R without using any formal logic. However, I leave those questions to one side in this paper.

7. For objections to the claim that the second-order logic needed to gain substantial results from Hume's Principle is epistemically innocent, see Shapiro and Weir [27].

8. So there is a link with Anselm since the objections bear a structural resemblance to objections made against Anselm's ontological proof of the existence of God. Gaunilo of Marmoutier famously objected to Anselm that his proof could be adapted to prove that the most excellent island exists and objectors following Gaunilo claimed that Anselmian arguments could be used to generate existence proofs for too many types of things.

9. Boolos [2], p. 214; see also [3]. For Field see [5], p. 158. Dummett also criticizes Wright on similar lines. He objects to the use of a method which is known, he alleges, to lead to disaster when one has given no principled explanation of the difference between the legitimate and illegitimate uses—a principled explanation amounting to more than just saying that no contradiction seems to follow in the legitimate case; see [4], pp. 188–89, 208.

10. There are set theories with classical background logics in which this holds too, for example, those of Church and Mitchell for which see Forster [7], especially Chapter 4. A more natural such theory, arguably, is Oberschelp's *Set Theory over Classes* [20].

11. More generally one can find sets, or proper classes of principles, such that taken singly or in pairs they are consistent but the set or class as a whole is unsatisfiable. See Fine [6], p. 514.

12. If P is unsatisfiable the principle is logically equivalent to Axiom V; conversely, any model in which P holds can be expanded to satisfy the principle by assigning some one object to $\alpha x Xx$, for every X, so that $\alpha x Xx = \alpha x Yx$ holds universally.

13. Here '$\vdash$' represents provability in standard pure second-order logic with the full impredicative Axiom Schema of Comprehension.

14. Similarly distraction principles with Bad as 'at most α' have models in which there is just one abstract given by the principle, the proper class abstract, at all cardinalities $\leq \alpha$. Where α is finite and nonzero, there is also a bizarre model of size $\alpha + 1$ for Bad = [at most α] in which there are two classes, the *bad* proper class and the *good* universal set. These two are coextensional, with $x \in y$ defined by $\exists F(y = \{x : Fx\} \,\&\, Fy)$, so the axiom of extensionality fails for this distraction principle though it holds for any distraction principle (trivially in this case) when relativized to sets.

15. Here I am defining a strong inaccessible to be a regular limit cardinal $\aleph_\lambda$ with λ a nonzero limit which is such that $\aleph_\lambda > 2^\kappa$ for all $\kappa < \aleph_\lambda$. Define $\aleph_\lambda$ to be nearly strong if and only if the above holds with the last clause amended to $\aleph_\lambda \geq 2^\kappa$ for all $\kappa < \aleph_\lambda$, that is, $\aleph_\lambda$ can be "caught"—but not overtaken—from below using 2^x, that is, the powerset operation. See [26], p. 316.

16. Cf. Weir [32], Appendix I. If we let Bad be inaccessible then subsets will not hold in general: for example, if the cardinality of the model is θ_1, that is, the second inaccessible, then there will be sets of size $< \theta_1$ but $> \theta_0$ which have θ_0 subextensions which are not sets. We can get around this by letting Bad be inaccessible &Big; alternatively we could use ⌜exactly inaccessible$_n$⌝ for Bad. If the generalized continuum hypothesis is true then Bad as exactly α, where α is a regular cardinal, (e.g., $\aleph_{n+1}$) will have models too as there will be exactly α smaller subsets of α.

17. In fact, the semantics of classical first-order logic is only left "unscathed" by mathematics if one accepts some theory such as ZF and uses it to provide the model-theoretic semantics for the logic. Thus intuitionists criticize classical first-order logic for mathematical reasons, taking mathematics to be more fundamental than logic and radical finitists may hold that ⌜there are no more than n things⌝ is a logical truth, for sufficiently high n. Similarly a "finitistic neo-neo-Fregean(!)" who held to the distraction principle P_ω with Bad as [exactly $\aleph_0$-sized] would reject as unintelligible much of first-order model theory (the compactness theorem, for example) since only finite sets of wffs, only finite models exist, and so forth.

18. Such an argument will not impress a Quinean empiricist about mathematics of course.

19. Or if one is unhappy with the use of proper names, replace the consequent 'Clinton is not an adulterer' with, for example, 'everything has zero mass'. HP + θ entails the Σ-restriction of the conclusion, namely, every nonabstract is of zero mass, but θ alone does not entail that everything is of zero mass.

20. See [38], fn. 21, p. 319. Wright's amended requirement, however, seems to me to be too restrictive—he limits unnecessarily the criterion to theories which are consistent with the abstraction principle. But suppose our theory T is an *ultraquantized* scientific theory which holds that the universe contains exactly 10^{50} objects, so one inconsistent with HP. Nonetheless the relativization $T^{\sim M}$ is perfectly consistent with HP; it holds only that there are exactly 10^{50} nonmathematical objects.

