12

Implicit Definition and the A Priori

Bob Hale and Crispin Wright

1. INTRODUCTION

An explicit definition aims to supply a semantically equivalent[1] expression of the same syntactic type as its definiendum. Implicit definition, taken as the complement of explicit, embraces a variety of subtypes. What all have in common is the idea that we may fix the meaning of an expression by imposing some form of constraint on the use of longer expressions—typically, whole sentences—containing it. On one traditionally influential conception, this constraint is imposed by the (putatively) *free stipulation* of the truth of a certain sentence, or range of sentences,[2] embedding the definiendum and composed of otherwise previously understood vocabulary.

Our interest here is in a general and a more specific issue about the role and utility of implicit definition. The general issue is whether, and if so under what conditions, the meanings of any significant class of expressions—for instance, logical constants, basic terms of fundamental mathematical theories, or theoretical terms of empirical science—might be constituted by implicit definitions; the more specific issue is whether, if so, such definitions have a role to play in a satisfactory account of the possibility of a priori knowledge of logic and mathematics. We shall refer to the thesis that at least some important kinds of non-inferential a priori knowledge are founded in implicit definition as *the traditional connection*.

We have been much helped by reactions to earlier versions of this paper from and/or discussion of the issues with Jim Edwards, Paul Horwich, Gary Kemp, Jimmy Lenman, Fraser MacBride, Andrew McGonigal, Christopher Peacocke, Adam Rieger, Pat Shaw, and Nick Zangwill. The paper was prepared during Bob Hale's tenure of a British Academy Research Readership; he is grateful to the Academy for its generous support. Crispin Wright gratefully acknowledges the support of the Leverhulme Trust.

[1] Typically, synonymous—but depending on the purposes in hand, some weaker kind of semantic equivalence (such as co-reference) may suffice.

[2] For simplicity, we suppress this qualification in the sequel.

Affirmative answers to both the general and the specific question have found supporters throughout analytical philosophy's first century.[3] In particular, Gentzen's idea, that the meanings of the logical constants should be regarded as implicitly defined by the stipulation of the usual rules for their introduction and elimination in inferential contexts, has been accepted by much of the most important recent philosophical reflection on the epistemology of logic.[4] We ourselves have canvassed a similar, neo-Fregean, view of certain classical mathematical theories as founded on the stipulation of *abstraction principles*—principles like Hume's Principle[5] which share the form of Frege's ill-fated Basic Law V but which, unlike that principle, may reasonably be regarded as consistent.[6] However, there has also been no shortage of antagonists, recent and contemporary. The idea that the holding of certain sentences 'true by convention' might somehow provide a foundation for a priori knowledge generally has been regarded with suspicion ever since Quine's 'Truth by Convention',[7] while 'Two Dogmas of Empiricism'[8] sowed the seed for a widespread scepticism, persisting to this day, not just about analyticity and the a priori but about the very notion of meaning which Carnap and the other early friends of implicit definition thought such definitions might determine.

Quine's more general scepticism is not on our agenda here—though we record the opinion that the two principal lines of argument in 'Two Dogmas' (that the notion of analyticity resists all non-circular explanation, and anyway fails to accommodate the revisability in principle of all statements that participate in total empirical science) respectively impose a quite impossible standard for conceptual

[3] For a very comprehensive discussion, see J. Alberto Coffa, *The Semantic Tradition from Kant to Carnap: To the Vienna Station* (Cambridge: Cambridge University Press, 1991).

[4] For instance, it is supported, notwithstanding important differences, by each of Michael Dummett, Christopher Peacocke, and Paul Boghossian—see e.g. Dummett: *The Logical Basis of Metaphysics* (London: Duckworth, 1991); Peacocke: 'Understanding Logical Constants: A Realist's Account', *Proceedings of the British Academy* 73 (1987), *A Study of Concepts* (Cambridge, Mass.: MIT Press, 1992), 'Proof and Truth' in John Haldane and Crispin Wright (eds.), *Reality, Representation and Projection* (New York and Oxford: Oxford University Press, 1993); 'How are A Priori Truths Possible?' *European Journal of Philosophy* 1 (1993); Boghossian: 'Analyticity Reconsidered', *Noûs* 30 (1996); 'Analyticity' in Bob Hale and Crispin Wright (eds.), *A Companion to the Philosophy of Language* (Oxford: Blackwell, 1997).

[5] That is, the principle—under discussion in Frege's *Grundlagen* §§63–7, where Hume is (somewhat generously) credited with having recognized its correctness—that the number of Fs is the same as the number of Gs iff there is a one–one correlation between the Fs and the Gs.

[6] More precisely, the system got by adjoining Hume's Principle to second-order logic can be shown to be consistent relative to second-order arithmetic. See George Boolos's 'The Consistency of Frege's *Foundations of Arithmetic*' in J. Thomson (ed.), *On Being and Saying: Essays in Honor of Richard Cartwright* (Cambridge, Mass.: MIT Press 1987): 3–20, along with other works to which Boolos there makes reference, and the first appendix to George Boolos and Richard Heck, 'Die Grundlagen der Arithmetik §§82–3', in Matthias Schirn (ed.), *The Philosophy of Mathematics Today* (Oxford: Clarendon Press, 1998).

[7] Originally published in O. H. Lee (ed.), *Philosophical Essays for A. N. Whitehead* (New York: Longmans, 1936).

[8] Originally published in *Philosophical Review* 60 (1951) and subsequently reprinted in Quine's *From a Logical Point of View* (Harvard: Harvard University Press, 1953).

integrity and confuse what is analytic, or known a priori,[9] with what is indefeasibly certain. However, we wholly endorse what we take to be one principal point of 'Truth by Convention': even if conceiving of certain axioms and first principles[10] as implicit definitions may provide the means to explain how a priori knowledge is possible of the truth of the propositions which they express, this could not possibly be the model of *all* a priori knowledge—for it has nothing to offer if we seek to understand our capacity for novel inference and our recognition of hitherto unratified logical consequences.

This limitation needs to be acknowledged from the start. But some recent writers have challenged the capacity of implicit definition to sustain even the epistemologically more modest role of underwriting local cases of the traditional connection. Certainly, it can seem prima facie puzzling how exactly the matter is to work. We take some sentence containing—in the simplest case—just one hitherto unexplained expression. We stipulate that this sentence is to count as true. The effect is somehow to bring it about that the unexplained expression acquires a meaning of such a kind that a true thought is indeed expressed by the sentence—a thought which we understand and moreover know to be true, without incurring any further epistemological responsibility, just in virtue of the stipulation. How does this happen? Paul Horwich, for one, has argued that it doesn't happen: that there is no way whereby, merely by deciding that a sentence containing some hitherto undefined expression is to count as true, we can be assured not merely that that expression takes on a meaning but, more, that the meaning it takes on so configures itself in tandem with our previous understanding of the remaining part of the sentence that the whole expresses a true proposition.[11]

Horwich's scepticism is less radical than Quine's. He believes in meanings, proposes his own account of their constitution in terms of use, and does see implicit definition as playing a role in the determination of meanings. But—and here's the rub—what plays this role, in his view, is our (mere) *acceptance* of the sentences which implicitly define a given expression. This acceptance in turn may contribute towards determining a meaning-constituting pattern of use for that expression. But there can be no a priori likelihood, in Horwich's opinion, that it is acceptance of a *truth.* That is a further matter, to be determined by how things fare with the theory in which the definitions in question participate.[12]

[9] Quine, to the detriment of his polemic, made no distinction between these of course.

[10] By no means all, it should be noted. It would not be plausible, for instance, to see the meanings of any expression as constituted in implicit definition if any matrix for an implicit definition—the previously understood sentence-frame in which the definiendum is placed—would have to contain an equivalent expression. (That will be the fate of expressions for the conditional, on the conception of implicit definition on which we will eventually converge below.)

[11] Cf. Paul Horwich, *Meaning* (Oxford: Clarendon Press, 1998: ch. 6, or his 'Implicit Definition, Analytic Truth and Apriori Knowledge', *Noûs* 31 (1997): 423–40.

[12] Cf. *Meaning*: 8, 143.

This view of the matter seems to incorporate one definite advantage. An apparent drawback of the idea that knowledge of implicitly definitional postulates is a priori knowledge is that it seems to leave no room for the suggestion that it is the very hypotheses of empirical scientific theory which are the prime determinants of the meanings of the theoretical terms which they contain. Horwich's position, by contrast, is designed to allow our mere acceptance of such hypotheses to fit them with a meaning-determining role, without compromise of their being hostage to subsequent empirical evidence. There is a point here to reckon with. Although considerations of space will prevent us from considering the issues to do with empirical theoretical terms in any detail, we accept it as a constraint upon any account of implicit definition designed to underwrite the traditional connection that it should somehow match this apparent advantage[13] of Horwich's account—that it should be broadly consonant with the idea that it is via their role in empirically revisable scientific theories that theoretical terms (at least partially) acquire their meaning.

Our programme in the sequel is as follows. In Section 2, we will review Horwich's objections to the traditional view of implicit definition as underwriting at least some a priori knowledge. It seems to us that the critical thrust of these objections entirely depends upon a certain model of what, if the traditional conception of implicit definition[14] is to be sustained, such definitions must accomplish—a model which we believe should be rejected. In Section 3 we shall outline a different conception of how the traditional connection had best be supposed to work and, in Section 4, begin to explore some of its ramifications for the metaphysics of meaning, specifically, for the question: when does fixing a use explain a meaning? In Section 5, we shall—all too briefly—sketch an approach whereby a conception of implicit definition which underwrites the traditional connection might also, without compromise of their empirical revisability, allow scientific theories themselves a role in determining the meanings of their distinctive vocabulary. Finally, in Section 6 we will connect that discussion with objections proposed by Hartry Field and Kit Fine[15] to our preferred view of the status of neo-Fregean abstraction principles.

[13] 'Apparent' because we don't think Horwich actually does secure this advantage—see Sect. 5 below.

[14] i.e. as proceeding through stipulation of the truth of some sentential context(s) of the definiendum.

[15] Hartry Field, 'Platonism for Cheap? Crispin Wright on Frege's Context Principle', *Canadian Journal of Philosophy* 14 (1984): 637–62, repr. in Field's *Realism, Mathematics and Modality* (Oxford: Blackwell, 1989): 147–70; Kit Fine, 'The Limits of Abstraction' in Matthias Schirn (ed.), *The Philosophy of Mathematics Today* (Oxford: Clarendon Press, 1998): 503–629.

2. A MISGUIDED MODEL

Horwich asks:

> How exactly could it come about that our regarding a certain (perhaps conjunctive) sentence '#*f*' as *true* would provide the constituent '*f*' with a meaning, and what exactly is the meaning that '*f*' would acquire?[16]

He supplies on behalf of the view he wishes to oppose the answer that 'the decision to regard "#*f*" as true is, implicitly, a decision to give "*f*" the *meaning it would need to have* in order that "#*f*" be true'.[17] A number of problems are then, of course, immediately suggested by the reference in this answer to '*the* meaning "*f*" would need to have for "#*f*" to be true'. What ensures that there *is* any such a meaning? And what ensures that there is *just one* such meaning? For instance, the possibility cannot be discounted that, since the procedure of implicit definition assumes that the matrix '#–' already has a specific meaning, there simply is no meaning that '*f*' could take that would render '#*f* true; '# –' might be such that no matter what meaning '*f*' had, '#*f*' would be false. One not very interesting case would be if '#_' were a conjunction one of whose conjuncts is a closed and false sentence—Horwich offers 'Snow is green and the moon is _'. A more interesting, classic example would be provided by the attempt to fix the meaning of a functional expression, intended to denote a function from properties to objects, by means of the matrix:

$$\forall F \forall G(\ldots F = \ldots G \leftrightarrow \forall x(Fx \leftrightarrow Gx))$$

from which, by applying the matrix to the predicate R defined by $Rx \leftrightarrow \exists F(x = \ldots F \wedge \neg Fx)$, we can derive Russell's contradiction: $R(\ldots R) \leftrightarrow \neg R(\ldots R)$. Here it would obviously be of no avail to attempt to stipulate that '*f*' is to mean just what it needs to mean in order to render:

$$\forall F \forall G(f(F) = f(G) \leftrightarrow \forall x(Fx \leftrightarrow Gx))$$

true.

One might quite properly say that in this and similar cases (for instance, Prior's connective 'tonk') that there *is* no meaning for '*f*' which renders '#*f*' true, because there *can* be none. But this way of expressing the matter has dangerous associations, to which Horwich succumbs. The danger is that of slipping into a picture—which will strike anyone but an extreme realist about meanings as tendentious—of implicit definition as aimed at hitting on some (unique) pre-existing meaning, as if it were akin to an attempt at *reference-fixing.* If some London policeman in the 1890s were to say, 'Let's call the perpetrator of these

[16] *Meaning*: 132. [17] Ibid. 132–3 (our emphasis).

ghastly crimes, "Jack the Ripper", the success of his proposal would naturally be hostage to the truth of the assumption that there was indeed a (unique) perpetrator of the murders—that the grievously disfigured corpses were not the results of some bizarre series of accidents, for instance. Likewise when someone stipulates, 'Let "*f*" have whatever meaning it would need to have in order that "#*f*" be true', the proposal is hostage to the possibility of there *being* such a meaning—a hostage that may not be redeemed in certain cases, as just noted. But the analogy obscures the point that, for any but the extreme realist, the existence of an appropriate meaning is not an antecedent fact of metaphysics, so to speak, but—(to oversimplify horribly; qualifications to follow)—a *matter of our intent*. In brief, it is a matter not of hoping that a platonic realm of meanings can put forward a denizen to satisfy the condition expressed by the stipulation but a matter merely of whether the stipulation *succeeds in avoiding certain operational drawbacks*—of which inconsistency is one (but only one) glaring example. If it so succeeds, our intention to fix a meaning by the stipulation will suffice to ensure that there *will be* a suitable meaning—there is no additional requirement of cooperation by a self-standing realm of meanings. This broad orientation is, we believe, crucial if justice is to be done to the traditional connection. We will prefigure some of the necessary detail below.

To forestall misunderstanding: we are not denying that implicit definitions have any genuine 'existence problem' to surmount. But if the problem were that of getting a guarantee that such a definition somehow successfully alighted upon a preconstituted meaning, it is utterly obscure how it might be solved—except by finding an *interpretation* of the definiendum: some paraphrase in independent semantic good standing (which would then of course make a mere detour of the implicit definition—this crucial point will recur in the sequel). The genuine issue of existence is rather whether the implicit definition serves so to direct and constrain the use of the definiendum that it can participate as an element in a successful communicative practice. This broadly Wittgensteinian orientation is, of course, barely a first step towards understanding how implicit definition should work. But it is curious that Horwich, who officially proposes a 'use theory' of meaning, should sustain objections to the traditional connection which seem to presuppose something quite at odds with Wittgenstein.

It is similar with what Horwich calls 'the uniqueness problem'. Again, the stipulation above for 'Jack the Ripper' is hostage to the existence of a *single* perpetrator of the crimes; if they were the work of a team, or a number of copycats, or coincidence, the stipulation fails to fix a referent. Likewise according to Horwich's view[18] when we say that '*f*' is to have the meaning it needs to have for

[18] That is, according to the view Horwich takes of what has to be the case, if the traditional conception of implicit definition as proceeding through stipulation of the truth of some sentential context(s) is to be upheld—we are not, of course, speaking his own 'use-theoretic' view: the view about how implicit definition works which Horwich himself espouses and sets against the traditional conception. Our subsequent use, throughout this section, of 'Horwich's model of

'#*f*' to be true, we presuppose not only that there is at least one meaning such that if '*f*' has that meaning, '#*f*' is true, but that there is *at most* one such meaning. One who proffers an implicit definition seeks to single out that unique meaning whose interpolation, so to speak, into the incomplete thought expressed by '#–' results in a true proposition. If things go well, there will be one and only one such item in the realm of meanings. But things may not go well—there may either be no item in the realm of meanings which can combine with the meaning of '#–' to give a true proposition, or there may be more than one. The uniqueness problem is then conceived as the problem of somehow getting an assurance against the latter eventuality. It is not easy to see where such assurance might come from.

This sense of the problem seems again to be largely inspired by a combination of platonist imagery and a misconceived analogy between implicit definition and reference-fixing. As with the 'existence problem', we do not dispute that there is also a *potentially* genuine issue near by—one detachable from Horwich's model. This issue, of course, is that of *indeterminacy* and it challenges not just implicit definition, as traditionally viewed, but all forms of definition and concept-fixing. A good explanation of the meaning of an expression must place real but satisfiable constraints on the explanandum. An attempted definition may thus be flawed by overconstraint—unsatisfiability—or by *underconstraint*; for instance—to take the limiting case—if '#–' is such that '#*f*' is *bound to be true*, whatever '*f*' means. An implicit definition would fail in this way, if we inadvertently chose for '#–' something which was, in fact, a valid schema—in that case, the attempted definition would abort because it would place no restriction, beyond that implicit in the syntactic type of '*f*', on what that expression could mean. But less dramatic degrees of failure in the same dimension are obviously possible; a definition may impose some but insufficient constraint on the use, or interpretation, of the definiendum. (Consider for instance the stipulation, 'Some plane figures are *f*.)

We shall return to the issue of sufficiency of constraint in Section 4, but two brief observations are in order now. First, there is no absolute level of determinacy, independent of our purposes and the particular context, to which explanations of meaning must attain, if they are to be adequate. What should count as *insufficient* constraint on a definiendum is relative to the context and purposes of the definition. It is no objection to an implicit definition that it fails to discriminate among various more specific interpretations of its definiendum, if context and purposes do not demand finer discrimination. Second, no more determinacy can reasonably be demanded of meanings purportedly fixed by implicit definition than belongs to meanings in general. Any suggestion that a special 'uniqueness

implicit definition', 'the conception of implicit definition with which Horwich is working', and cognate phrases should likewise be understood as making reference to his interpretation of the presuppositions of the view he is attacking, not his positive view. In our view, the opposition Horwich sets up between the traditional conception of implicit definition as proceeding through stipulation and a conception of it as working by fixing a pattern of use for the definiendum is a false one.

problem' attends implicit definition, as traditionally viewed, accordingly owes an argument that some special and pernicious sort of indeterminacy afflicts the attempt to fix meanings by implicit definition—something unmatched by other modes whereby meanings are learned, including the most basic and informal instruction we each receive in our mother tongues. Such an argument, therefore, must not merely consist in the application to implicit definitions of considerations which are easily adapted to the cause of a more general meaning-scepticism. No such specific case has, to the best of our knowledge, ever been made.

The existence and uniqueness problems are two of four objections which Horwich lodges against the traditional account. The others are what he calls the 'possession problem' and the 'explanation problem' respectively. He writes:

> Even if there exists one and only one meaning '*f* [*sic*] to have for '#*f* to be true, can we be sure that '*f* comes to possess it? And supposing it does come to possess it, what would explain how this happens?[19]

With only marginally less brevity, he 'elaborates' the possession problem thus:

> In general, if you want to give a particular object a particular property, it will not be enough to say 'Let it have that property': rather (or, in addition), you have to do something. You can't make a wall red just by saying 'Let it be red'—you have to paint it. Well, this point applies to meaning properties too. You cannot give a word a certain meaning by declaring, however earnestly or passionately, that it has that meaning. Something more must be done.[20]

Since these are meant to be new and additional problems, we can assume the case is one for which existence and uniqueness are resolved; a case where we somehow have assured ourselves that the expression, 'that (unique) meaning which suffices for "#*f*" to express a truth', does have a unique, determinate referent. The alleged further difficulties would then seem to have to do with exactly how it may be brought about by stipulation that '*f*' takes on *that* item as its meaning—or with explaining how the trick was pulled if it does. But why should either point give rise to any difficulty?