21. Syntactic conservativeness and semantic conservativeness are independent of one another since both are formulated in terms of conditionals of the form if $\mathrm{T}^{\sim M}$, P entails $C^{\sim M}$ then T entails C, for the appropriate notion of entailment and to get from one to the other one needs a completeness result for at least one component of the conditional, a result which fails for standard second-order consequence.

22. Perhaps, though, the neo-Fregean can claim that only those sentences provable from analytic principles can be known so that the Gödel sentence for at least one formal proof system must be unknown (but perhaps reasonably believed?) by us. See Shapiro [25] for more on the problems incompleteness results pose for neo-Fregeans who accept the Dummett/Prawitz program of harmony constraints on introduction and elimination rules in acceptable proof systems. See also [12], pp. 4–5, fn. 5.

23. *Logical* abstraction principles, containing no nonlogical vocabulary on the right-hand side of the equivalence, are of this nature—if they hold in one domain of cardinality α they hold in all domains of that cardinality. See [6], pp. 509, 552.

24. This result shows that New V is deductively conservative on the modified Caesar-neutral criterion; however, it is deductively nonconservative on the Field criterion since one can derive global well-ordering WO from it—see [26], §3. For on the Field criterion we restrict WO to $\mathrm{WO}^{\sim M}$ by restricting the domain to the nonmathematical individuals; but we can still prove from New V what cannot be proven outright, that there is a well-ordering over that domain, namely, the restriction of the well-ordering over the universe. But on the modified Caesar-neutral criterion, we restrict to $\mathrm{WO}^{E,F}$ and this now states that there is a well-ordering R *which satisfies the second-order property* F, and which well-orders the domain E; and this we cannot prove from New V. The proof of WO from New V shows that New V is semantically nonconservative on the Field criterion, *if we suppose the falsity of WO*, another example of the nonneutrality of mathematical consequence.

25. Using Scott's "trick" of defining the cardinal $|x|$ of x as the set of all sets of least rank equinumerous with x, a first-order form of Hume's Principle can be derived from ZF, though Lévy [17] proved that Hume's Principle is not syntactically conservative vis à vis first-order ZF minus foundation, or ZF plus arbitrarily many urelements.

26. Wherever possible I will omit the superscripts and subscripts, which are metatheoretic notation indicating order.

27. See [26], §4.2.

28. Probably the neatest way to do this, and to handle variable-binding, is by use of λ terms, but to avoid even more clutter I will forbear from adding those.

29. In the Field criterion case we add β new individuals to create d^* and the function f maps $|D^*_n|$ into $d^* - d$.

30. Note in particular that if B is any abstraction principle of $\mathcal{L}$ true in M, B^* will be true in M^*.

31. In "Implicit definition and the a priori" [11], the authors assimilate abstraction principles, not to primitive inference rules, but to implicit definitions, for instance, of scientific terms. The claim that concerns of 'justificatory modesty' and 'paradox-exploitation' have a role to play here is not as implausible as it would be in the case of primitive inference rules, but is still, I think, implausible. I discuss the appeal to implicit definition a little further below in Section 7.

32. Thus we have a recipe for creating trivial abstractions which are stable from cardinality κ up, where κ is such that there is a formula φ of our language (which will play the *left disjunct role*) true in all and only models $\geq \kappa$.

33. Neo-Fregean approaches to real analysis are to be found in [13].

34. Recall that for simplicity I am identifying properties with extensions: first-order properties are simply subsets of the domain of individuals, and so on.

35. We could extend this into the transfinite by introducing predicates of all ordinal type $< \alpha$, for some fixed ordinal α, and letting $F^{\beta}(G^{\lambda})$ be well-formed where $\gamma < \beta$.

36. Or rather with 'Bad' replaced by a formula which ZFC theorists can translate as 'has cardinality $\beth_\omega$'. Bronagh herself might well reject that translation because she might deny that the universe has a cardinality.

37. A variant on this response is to appeal to Dummett's notion of "indefinite extensibility." We should see the total mathematical universe as an "indefinitely extensible totality" so that no fixed set of abstraction principles captures it. I am very sceptical about the possibility of putting Dummett's notion to any such use: see [32], §3.i.

38. Though of course the favored abstraction principle may entail that ZFC is a correct theory of pure sets, as far as it goes.