We suggest not merely that there is no such further difficulty but that to suppose otherwise is actually inconsistent with the 'reference-fixing' model of implicit definition in force in Horwich's critique of the traditional conception of such definition as proceeding through stipulations of truth. Consider again the stipulation of our imaginary Victorian detective for 'Jack the Ripper'. If there are no problems about existence or uniqueness—if the crimes in question are indeed the work of a single killer—how could the stipulation fail? (We need not bother with the irrelevant possibility that it fails to be taken up by others). What residual issues could there be about what makes it the case that 'Jack the Ripper' has the referent it does, or about what explains the connection between the name and the referent? By hypothesis, that referent is the unique object satisfying the reference-

[19] *Meaning*: 134. [20] Ibid. 134–5.

fixing condition that was agreed on for the name. What further problem can there possibly be?

Well actually, there *is* a further problem for the reference-fixing model—though it is not clearly identified in Horwich's discussion. Grant that the detective's stipulation fixes a referent for 'Jack the Ripper'. Still, he is no nearer knowing *who the Ripper is*. But then the stipulation of '#*f*' as true, conceived as working in the way Horwich supposes it must work if the traditional view is to be viable, could indeed fix a *referent* for the expression, 'that (unique) meaning which'—assuming that '#–' retains its prior meaning—'suffices for '#*f*' to express a truth', without making it clear *what meaning that is*. So someone who *perfectly understood the stipulation* might yet not achieve an understanding of '*f*'. And that is to all intents and purposes for the implicit definition to fail—for in that case it merely *fixes* a meaning without explaining it. The real residual problem for a proponent of Horwich's model of implicit definition is thus not, as he supposes, to explain the connection which the stipulation establishes between '*f*' and its meaning—that is no more problematic than in the case of 'Jack the Ripper'—but rather to make out the capacity of implicit definitions genuinely to *explain* meanings, to impart understanding. We might call this the *understanding* problem.

It is a critical problem—on the conception of implicit definition with which Horwich is working. For think what would be required to resolve it in a particular case. In order to identify the referent of 'Jack the Ripper', explained as indicated, an agent will have to determine who committed the crimes in question, that is, will have to bring the perpetrator under some canonically identifying concept: 'the fifth from the left in that line-up', for instance, or 'the Prince of Wales'. By analogy, in order to arrive at an understanding of the meaning of '*f*', explained by its implicit definition as conceived by Horwich, a thinker will have to bring the referent of 'that (unique) meaning which suffices for "#*f*" to express a truth' under some canonically identifying concept: that is, identify it as the referent of some expression whose own meaning in turn somehow serves to disclose what meaning it is. But that is tantamount to the demand that successful, implicit definition requires a recipient to have—or to have access to—*independent* resources sufficient for an *explicit* definition of the definiendum. Yet it was all along an absolutely crucial point about implicit definitions, as traditionally conceived, that they were to serve in cases—fundamental mathematical and logical concepts, and scientific-theoretical terms—where no resources for (non-circular) explicit definition were available. Horwich's model is thus implicitly in tension with this absolutely crucial point. We take that to be a decisive objection to it.

In summary: suppose that the matrix '#–' is already understood and that '#*f*' is laid down as an implicit definition of '*f*'. In Horwich's view, the definitional purport of this performance—on the traditional conception he opposes—is best captured by modelling it on the resolution: '*f*' is to have that (unique) meaning

which suffices for '#*f*' to express a truth.[21] Horwich uses this picture of the matter to generate four general difficulties for what he regards as the traditional view of implicit definition, and for its capacity to contribute towards the explanation of non-inferential a priori knowledge. Two of these problems—existence and uniqueness—do indeed arise on the model in question, and we contend that their solution—in the form in which they so arise—is to reject that model. The remaining pair—possession and explanation—are, we have suggested, spurious even on the terms of the model. However there is a further genuine problem which arises on the model—the 'understanding' problem—whose solution will once again be, we contend, to give up the model itself.

These conclusions are wholly negative. The discredited model has it that implicit definition is a kind of reference-fixing, whose candidate referents in any particular case compose a population of predeterminate meanings. The questions must arise: if that model is askew, then with what should it be opposed—what is the right way of conceiving of the workings of a successful implicit definition—and can the traditional connection with a priori knowledge still be vindicated? One clear desideratum to have emerged is that a satisfying account of explanation via implicit definition must leave room for the capacity of such explanations to *invent* meanings, to bring us to a competence with concepts which we previously simply did not possess and for which we have no other means of expression save by the terms implicitly defined. Since a satisfying account of how else a successful implicit definition might achieve that effect must presuppose a proper characterization of the effect—our coming to a genuine understanding of the definiendum by way of grasp of a novel concept—the issue raised by the first question is, potentially, a very large one, demanding no less than an account of when it is correct to regard an expression as genuinely possessing a meaning in the first place. While that is, naturally, too large an issue for the present discussion, it is possible to outline at least some of the constraints which it seems clear that a good implicit definition, viewed as introducing a novel concept, should observe. That will be a task for Section 4. First we need to consider afresh the question how, in general terms, the traditional connection between non-inferential a priori knowledge and implicit definition should be conceived as working.

3. THE TRADITIONAL CONNECTION

How, just by stipulating that a certain sentence, '#*f*', is true—where '#–' is already understood, and '*f*' is a hitherto contentless expression determined only to be of some specific syntactic category apt for a well-formed completion of '#–'

[21] Where this in turn is construed platonistically, along the lines suggested by the reference-fixing model.

—is it supposed to be possible to arrive at an a priori justified belief that #*f*? Well, the route seems relatively clear *provided* two points are granted: first that a stipulation of the truth of the particular '#*f*' is so much as properly possible—that the truth of that sentence is indeed something which we can settle at will; and second that the stipulation somehow determines a meaning for '*f*'. If both provisos are good, it will follow that the meaning bestowed on '*f*' by the stipulation cannot be anything other than one which, in conjunction with the antecedent meaning of the matrix, '#–', results in the truth of the sentence in question. For it is the very stipulative *truth* of the sentence which determines what meaning '*f*' acquires. Moreover, if the stipulation has the effect that '*f*' and hence '#*f*' are fully *understood*—because we now understand not merely the matrix but the definiendum, and the significance of their syntactic combination—then nothing will stand in the way of an intelligent disquotation: the knowledge that '#*f*' is true will extend to knowledge that #*f*. In other words: to know both that a meaning is indeed determined by an implicit definition, and what meaning it is, ought to suffice for a priori knowledge of the proposition thereby expressed.[22]

Everything thus depends on the two provisos noted: that it is—at least sometimes—possible to bring it about that an antecedently partially understood sentence is true just by stipulating that it is to be so, without further explanatory or epistemic obligation, and that in at least the best such cases, the result is something which does indeed *fully explain* a meaning for the definiendum. Both provisos are clearly substantial. We consider them in turn.

For all we have so far said, '#–' can be *any* previously understood matrix and '*f*' *any* expression of a nominated kind which is syntactically appropriate to generate a well-formed sentence if appropriately introduced into '#–'. As stressed

[22] This model of how a stipulation that a given partially understood sentence is true may eventuate in a priori knowledge of a proposition expressed by it is offered as a piece of reconstructive epistemology, not a psychological hypothesis. As a parallel, consider the situation when someone with a certain item of knowledge is inclined to accept certain of its—perhaps not wholly obvious—consequences seemingly spontaneously, without being able to reconstruct an explicit derivation of them. Even if the most he can offer by way of explicit justification is quite lame, we should not scruple to regard these consequential beliefs of his as knowledge just on that account: it should be enough if it is plausible that his general intelligence is somehow sensitive to the obtaining, or not, of the relevant consequence relations and if a fully explicit account of them is available in principle. Likewise, the actual phenomenology of a priori knowledge of first principles to which the proposed model applies might bear little relation to the explicit detail of the model, involving no more than immersion in a practice in which the implicitly definitional sentences are unswervingly accepted and the dawning of simultaneous impressions of comprehension of the definiendum and of the evident correctness of those sentences. It would nonetheless be appropriate to regard this as involving a priori knowledge, justifiable along the lines illustrated by the model, provided the practice of unswerving acceptance of the sentences in question was best construed as effectively stipulative and, so conceived, the stipulations in question satisfied such additional constraints on successful implicit definition as a full working-out of the model will involve. (In recently fashionable terminology, this is to suggest that the model reconstructs an *entitlement* to knowledge, rather than a justification which ordinarily competent thinkers would need to be able to offer in order to count as knowledgeable.)

above, not every such pair serves the construction of a sentence which we are at liberty, consistently with our prior understanding of '#–', merely to stipulate to be true. Trivially, '#–' may be such that no well-formed completion of it by something of the syntactic type of '*f* *can* express a truth (it may be a contradictory predicate, for example, and '*f* a term). More interestingly, the syntactic type of '*f* may demand that the truth of '#*f* carries some existential or referential implication which cannot be vouchsafed just by a stipulation—let '*f* be 'Jack the Ripper', for instance, and let '#–' be '– is the perpetrator of this series of killings'. To attempt to lay it down a priori that the sentence, 'Jack the Ripper is the perpetrator of this series of killings', is to express a truth would be merely presumptuous—for we could have no a priori entitlement to the presupposition that 'the perpetrator of this series of killings' refers at all. The properly modest stipulation would rather be that 'If anyone singly perpetrated these killings, it was Jack the Ripper' (read so that its gist is just that of 'We hereby dub the perpetrator, if any, of all these killings: Jack the Ripper.')[23]

Let us call *arrogant* any stipulation of a sentence, '#*f*', whose truth, such is the antecedent meaning of '#–' and the syntactic type of '*f*', cannot justifiably be affirmed without collateral (a posteriori) epistemic work. The traditional connection, then, between implicit definition and the a priori requires that at least some stipulations not be arrogant. How can we circumscribe the arrogant ones so as to hive off a remainder which are safe for cost-free stipulation?

A natural way of trying to capture the moral of the Ripper example would be to say that like any purported definition, implicit definition can only bestow *sense*, not reference: we may, if we please, decide to confer a meaning of a certain kind on a term, but we cannot just by a decision confer a *reference* on a term—for that, we need in addition that a cooperative world puts forward an appropriate referent. However given any semantic framework which, like Frege's, extends the notion of reference or semantic value—*Bedeutung*—to other kinds of expression besides singular terms, this cannot be a happy way of putting the point. If we determined, for instance, that 'lupina' was to mean the same as 'female wolf',

[23] 'Jack the Ripper', so introduced, would qualify as what Gareth Evans called a Descriptive Name (compare his own 'Julius' and 'Deep Throat'—for which see *The Varieties of Reference* (Oxford: Oxford University Press, 1982): 31–2, 35–8, 47–51, 60–1). In Evans's view—and in contrast with the descriptions by which they are introduced—such names combine the feature of rigid designation (Evans would have preferred simply 'designation') with—like those descriptions and unlike what he calls Russellian names—the possession of a sense that can survive failure to refer. A theorist who regards this combination of properties as in tension will be unhappy with the suggestion that the Jack the Ripper stipulation can be saved from presumption by conditionalization—unless, of course, 'Jack the Ripper' is conceived as a mere shorthand for its reference-fixing description. But we here assume the broad correctness of Evans's view of such names. It is implausible to suppose that, in the event that nobody murdered the women in question, the intelligent consideration of sentences embedding 'Jack the Ripper' so introduced would involve the mere illusion of thought; but the modal properties of thoughts embedding the name are such that it is equally implausible to think of it as a shorthand description.

then that would suffice for the sentence 'A female wolf is a lupina' to express a truth. And in order for that to be so, according to any broadly Fregean semantic framework, the common noun, 'lupina' would have to have acquired a *Bedeutung* as well as a sense. So a stipulative decision—in this case an explicit definition—*can* after all confer reference, as well as sense. That the meaning of '#–' and the syntactic type of '*f*' are such that the truth of '#*f*' would demand that '*f*' has a *Bedeutung* need not of itself require that the stipulation of that sentence as true would be arrogant.

What this reflects is the, plausibily, different relationship in which sense stands to reference in the respective cases of singular terms on the one hand and other kinds of expression—predicates, relations, quantifiers, and functional expressions—on the other. In the case of the former, fixing a meaning must involve establishing a condition on the identity of the referent if any; but (as in the case of definite descriptions and abbreviations of them) establishing such a condition may *fall short of* establishing an actual reference, or it may (as arguably with most proper names and demonstratives) *presuppose* successfully establishing a reference. In neither scenario can one permissibly just stipulate as true any sentence whose truth would require that the term in question refers. But with incomplete expressions, the suggestion may be, the matter stands differently: here to establish a meaning is to establish a *Bedeutung* too—there is no question of our attaching a clear satisfaction-condition to a predicate, for instance (or a clear condition on the identity of the value of the function denoted by an operator for a given argument)—yet somehow failing to supply such expressions with a *Bedeutung*. But nor, in the case of such incomplete expressions, is the sufficiency of sense for possession of a *Bedeutung* to be compared to the situation with most proper names, where meaning is conferred *via* establishing a referent. Rather—whatever *Bedeutung* is held to consist in for such expressions—one automatically confers a *Bedeutung* upon them by settling their meaning (contrast the situation with Russellian names where one confers a meaning—sense—upon them by settling their *Bedeutung*).

It would take us too far afield to review this disanalogy properly. But even if we grant that it is correct, what does it tell us about the demarcation of arrogance? Clearly it will not suffice for the stipulation of '#*f*' to avoid arrogance that neither '*f*' nor '#–' be a singular term. There are many counter-examples: for instance, a set of legitimate stipulations serving to explain the satisfaction-conditions of a hitherto undefined predicate would become arrogant if merely enlarged by the additional stipulation that something falls under that predicate. On the other hand, we had the example, not withstanding its embedded occurrence of the singular term 'Jack the Ripper', of the arguably perfectly legitimate stipulation:

> If anyone singly perpetrated all these killings, it was Jack the Ripper.

Here the crucial point—what makes the stipulation acceptable—has to do not with any special relationship between sense and reference characteristic of

expressions other than singular terms—for the example precisely concerns a singular term—but with the fact that the stipulation is *conditional* in such a way that no commitment to successful reference on the part of the definiendum is entrained.

One attractive general suggestion[24] would be that all *any* definition—implicit or otherwise—can legitimately (non-arrogantly) do is to fix necessary and sufficient conditions on the identity of the *Bedeutung*, if any, of its definiendum: to determine how it has to be with an entity of the appropriate kind if it is to be the *Bedeutung* of the defined expression. That much may then suffice or not suffice, depending on the syntactic type of that expression, for the existence of such a *Bedeutung*. What specific constraints this suggestion would impose on the implicit definition of predicates, functional expressions, quantifiers, and so on will naturally turn on how one conceives of the *Bedeutungen* of such expressions. But one would expect that a stipulation of *satisfaction-conditions* would be the general pattern—one would stipulate what it would take for an object to satisfy the predicate in question, or for a concept to satisfy the quantifier in question, or for an object to satisfy any definite description formed by completing the functional expression by a term standing for an entity in the relevant range of arguments. All of these stipulations could be given in the form of conditionals. So since a non-arrogant stipulation of the condition on the identity of the referent of a singular term will likewise naturally assume a conditional shape—as witness the Ripper example—our working suggestion is going to be that, in order to avoid arrogance in implicit definition, and irrespective of the syntactic type of definiendum involved, it will in general be sufficient to restrict attention to sentences which are appropriately *conditional*.

Schematically, let S(*f*) be a sentence (or type of sentence) embedding one or more occurrences of the definiendum '*f*'. Outright, categorical stipulation of the truth of S(*f*) may well be arrogant.[25] There will, however, be no arrogance in what we shall call the *introductory stipulation* that the truth of some other sentence or sentences (of specified type(s)), S_I, is to suffice for that of S(*f*). Further—provided that our stipulations taken as a whole are conservative, in a sense shortly to be explained[26]—there will, likewise, be no arrogance in the *eliminative stipulation* that the truth of S(*f*) is to be itself sufficient for that of a certain other sentence or sentences (of specified type(s)), S_E. Equivalently, we may—subject to the same proviso—non-arrogantly stipulate the truth of *introductory* and/or *eliminative conditionals*, $S_I \rightarrow S(f)$ and $S(f) \rightarrow S_E$.[27] A simple example is provided

[24] We are here indebted to Christopher Peacocke.

[25] It *need* not be arrogant, since S(*f*) may be already conditional in form, or equivalent to a conditional.

[26] See the opening paragraphs of Sect. 4, below.

[27] In the special case $S_I = S_E$, implicit definition of *f* may assume the form of a biconditional stipulation—as with implicit definition of the direction-operator by stipulation of the Direction equivalence: 'The direction of line *a* = the direction of line *b* ↔ lines *a* and *b* are parallel.'

by the obvious introductory and eliminative stipulations for material conjunction—that the truth of *A* together with the truth of *B* is to suffice for that of '*A* and *B*', and that the truth of '*A* and *B*' is to suffice for that of *A*, and likewise for that of *B*.[28]

Now to the second proviso demanded by the traditional connection: that the stipulation somehow determine a meaning for its hitherto undefined constituent. The crucial point here is the correct reading of 'determine'. One lesson of our consideration of Horwich's discussion, most especially of the 'understanding' problem which emerged, was that the success of an implicit definition cannot consist merely in its *indirect* specification of a suitable meaning for its definiendum. The process of stipulation must somehow be conceived not merely as fixing the truth of the stipulatum but as *conveying* a meaning perfectly adapted to ensuring *both* that '*f*' so contributes to the meaning of '#*f*' that that sentence comes to express a truth *and* that we understand what meaning that is. Here 'conveying' a meaning cannot merely involve: making it obvious how the definiendum is to be *interpreted*—where interpretation is a matter of translation into some independently accessible vocabulary. If that were all that was possible, then when '*g*' was such an interpretation of the definiendum, the proposition expressed by '#*f*' could be as well expressed by '#*g*', and the recognition of the adequacy of the translation would presumably depend on *antecedent* knowledge of the truth of the latter. So absolutely no connection would have been made between the stipulative introduction of '*f*' and the possibility of a priori knowledge of that proposition. It follows that if such a connection can indeed be made, then the implicit definition of '*f*' must convey its meaning not in the sense of making it obvious and inescapable how to interpret it—how to locate its meaning among those expressible in an independently available vocabulary—but in the sense of explaining that meaning to someone who has no other means of expressing it and is, in that sense, so far innocent of the conceptual resource which the implicit definition affords.

The traditional connection thus demands that a good implicit definition can *invent* a meaning—that it is at the service of a *first-time construction* of a meaning, one which may but need not be accessible in other ways—where to invent a meaning, we suggest,[29] can only be to bring it about that some expression has a novel but intelligible pattern of use, perhaps unmatched by any expression either already available in the language or explicitly definable by means of such. To invent a meaning, so conceived, is to fashion a concept: it is to be compared to making a mould and then fixing a certain shape-concept by stipulating that its

[28] The introductory conditional in this case may be taken as either '$A \rightarrow (B \rightarrow (A \text{ and } B))$' or '$B \rightarrow (A \rightarrow (A \text{ and } B))$' assuming permutation of nested antecedents of conditionals; there will be two eliminative conditional schemas: '$(A \text{ and } B) \rightarrow A$' and '$(A \text{ and } B) \rightarrow B$'. Note that each of *A* and *B* may contain occurrences of 'and', although they will not do so in the basic cases; the example illustrates why it would be wrong to require S_I and S_E to be free of occurrences of *f*.