39. The basic strategy here is to order the terms of the language and assign each predicate φx its own eigen object in a countably infinite domain; for classes one then assigns $\{x : \varphi x\}$ that object unless φ is coextensive with an earlier term ψ, in which case $\{x : \varphi x\}$ is assigned the same referent as $\{x : \psi x\}$. For any other abstraction principle with its operator $[x : \theta x]$ one proceeds in the same way but substitutes φ bears R to ψ for φ is coextensive with ψ, where R is the equivalence relation on properties generated by the (logical) right-hand side of the abstraction principle for $[x : \theta x]$; so long as one's logic satisfies an extensionality principle as pure second-order logic plus logical abstraction principles does—cf. [6], p. 555—R cannot be more fine-grained than coextensionality.

References

[1] Boolos, G., "Saving Frege from contradiction," pp. 171–82 in *Logic, Logic, and Logic*, Harvard University Press, Cambridge, 1998. Zbl 0972.03502. MR 1376407. 17, 18, 42

[2] Boolos, G., "The standard of equality of numbers," pp. 202–19 in *Logic, Logic, and Logic*, Harvard University Press, Cambridge, 1998. Zbl 0972.03504. MR 1376398. 14, 42

[3] Boolos, G., "Whence the contradiction?" pp. 220–36 in *Logic, Logic, and Logic*, Harvard University Press, Cambridge, 1998. Zbl 0972.03505. MR 1376398. 19, 42

[4] Dummett, M., *Frege. Philosophy of Mathematics*, Duckworth, London, 1991. 41, 42

[5] Field, H., *Realism, Mathematics and Modality*, Basil Blackwell, New York, 1989. MR 92b:03003. 21, 42

[6] Fine, K., "The limits of abstraction," pp. 503–629 in *The Philosophy of Mathematics Today*, edited by M. Schirn, Oxford University Press, New York, 1998. Zbl 0922.03011. MR 2000g:03009. 27, 32, 35, 43, 44, 45

[7] Forster, T. E., *Set Theory with a Universal Set*, vol. 20 of *Oxford Logic Guides*, The Clarendon Press, New York, 1992. Zbl 0755.03029. MR 94b:03087. 42

[8] Frege, G., *The Foundations of Arithmetic. Die Grundlagen der Arithmetik*, Northwestern University Press, Evanston, 1974. Translation by J. L. Austin. Zbl 0041.14701. MR 50:4227. 42

[9] Garland, S. J., "Second-order cardinal characterizability," pp. 127–46 in *Axiomatic Set Theory*, edited by T. Jech, American Mathematical Society, Providence, 1974. Zbl 0319.02065. MR 54:4982. 17

[10] Hale, B., *Abstract Objects*, Blackwell, Oxford, 1987. 42

[11] Hale, B., and C. Wright, "Implicit definition and the a priori," pp. 286–319 in *New Essays on the A Priori*, edited by P. Boghossian and C. Peacocke, The Clarendon Press, Oxford, 2000. 13, 45

[12] Hale, B., and C. Wright, *The Reason's Proper Study*, The Clarendon Press, Oxford, 2001. 41, 42, 44

[13] Hale, B., "Reals by abstraction," *Philosophia Mathematica*, vol. 8 (2000), pp. 100–23. Zbl 0968.03010. MR 2001i:03015. 42, 45

[14] Heck, R. G., Jr., "On the consistency of second-order contextual definitions," *Noûs*, vol. 26 (1992), pp. 491–94. MR 95a:03013. 16

[15] Heck, R. G., Jr., "The consistency of predicative fragments of Frege's *Grundgesetze der Arithmetik*," *History and Philosophy of Logic*, vol. 17 (1996), pp. 209–20. Zbl 0876.03032. MR 98i:03079. 40

[16] Heck, R. G., Jr., "Finitude and Hume's Principle," *Journal of Philosophical Logic*, vol. 26 (1997), pp. 589–617. Zbl 0885.03045. MR 98m:03117. 22, 42

[17] Lévy, A., "The definability of cardinal numbers," pp. 15–38 in *Foundations of Mathematics*, edited by J. Bulloff, T. Holyoke, and S. Hahn, Springer, New York, 1969. Zbl 0182.01203. MR 39:3987. 44

[18] Moore, G. H., "Beyond first-order logic: The historical interplay between mathematical logic and axiomatic set theory," pp. 95–137 in *History and Philosophy of Logic, Vol. 1*, edited by I. Grattan-Guinness, Abacus, Tunbridge Wells, 1980. Zbl 0495.01007. MR 82j:01057. 22

[19] Moore, G. H., "The emergence of first-order logic," pp. 95–135 in *History and Philosophy of Modern Mathematics*, edited by W. Aspray and P. Kitcher, Minnesota Studies in the Philosophy of Science. XI, University of Minnesota Press, Minneapolis, 1988. Zbl 0687.01011. MR 89i:03001. 22