[29] The suggestion is, of course, entirely in the spirit of Horwich's own official view.

instances comprise just those objects which fit the mould (or are of the same shape as something which does). There is a sense in which the shape—the bare possibility of matter so configured—existed all along. We did not create *that* possibility. But we did create a concept of that shape (whether or not we also fixed the meaning of a word to be associated with it). It would make no sense for someone who followed the performance to doubt that there is any such shape—we displayed the shape in fixing the concept of it. In rough analogy, we must so conceive implicit definition that—in the best case—it makes no sense to doubt that there is a meaning taken on by the defined expression, not because the meaning in question allows of independent specification but because it has somehow been *fully explained* in the very process that creates it. Such an explanation can only consist in the fact that the implicit definition determines—or plays a part in a more general framework which determines—a pattern of use which is *fully intelligible without further interpretation*.[30]

Under what conditions would an explanatory stipulation achieve that end? Our approach in what follows will be piecemeal: to list a range of pitfalls—a range of specific ways in which an attempted stipulation of use might fail to fix anything which intuitively amounted to a novel meaning. Our underlying suggestion will be, of course, that an implicit definition which avoids these pitfalls will indeed have what it takes to instantiate 'the best case' of use-fixing, and thereby to sustain the traditional connection with non-inferential a priori knowledge. But we will not attempt to prove the sufficiency of our proposals; our project here is merely to outline a conception of implicit definition which retains the idea that it works through stipulation of the truth of suitable sentential contexts and (thus) has a chance of sustaining the traditional connection.

4. KNOWING A USE AND UNDERSTANDING A MEANING

When does a definition—of any kind—so fix a use that it genuinely explains a meaning?

The first requirement is, evidently, that of *consistency*. The inconsistency in Frege's Basic Law V, for instance, has the effect—assuming an underlying logic which, like classical or intuitionist logic, sustains *ex falso quodlibet*—that no particular use of the terms it introduces can be stably defended as correct (except in a sense of 'correct' which allows both an assertion and its denial to be correct). No doubt it would be too brisk to insist that any such inconsistent stipulation must fail *altogether* to fix a meaning. For instance, the explicit definition of a predicate, '*f*', as follows:

[30] The preceding reflections apply to implicit definition what we take to be the principal point of the Delphic-seeming remark at *Investigations* §201: 'What this shows is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases.'

x is $f =_{df}$ x is not true of itself[31]

works fine right up to the point where '*f*' is taken to lie within the range of 'x'. And because a similar move is required in order to elicit Russell's paradox from Basic Law V, it's tempting to say that the stipulation of that axiom can work as a *partial* explication of the notion of set, or extension of a concept, which must then however be further modified if a stable characterization is to eventuate. However, the point remains that a prime requirement on any coherent determination of the use of an expression, so on any implicit definition in particular, is consistency. An inconsistent stipulation cannot determine a coherent pattern of use, even if it somehow provides a glimpse of a coherent partial reconstruction of itself; and such stipulations cannot, of course, assist with the project of explaining non-inferential a priori knowledge (since that must require that we know the *truth* of the sentence which gives the definition!).

The traditional connection requires that we avoid arrogance. A good implicit definition has to be something which we can freely stipulate as true, without any additional epistemological obligation. This demands that, in the best case, such a purported definition must be not merely consistent but *conservative*: it must not introduce fresh commitments which (i) are expressible in the language as it was prior to the introduction of its definiendum and which (ii) concern the previously recognized ontology of concepts, objects, and functions, etc., whatever in detail they may be.[32] For if it does so, then an entitlement to accept it must await the

[31] This is of course the so-called *heterological* paradox.

[32] Note the thrust of clause (ii). It is our view that a stipulation *may* have consequences which can be expressed in the antecedent language, and to which there need have been no previous commitment, without compromise of its legitimacy *provided* the truth of these consequences makes no demands on the previously recognized ontology. If we omitted clause (ii), characterizing conservativeness purely in terms of innocence of novel consequences expressible in the antecedent language, then Hume's Principle, for instance, would be non-conservative. For it entails that the domain of objects is infinite—a claim which can be expressed purely in terms of the antecedent logical vocabulary which the Principle utilizes and to which there might well have been no commitment before. But the presence of clause (ii) restores the prospect of the conservativeness of the Principle, since its entailment of the infinity of the domain makes no demands on the previously recognized ontology, whatever it may have been, but is sustained by the objects—numbers—to which the Principle introduces means of reference.

Note also that the relevant notion of commitment here need not be proof-theoretic—it should be enough to exonerate an implicit definition from the charge of non-conservativeness if any fresh commitments it discloses are fresh only in the sense that they were not deducible using previously recognized machinery, but are indeed semantic consequences of previous theory. (However that fact will of course have to be recognized *somehow* if we are to be entitled to stipulate the definition.)

There are a number of subtleties that need exploring here. A more rigorous characterization of the kind of conservativeness which, we believe, is relevant to the case of abstraction principles may be found in C. Wright, 'On the Philosophical Significance of Frege's Theorem' in Richard G. Heck, Jr. (ed.), *Language, Thought, and Logic: Essays in Honour of Michael Dummett* (Oxford: Oxford University Press, 1997) and 'Is Hume's Principle Analytic?', *Notre Dame Journal of Formal Logic* 40:1 (1999).

ratification of those commitments and thus cannot be purely stipulative. Consider for instance the stipulation, 'The Moon is *f*',[33] viewed as an attempt implicitly to define a predicate '*f*'. *One* thing wrong with it is its non-conservativeness: its truth demands that the Moon exists, and so it can be known only if it is known that the Moon exists. But the kind of stipulation which is to be at the service of an account of knowledge a priori must be innocent of such collateral non-a-priori commitments. Notice, however, that this requirement is imposed by the traditional connection, rather than our leading issue in this section: the demands of genuine definition. Non-conservativeness need not, in and of itself, amount to a shortcoming in the ability of an implicit definition to fix a meaning for its definiendum. If 'The Moon is *f*' comes short in that respect too, it does not come relevantly shorter than the presumably conservative 'If the Moon exists, then it is *f*'—compare 'If anyone singly perpetrated all these killings, it was Jack the Ripper' and 'Jack the Ripper is the perpetrator of this series of killings': the latter is non-conservative, but its definitional purport is exactly that of the former, and presumably quite sufficient.

A natural suggestion is that 'If the Moon exists, then it is *f*' leaves open too wide a latitude of interpretation for '*f*'—it could mean e.g. any of 'orbits the Earth', 'is broadly spherical', 'is largely composed of rock', 'is a satellite of some planet', and so on. But it would be dangerous to insist that a term is properly defined only if it resists alternative interpretation, since Quine and others have contrived to open a standing concern about the extent to which such resistance is shown by the vast majority of (by ordinary criteria, perfectly well-understood) expressions of natural language. A better diagnosis of the shortcomings of the example sees them as having less to do with the scope it leaves us in interpreting '*f*' in ways consistent with the stipulation than with how little that stipulation tells us about the truth-conditions of *other* kinds of context which, such is the syntax of '*f*', a good definition of it should put us in position comprehendingly to construct. What are the truth-conditions of 'Saturn is *f*', 'This cricket ball is *f*', or 'That bust of Sophocles is *f*'? No basis has been provided for an identification of the truth-conditions of these claims, let alone for a determination of their truth-values.

This kind of shortcoming is closely connected with what Gareth Evans called the *Generality Constraint*. In Evans's view,[34] a subject can be credited with grasp of any particular thought only if creditable with a grasp of each thought in a relevant *local holism*. For instance, a thinker who grasps the thought, '*a* is *F*', must likewise grasp each thought of the form, '*b* is *F*', where '*b*' is any singular term which he understands and which stands for something of which '*F*' is signifi-

[33] Suggested by the more complicated example—'Snow is green and the moon is *f*'—which Horwich uses (*Meaning*: 133) to illustrate his 'existence problem'. Horwich's example is, of course, equally objectionable on grounds of non-conservativeness.

[34] See *The Varieties of Reference*: 100–5.

cantly predicable; and likewise he must grasp any thought of the form '*a* is *G*', where '*G*' expresses any property significantly predicable of items of the kind to which the referent of '*a*' belongs. The principle is plausible enough—though, taken at the level of thoughts, it has metaphysical resonances which are missing from the corresponding principle governing sentence-understanding. And the latter principle, for its part, is surely incontrovertible. For sub-sentential understanding is essentially implicitly *general*: to grasp the meaning of any sub-sentential expression is implicitly to grasp its contribution to the meaning of any sentence in which it can significantly figure. Accordingly one must, in grasping the meaning of any sub-sentential expression, '*f*', thereby grasp the meaning of any significant sentence, '#*f*', whose matrix '#–' one already understands. It is merely a special case of this that one who understands a term, '*a*', should understand each significant predication on '*a*' of any predicate already understood by him; or that one who understands a predicate, '*F*', should understand any significant sentence in which it is applied to a term, '*a*', which he understands. It is a platitude that to understand any sentence is to understand the meanings of its constituents and the significance of the way in which they are put together. Since the latter items of understanding are essentially general, it follows that to understand any sentence is to understand that range of significant sentences which can be derived from it by permutation of understood constituents.

The point, then, applies not just to the constituents—terms and predicates—of expressions of singular thought on which Evans was focusing, but to expressions of every syntactic type (and—when the Generality Constraint is conceived, as by Evans, as applying directly to thoughts as such—to the conceptual abilities appropriately corresponding to an understanding of such expressions). The bearing on successful implicit definition is clear. If the stipulation of '#*f*' really is to fix the meaning of '*f*', then it must succeed in explaining a pattern of use for that expression which complies with the Generality Constraint: that is to say, the implicit definition must put a recipient in a position to understand any well-formed sentence, '. . . *f* . . .', whose matrix, '. . . — . . .', is intelligible to him. This requirement may indeed be relaxed to allow cases where the implicit definition does not achieve this effect single-handedly, as it were, but forms an integral part of an explanatory complex whose overall effect complies with the Generality Constraint. But if a stipulation of '#*f*' falls short of fixing the meaning of an appropriately general range of contexts even when allied to other explanatory moves, then—even allowing that we may nevertheless have fixed the use of a wide class of sentence types involving '*f*'—a recipient will naturally feel that she does not really associate a meaning of '*f*', even though competent in the practice in which its use is a part, as far as it goes.

'*Any* well-formed sentence, ". . . *f* . . .", whose matrix, '. . . — ". . .", is intelligible to him'? There are plenty of examples of well-formed sentences containing only perfectly well-understood constituents which nevertheless make no (literal) sense. The significant application of a predicate, for instance, is often

restricted to one particular category of object, and one might well suspect that virtually all expressions have a limited *range of significance*—a limited range of sentential matrices in which it so much as makes sense to introduce them—so that the proper demand imposed by the Generality Constraint on the definition of an expression is only that it bestow understanding of any sentence resulting from combining the definiendum with an understood matrix encompassed in its range of significance. While it is right to regard it as a condition for the success of an implicit definition of an expression of any syntactic type that it bestow a sufficiently general mastery of the definiendum, it will thus be a substantive question in any particular case how general 'sufficiently general' is. Manifestly, however, wherever the line is drawn, 'If the Moon exists, then it is *f*' falls short.

As remarked, an implicit definition may be only part of the apparatus whereby the meaning of a term is fixed. Sometimes the remaining apparatus will itself consist in further implicit definitions. A further requirement which it seems should be imposed in such a case is, to put it very loosely, that various ingredient definitions pull in the same direction so that we do not have a situation where some members of the network of implicit definitions make a mystery of the others. The most familiar examples of violation of this constraint are cases where the inference rules which are stipulated for a logical connective are in *disharmony*. One possible disharmonious scenario is where the consequences which are stipulated to be permissibly inferable from premisses containing the connective in question do not allow of independent justification by reference to the type of ground which has been stipulated as sufficient in turn to justify premisses of that type; that is, most simply, a scenario where the *elimination* rule(s) for the target connective allow(s) the derivation from a premiss in which it is the main connective of conclusions which cannot independently be inferred from anything stipulated via the *introduction* rule(s), to be logically sufficient for such a premiss. In such a case, however, the use in tandem of the introduction and elimination rules will be non-conservative. The need for an additional constraint of harmony, over and above conservativeness, arises from the obverse scenario: that where, in the simplest case, the strongest conclusion permissibly inferable by an application of the elimination rule(s) is *weaker* than can independently be inferred from the type of premiss stipulated as sufficient by the introduction rule(s)—when intuitively, *more* ought to be inferable from a premiss of the relevant kind than the elimination rule allows.

Suppose for example we introduce a quantifier, '%', in first-order arithmetic whose introduction rule coincides with that for the universal quantifier, but whose elimination rule:

$$\frac{\Gamma \vdash \%x\mathrm{A}x}{\Gamma \vdash \mathrm{A}t}$$

is so restricted that 't' is required to have an *even number* as its referent. The effect then is that whereas, by the introduction rule, it will be necessary in order to establish '%xAx' on a given set of premises, to show that they suffice for an

arbitrary number to be A, the permissible conclusion from '%xAx' so established will be restricted to instantiations to even numbers. This pairing of rules for the quantifier '%' is conservative (so consistent) and seems in no way deficient in point of Generality.[35] Yet it would be natural to say that they nevertheless fail to determine any meaning for '%'. For if the introduction rule is justified, it can only be because grounds for each instance of 'Ax' are necessary before one can justify '%xAx'. And in that case it's unintelligible why the range of consequences which may permissibly be elicited from '%xAx' is explicitly restricted to *even* numbers.[36]

Such a schizoid pairing of the introduction and elimination rules for a connective provides a relatively simple and clear example of disharmony. But the point is quite general. Any satisfactory explanation of the use of a term must include provision both for the justification of statements featuring that term and for the use of such statements in the justification of others which don't. If it does not do that, it will be possible to find fault with it on grounds of Generality. But if provision is indeed supplied for both kinds of use, then there has to be a pitfall to be avoided of the kind just illustrated.

To summarize, then, the principal points of this and the preceding section:

(i) Implicit definition can underwrite (non-inferential) a priori knowledge only if it serves not merely to constrain the meaning of the definiendum in the way envisaged by the Horwichian, reference-fixing model, but to *explain* that meaning in such a way that it can be grasped by someone who antecedently lacks (the resources to define) the concept which the definiendum thereby comes to express.

(ii) Our suggestion is that this achievement may be secured provided that the definition creates a pattern of use in such a way that appropriate constraints of (at least) Conservativeness (and hence consistency), Generality, and Harmony are all satisfied.

(iii) It cannot in general be a requirement on successful implicit definition that it puts us in position—as we feel—*successfully to interpret* the definiendum; interpretation must perforce draw on independently given conceptual resources and, as we have stressed, it is integral to the interest of the notion of implicit definition that it be possible to think of some concepts as given by such means and

[35] Since uses as both premisses and conclusions have been provided for every type of sentence of the form '%xAx' where 'Ax' is any arithmetical open sentence; thus no less Generality has been achieved than we have for 'All numbers . . .' or 'Some number . . .'.

[36] Similar examples are given by Christopher Peacocke in his 'Proof and Truth' (cited in n. 4, see 167–8)—where he considers the possibility of a quantifier Qx . . . x . . . stipulated to have the same introduction rule as the usual existential quantifier, but no elimination rule at all (or, in a variant on the example, a restricted version of the usual elimination rule, allowing only conclusions shorter than some pre-assigned finite length)—and, more recently, in his chapter on the Philosophy of Language in A. Grayling (ed.), *Philosophy 2* (Oxford: Oxford University Press, 1998) where he considers (at 98–100) a connective *vel* having the same introduction rules as truth-functional disjunction, but a version of the elimination rule which permits only atomic conclusions.

no other. The proper intelligibility of an implicit definition has to consist not in its interpretability but in the type of constraints it imposes on the use of the definiendum.

(iv) Perhaps additions will need to be made to the constraints just reviewed before we will have something sufficient for the full intelligibility of the defined term. But we see no reason for pessimism that such a complete set of constraints can be given. In that case we shall have a framework of conditions such that any implicit definition which meets them will have two features: first, anyone who understands the syntactic type of the definiendum and the matrices with which it is configured in the sentences stipulated to be true, will be put in position to understand those sentences—to grasp the thoughts which they express—by their very stipulation; second, it will be integral to the meaning thereby determined that these sentences are indeed true—for the pattern of use which they demand, and which is in turn essential to the meaning successfully bestowed on the definiendum, is precisely a function of their being stipulated to be so. So a thinker who is party to the stipulative acceptance of a satisfactory implicit definition is in a position to recognize both that the sentences involved are true—precisely because stipulated to be so—and what they say. That will be to have non-empirical knowledge of the truth of the thoughts expressed.

5. THE IMPLICIT DEFINITION OF SCIENTIFIC-THEORETICAL TERMS

In the light of the foregoing, the idea that scientific theoretical terms are implicitly defined by (some proper subset of) the hypotheses of the theories in which they occur is in difficulty on two different counts. First, it is in immediate tension with the requirement of conservativeness in implicit definitions; scientific theories would be of little interest to us if they did not, in conjunction with certain observational data, provide the resources for the prediction of *new* observational consequences—consequences expressed in language innocent of the terms which they putatively define—which were not deducible before. Second, as noted at the beginning, space must be provided for a theory allegedly implicitly defining a term or range of terms, to be *disconfirmed* (i.e. shown to be probably false) by independently accumulated evidence. Yet if the meaning conferred by implicit definition on a term is always a meaning it acquires by stipulation of the truth of a sentence or sentences in which it is a constituent, any such provision is preempted. For in that case, there should be no explaining how subsequently gathered evidence might show the defining theory to be *false*—since that would be, *per impossibile*, a situation inconsistent with its retention of its meaning. We cannot—without lapsing into obvious incoherence—simultaneously hold that '*f*' has the meaning conferred on it by the stipulation that '#*f*' true, but that, so understood, '#*f*' has turned out to be false.

These difficulties do not show that the standard account—that implicit definition proceeds through stipulation of the truth of certain sentences—must be abandoned altogether, or even that it cannot be retained, in essentials, for the scientific case.[37] All that strictly follows from them is that, whatever the correct account of the idea that scientific terms may derive their meanings from their theoretical role, it cannot be right to conceive of implicit definition in scientific contexts as working through the stipulation (or acceptance) of the relevant theory *itself*—outright, unqualified stipulation of the truth (or acceptance) of the theory straightforwardly forecloses on the possibility of empirical disconfirmation. But this consequence can be accommodated by agreeing that the vehicle of implicit definition, in the scientific case, is *not* the scientific theory itself, but must rather be some *other* sentence so related to the theory that its stipulation or acceptance as true on the one hand suffices to confer meaning on the theoretical term to be defined but, on the other, remains uncompromised by recognition, a posteriori, of evidence telling against the theory. Schematically, what is stipulated as true (or firmly accepted) is not the theory T as such, but some sentence $\phi(T)$ which can remain in force through any empirical vicissitudes T may undergo. Our question should thus be: what, for given theory T, might plausibly serve as $\phi(T)$—the stipulative vehicle of the implicit definition?

Well, one natural suggestion—encouraged by a well-known tradition of theorizing about the manner in which theoretical scientific terms acquire meaning[38]—would appeal to the idea that we may view a scientific theory, embedding one or more novel theoretical terms, as comprising two components: one encapsulating the distinctive empirical content of the theory without deployment of the novel theoretical vocabulary, the other serving to fix the meaning(s) of the theoretical term(s) we seek to introduce. The theory's total empirically falsifiable content is, roughly, that there exist entities of a certain kind, viz. entities satisfying (a schematic formulation of) the (basic) claims of the theory. This can be expressed by the theory's *Ramsey sentence*, i.e. roughly, an existential generalization obtainable from the original formulation of the theory employing the new theoretical terms by replacing each occurrence of each new term by a distinct free variable of appropriate type, and closing the resulting open sentence by prefixing the

[37] Both difficulties feed on features—non-conservativeness and empirical falsifiability—special to the scientific case. Thus even if—as we do not accept—those difficulties enforce rejection of the standard account for that case, they do not straightforwardly generalize to other areas—centrally, logic and mathematics—in which implicit definition is practised. It would, accordingly, remain a perfectly viable option to uphold a mixed account, preserving the traditional connection for those latter cases in which it is plausible that our knowledge is a priori, whilst conceding that the standard account requires modification or replacement to accommodate implicit definition of scientific theoretical terms.