[20] Oberschelp, A., "Set theory over classes," *Dissertationes Mathematicae*, vol. 106 (1973), p. 62. Zbl 0343.02050. MR 47:8300. 42

[21] Parsons, T., "On the consistency of the first-order portion of Frege's logical system," *Notre Dame Journal of Formal Logic*, vol. 28 (1987), pp. 161–68. Zbl 0637.03005. MR 88h:03002. 40

[22] Priest, G., *In Contradiction. A Study of the Transconsistent*, vol. 39 of *Nijhoff International Philosophy Series*, Martinus Nijhoff Publishers, Dordrecht, 1987. Zbl 0682.03002. MR 90f:03007. 40

[23] Priest, G., "What is so bad about contradictions?" *The Journal of Philosophy*, vol. 95 (1998), pp. 410–26. MR 1638747. 40

[24] Shapiro, S., *Foundations without Foundationalism. A Case for Second-Order Logic*, vol. 17 of *Oxford Logic Guides*, The Clarendon Press, New York, 1991. Oxford Science Publications. Zbl 0732.03002. MR 93j:03005. 23, 42

[25] Shapiro, S., "Induction and indefinite extensibility: The Gödel sentence is true, but did someone change the subject?" *Mind*, vol. 107 (1998), pp. 597–624. MR 2000a:03013. 44

[26] Shapiro, S., and A. Weir, "New V, ZF and abstraction," *Philosophia Mathematica. Series III*, vol. 7 (1999), pp. 293–321. Zbl 0953.03061. MR 2000j:03006. 14, 18, 27, 33, 43, 44

[27] Shapiro, S., and A. Weir, "'Neo-logicist' logic is not epistemically innocent," *Philosophia Mathematica. Series III*, vol. 8 (2000), pp. 160–89. Zbl 0966.03002. MR 2001f:03023. 42

[28] Tennant, N., *Anti-Realism and Logic*, The Clarendon Press, Oxford, 1978. 42

[29] Tennant, N., "On the necessary existence of numbers," *Noûs*, vol. 31 (1997), pp. 307–36. MR 98j:03001. 42

[30] Tennant, N., *The Taming of the True*, The Clarendon Press, New York, 1997. Zbl 0929.03001. MR 98m:03007. 42

[31] Weir, A., "Classical harmony," *Notre Dame Journal of Formal Logic*, vol. 27 (1986), pp. 459–82. Zbl 0631.03003. MR 87m:03083. 23

[32] Weir, A., "Dummett on impredicativity," pp. 65–101 in *New Essays on the Philosophy of Michael Dummett*, edited by J. Brandl and P. Sullivan, vol. 55 of *Grazer Philosophische Studien*, Rodopi, Amsterdam, 1998. Zbl 0970.03017. MR 1761354. 43, 45

[33] Weir, A., "Naïve set theory is innocent!" *Mind*, vol. 107 (1998), pp. 763–98. MR 2000b:03024. 40

[34] Weir, A., "Naïve set theory, paraconsistency and indeterminacy. I," *Logique et Analyse. Nouvelle Série*, vol. 41 (1998), pp. 219–66. Zbl 01704401. MR 2002c:03051. 40

[35] Wright, C., *Frege's Conception of Numbers as Objects*, Aberdeen University Press, Aberdeen, 1983. 14, 42

[36] Wright, C., "On the philosophical significance of Frege's theorem," pp. 201–44 in *Language, Thought, and Logic: Essays in Honour of Michael Dummett*, edited by R. G. Heck, Jr., Oxford University Press, New York, 1997. Zbl 0938.03508. MR 2000h:03005. 14, 16, 19, 21, 28, 41, 42

[37] Wright, C., "On the harmless impredicativity of $N^{=}$ ('Hume's Principle')," pp. 339–68 in *The Philosophy of Mathematics Today*, edited by M. Schirn, Oxford University Press, New York, 1998. Zbl 0925.03022. MR 2001f:03024. 42

[38] Wright, C., "Is Hume's Principle analytic?" *Notre Dame Journal of Formal Logic*, vol. 40 (1999), pp. 6–30. Zbl 0968.03009. MR 2002a:03014. 28, 29, 30, 31, 32, 42, 44

Acknowledgments

An early version of this paper was read at the first Abstraction Day meeting at St. Andrews University in November 1998; a later version was read at the same venue in December 2000 under the auspices of the University's Arché Centre. I am grateful to all the participants at those talks and in particular to Bob Hale, Stewart Shapiro, and Crispin Wright for discussions then and on many other occasions. Further comments on later drafts by Julian Cole and an anonymous referee for this journal have also proved extremely helpful.

School of Philosophical Studies
The Queen's University of Belfast
Belfast BT7 1NN
NORTHERN IRELAND
a.weir@qub.ac.uk