[38] The tradition includes Bertrand Russell, *The Analysis of Matter*; F. P. Ramsey, 'Theories' in his *The Foundations of Mathematics and Other Logical Essays* (London: Routledge & Kegan Paul, 1931); Rudolph Carnap, *Der logische Aufbau der Welt* (Berlin: Weltkreis, 1928); and David Lewis, 'How to Define Theoretical Terms', *Journal of Philosophy* 67 (1970): 427–46.

requisite number of existential quantifiers. Thus if, focusing for simplicity on the case where a single new theoretical term, '*f*', is introduced, the undifferentiated formulation of the theory is '#*f*', then its empirical content is exhaustively captured by its Ramsey sentence, '$\exists x(\#x)$', where the new variable 'x' replaces '*f*' throughout '#*f*'. The new term '*f*' can then be introduced, by means of what is sometimes called the *Carnap conditional*:[39] '$\exists x(\#x) \rightarrow \#f$', as denoting whatever (if anything) satisfies '#–' (on the intended interpretation of the old vocabulary from which it is constructed). This conditional expresses, in effect, a convention for the use of the new term '*f*'. Being wholly void of empirical content, it *can* be stipulated, or held true a priori, without prejudice to the empirical disconfirmability of the theory proper.[40]

[39] It is called the Carnap conditional by Horwich. In the paper cited in the preceding note, David Lewis calls it the Carnap sentence.

[40] Horwich observes (*Meaning*: 135–6) that a proponent of the traditional conception of implicit definition might try to deal with his 'possession problem'—specifically, with the objection that stipulating a theory formulation '#*f*' as true improperly forecloses on the possibility of empirical disconfirmation—by invoking this strategy. As against this, he claims, first, that the requisite existential generalizations 'are of dubious coherence insofar as they invoke quantification into predicate positions'; and, second, that if one regards Carnap conditionals as the proper vehicle of implicit definition of *scientific* terms, 'it would seem natural to suppose that *all* implicit definitions have something like this form, including those of arithmetical, geometrical and logical terms', with the result that fundamental logical principles, for example, could not be known a priori through their being used implicitly to define the logical constants. Obviously the crucial issue, as far as Horwich's second claim goes, is whether there is any compelling reason to think that implicitly definitional stipulation must *always* assume the form of a Carnap conditional (or something not relevantly different). Only if so will the traditional connection be subverted. So far as we can see, however, Horwich gives no actual argument to the purpose. His appeal to the naturalness of his supposition leaves us completely unimpressed—since there are clearly key features of the scientific case which do not straightforwardly carry over to logic and arithmetic (unless one is *already* persuaded of the correctness of Quine's global empiricism). For further discussion bearing on this issue, see Sect. 6 below. Horwich's first claim seems to us to manifest a degree of ambivalence about the Ramsey+Carnap strategy which he can ill-afford. Notwithstanding the impression to the contrary conveyed by his claim (ibid. 138) that shifting from the standard model of implicit definition to a use-theoretic account simply sidesteps the difficulties he raises for the former, exactly the same need to make provision for empirical disconfirmation arises on a use-theoretic account as on the standard model. Merely taking the meaning for '*f*' implicitly specified by embedding it in a scientific theory '#*f*' as 'the meaning constituted by *regarding* "#*f*" as true'—rather than 'the meaning "*f*" would need to have in order that "#*f*" *be true*'—makes no essential difference. On the use-theoretic account, our meaning what we do by '*f*' is constituted by our basic acceptance of '#*f*'. Accordingly, we *cannot* reject '#*f*' and still mean the same thing by '*f*'. Any break with the basic regularity—which consists in accepting '#*f*' as true—would involve either changing '*f*''s meaning or rendering it meaningless. Hence the use-theoretic account can no more accommodate rejecting '#*f*' as false than can the unreconstructed truth-theoretic account. It makes no difference that the former speaks of a regularity of *accepting* '#f' rather than a *stipulation of the truth* of '#f' (as on the former); clearly what causes the problem is the fact that '*f*'s meaning is supposed to be constituted by accepting '#*f*'—whether this results from a stipulation or not makes absolutely no difference. It is quite unclear how a proponent of Horwich's use-theoretic account is to solve the problem without invoking the Ramsey+Carnap strategy—so Horwich had better hope his misgivings about the requisite second-order existential generalizations can be assuaged.

This proposal promises an agreeably neat and simple solution to the problem of squaring a uniform account of the workings of implicit definition with the demands of empirical scientific theory. We may allow in general that a sentence or range of sentences '#*f*' will serve as an *indirect* implicit definition of '*f*' provided there is a uniformly recoverable sentence 'ϕ(#*f*)' associated with '#*f*', which may be viewed as an implicit definition in the standard stipulational sense. In the scientific case, the resulting proposal is, we may always take this sentence to be the corresponding Carnap conditional, regarding the theory itself as an indirect implicit definition of its distinctive vocabulary. Stipulating the truth of such a conditional is clearly neutral with respect to the epistemological status of the associated scientific theory, which can remain—just as it should—a posteriori and subject to empirical disconfirmation. Moreover it is clear just how such empirical disconfirmation can run without interference with the stipulation which, on this account, fixes the meaning of '*f*'—briefly, evidence running counter to the theory is precisely evidence against the theory's Ramsey sentence ('$\exists$x(#*x*)'), i.e. evidence that tends to confirm its negation '$\neg\exists$x(#*x*)', and thus in turn, via the logically true contrapositive of the Carnap conditional, confirms the denial of '#*f*'. Finally there is no a priori reason why a Carnap conditional should not be conservative and, indeed, satisfy all the constraints on implicit definition reviewed earlier and necessary for the viability of the standard account.

It is certainly well beyond the scope of this paper to review this proposal in any detail. We have dwelt upon it briefly only because the manœuvre is quite well known. What is evident is that Carnap conditionals are not the only kind of conditional sentence by means of which the meaning of new theoretical terms might be thought implicitly to be determined. For instance, a theorist at work during the early stages of subatomic physics, asked what he meant by 'electron', might say: 'Well, I don't *know* that there are any such things as electrons, but if there are such things, this much, at least, is true of them [here he states some bundle of claims which he takes to be true of electrons, if there are such things].' That is, he might explain (or partially explain) what he means by 'electron', not by giving us the Carnap conditional: 'If there are any things satisfying such and such laws, then electrons do', but by means of a kind of converse of it: 'If there are electrons, they satisfy such and such laws.' More generally and formally, we might view implicit definition of a theoretical term '*f*' as proceeding through the stipulation, not of the Carnap conditional: '$\exists$x(#*x*) $\rightarrow$ #*f*', but rather through that of a conditional of the type: '$\forall$x(x = *f* $\rightarrow$ #*x*)'. This could be called the *converse-Carnap* proposal.

As the reader will observe, this alternative would lack none of the mentioned advantages of the proposal utilizing Carnap conditionals. And there are probably further possible approaches, all serving to scotch the idea, looming large in Horwich's discussion,[41] that the traditional conception of implicit definition just

[41] See esp. *Meaning*: 135–7.

confronts an impasse when it comes to recovering the notion that theoretical terms receive their meanings from the theories in which they feature. The general perspective invited, to summarize, is this. Let *T* be any (empirical or non-empirical) theory containing a term, '*f*', expressing some distinctive concept of the theory, and let '#*f*' express some selection of basic claims or principles of *T* in whose formulation '*f*' is employed—as it might be, the conjunction of axioms of some logical or mathematical theory, or some of the fundamental claims of some empirical scientific theory. Then the view that meaning is given to '*f*' by its role in *T*—that it is implicitly defined thereby—may in all cases proceed by way of the thought that we in effect stipulate as true some appropriately related sentence 'ϕ(#*f*)'. How precisely 'ϕ(#*f*)' is related to '#*f*' will then depend on the discipline to which '*f*'-talk is to be added, and perhaps on details of the particular case. In some—non-empirical—cases, 'ϕ(#*f*)' may just be '#*f*' itself. In scientific contexts—and others, if such there be, where it is not open to us to stipulate the truth of '#*f*' itself—it will be some other sentence, such as a Carnap conditional, or a 'converse-Carnap' conditional, or some other type of context embedding '#*f*' which is, by contrast, available for outright stipulation.

This is merely a direction. Much more work would be needed before one could be confident what if any particular form of stipulation might best serve as a suitable vehicle for implicit definition in scientific contexts quite generally. The following section will assume no more than that the prospects of an account along such lines are to be taken seriously.

6. ABSTRACTION PRINCIPLES

As indicated in our opening remarks, we believe that Fregean abstraction principles—roughly, principles of the general shape:

$$\forall\alpha\forall\beta\ (\S\alpha = \S\beta \leftrightarrow \alpha \approx \beta)$$

—stipulating that it is to be necessary and sufficient for the truth of identities featuring §-terms that their arguments α and β stand in the equivalence relation, $\approx$, may legitimately be viewed as implicitly defining the term-forming operator '§' and thereby a sortal concept covering the referents of the terms it enables us to form. In particular, we hold that the sense of the numerical operator 'Nx . . . x . . .' may be adequately explained by stipulating the truth of Hume's Principle, i.e. the generalized equivalence:

$$\forall F\forall G(NxFx = NxGx \leftrightarrow \exists R(\forall x(Fx \rightarrow \exists!y(Gy \wedge Rxy) \wedge (\forall y(Gy \rightarrow \exists!x(Fx \wedge Rxy))))).$$

If this view of the epistemological status of Hume's Principle can be sustained, then in view of its now well-known entailment in second-order logic of the

Dedekind–Peano axioms for arithmetic,[42] there is every prospect of a vindication of a kind of *neo-logicism*, at least as far as elementary arithmetic goes.[43]

The Dedekind–Peano axioms are only satisfiable in infinite domains. In 'Platonism for Cheap? Crispin Wright on Frege's Context Principle',[44] Hartry Field—convinced that no combination of logic plus acceptable explanation of concepts could have existential import—contended that Hume's Principle cannot be any sort of conceptual truth, suggesting that the closest one could legitimately come to it would be a conditionalized version:

(HP*) If numbers exist, then $\forall F \forall G(\mathrm{N}xFx = \mathrm{N}xGx \leftrightarrow F \text{ 1-1 } G)$.[45]

This, he allows, is—or can be—a conceptual truth; but, of course, it does not entail the existence of numbers and is utterly useless for the neo-logicist's purposes.

The suggestion that one should replace Hume's Principle by such a conditionalization of it would seem to confront an immediate difficulty stressed by one of us in previous work.[46] How are we to understand the antecedent condition? In rejecting Hume's Principle as known a priori, Field holds that the obtaining of a one–one correlation between a pair of concepts cannot be regarded as *tout court* sufficient for the truth of the corresponding numerical identity. So, since that was an integral part of the proposed implicit definition of the concept of number, his position begs some other account of that concept. Otherwise there would seem to be no space for an intelligible doubt about the existence of numbers (there being no concept in terms of which the doubt might be framed). More specifically, if we take it that the hypothesis that numbers exist may be rendered as '$\exists F \exists x\ x = \mathrm{N}yFy$', then in order to understand the condition under which Field is

[42] See Crispin Wright, *Frege's Conception of Numbers as Objects* (Aberdeen: Aberdeen University Press, 1983): 158–69; also the appendix to George Boolos, 'The Standard of Equality of Numbers' in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam* (Cambridge: Cambridge University Press, 1990): 261–77. The result—now usually called Frege's Theorem—is first explicitly noted in Charles Parsons, 'Frege's Theory of Number' in Max Black (ed.), *Philosophy in America* (London: Allen & Unwin, 1964): 180–203.

[43] Not, of course, logicism in its most exacting sense—roughly, definition of all arithmetic primitives in purely logical terms in such a way as to facilitate transcription of each theorem of arithmetic into a theorem of logic. What is in prospect, rather, is a vindication of the claim that the fundamental laws of arithmetic are provable on the basis of second-order logic supplemented only with a principle which, though not an explicit definition, may be accepted as an implicit definition, proceeding in terms of concepts of second-order logic—and hence as an explanation—of the general notion of (identity of) cardinal number. See *Frege's Conception of Numbers as Objects*, §§xvii, xviii, esp. 153–4, and 'The Philosophical Significance of Frege's Theorem', Sect. I.

[44] See 167–70 of the reprint of this paper in Field's 1989 collection.

[45] This is not Field's formulation, but differences are merely notational. Cf. Field (1989: 169).

[46] See Crispin Wright: *Frege's Conception of Numbers as Objects*: 148–52 and 'Why Numbers can Believably Be: A Reply to Hartry Field', *Revue Internationale de Philosophie* 42 (167): 425–73, see Sect. V.

prepared to allow that Hume's Principle holds, we must already understand the numerical operator. But it was the stipulation—unconditionally—of Hume's Principle, which was supposed to explain that operator. That explanation has lapsed; but Field has put nothing else in its place.

Our reflections towards the end of the preceding section may seem to put Field in a strong position to reply to this objection. Consider the system consisting of Hume's Principle and second-order logic as a 'theory', in a sense inviting its comparison to empirical scientific theories, whose capacity to introduce theoretical concepts by implicit definition is uncompromised by the fact that they may turn out to be false. Think of this theory as *indirectly* implicitly defining the concept of cardinal number. Think of the real vehicle of this definition as being not Hume's Principle but the corresponding inverse-Carnap conditional:

$$(\text{HP}^{**})\quad \forall F\forall G\forall u\forall v((u = \text{N}xFx \wedge v = \text{N}xGx) \rightarrow (u = v \leftrightarrow F \text{ 1-1 } G)).$$

Then the complaint that Field has put nothing in place of Hume's Principle to enable us to construe the condition on which he regards it as a priori legitimate to affirm Hume's Principle is met head-on. Hume's Principle is still the key to the explanation of the concept of number—albeit indirectly. But the real explanation is given by the stipulation of HP**—a principle which stands fruitful comparison with the inverse-Carnap conditionals which we suggested might plausibly serve as the vehicles of the implicit definition of scientific theoretical terms, and which there is no reason to doubt meets all the conditions on legitimate implicit definition earlier discussed—or at least all those which Hume's Principle itself meets—while avoiding its objectionably abstract ontology. HP**, Field may say, tells us what numbers are in just the way that the inverse-Carnap conditional for any (other) scientific theory tells us what the theoretical entities it distinctively postulates are—by saying what (fundamental) law(s) they must satisfy, if they exist. That there are numbers is itself no conceptual or definitional truth—it is, rather, the content of a theory (in essence, the Ramsified version of Hume's Principle: $\exists\eta\forall F\forall G(\eta F = \eta G \leftrightarrow F \text{ 1-1 } G)$ which may perfectly well be—and in Field's view is—false.[47]

There is this much merit in the proposal: *if* there is good reason to insist that an implicit definition of the numerical operator should proceed, not through an outright stipulation of Hume's Principle, but through something more tentative, then it would seem that one appropriate shape for the stipulation is an

[47] This idea—or something closely akin to it—seems to be what Kit Fine has in mind when he suggests, in his 'The Limits of Abstraction', that Field can explain number using Hume's Principle whilst denying the existence of numbers 'by treating Hume's Law as an explanation of a variable number operator. The existence of numbers may then intelligibly be denied'—Fine claims—'since the denial simply amounts to the claim that there is no operator that conforms to the Law. If we regard Hume's Law as part of a "scientific" theory, then this response is equivalent to a Ramsey-style treatment of theoretical terms' (cf. 'The Limits of Abstraction': 524, fn. 10).

'inverse-Carnap' conditional of the kind suggested. The question is whether there is any such reason.

Can the comparison with the empirical scientific case provide one? Conditionalization is there called for in order to keep open the possibility of empirical disconfirmation. Fixing the meaning of '*f*' by stipulating the truth of the whole conditional '$\forall x(x = f \rightarrow \#x)$' leaves room for acknowledgement that the antecedent (more precisely, its existential generalization '$\exists x\ x = f$') might turn out false—grounds to think it false being provided by empirical disconfirmation of the consequent (more precisely, of the theory's Ramsey sentence '$\exists x(\#x)$'). Could an insistence that the numerical operator be implicitly defined, not by laying down Hume's Principle itself, but by HP**, be provided with an entirely analogous motivation?—by the need somehow to allow for the possibility that there might turn out to be no numbers?

Well, *how* might that turn out to be so? The suggested parallel with the scientific case would seem to require that such reason take the form of *disconfirmation* of Hume's Principle. But this looks hopeless. If, as it seems reasonable to suppose, Hume's Principle is (relevantly) conservative, straightforward empirical disconfirmation in not in prospect (since it will have no proper empirical consequences). How else might it fare badly?—might we somehow detect a misfit between its right- and left-hand sides in some particular case? That would have to involve 'observing' *either* that, whilst there was indeed a one–one correspondence between the instances of some concepts F and G, the number of Fs was distinct from the number of Gs, *or* that, whilst the number of Fs is indeed the number of Gs, there is no one–one correspondence. But to make either 'observation', we would require an independent criterion by which the identity or otherwise of the number of Fs with the number of Gs could be determined—precisely what we lack, if the meaning of the numerical operator is fixed only by the suggested stipulation of HP**.

It does not, of course, follow that there cannot be other reasons to insist upon a more cautious, conditionalized form of stipulation. We have accepted that if a stipulation is to serve as the basis of an item of a priori knowledge, in accordance with the traditional connection, then it must avoid arrogance. The stipulation of the relevant sentence as true ought not to require reference for any of its ingredient terms in any way that cannot be ensured just by their possessing a sense. If, then, an outright stipulation of Hume's Principle would involve arrogance, that would—or at least might—provide an independent reason for insisting that any acceptable explanation of the numerical operator should proceed, instead, through a conditionalized version of the principle.

'At least might'? We here leave open the question whether—at least in certain types of context (like axiomatic set-theory maybe)—arrogant stipulation might be acceptable after all. For even if not, it will follow that a conditionalized version of Hume's Principle, like HP**, is the strongest legitimate stipulation in its neighbourhood only if Hume's Principle itself is indeed arrogant. And if it suffices for

the avoidance of arrogance that the import of a stipulation may be parsed, as suggested earlier, into introductory and/or eliminative components, all conditional in form, prescribing which true statements free of occurrences of the defined vocabulary are to be respectively necessary and/or sufficient for true statements variously embedding it, then it is a sheer mistake to think that Hume's Principle is arrogant. The principle does not just assert the existence of numbers as 'Jack the Ripper is the perpetrator of this series of killings' asserts the existence of the Ripper. What it does—if all goes well—is to fix the truth-conditions of identities involving canonical numerical terms as those of corresponding statements of one–one correlation among concepts (compare the schematic stipulation: '*a* is the single perpetrator of these killings if and only if *a* is Jack the Ripper'). So it seems quite misdirected to complain that in stipulating the truth of the principle, we are somehow illicitly attempting simply to stipulate numbers into existence. The effect of the envisaged stipulation is, rather, to ensure that it is sufficient (as well as necessary) for the truth of identities linking numerical terms (and so for the existence of referents for those terms) that corresponding statements of one–one correlation hold true.[48] In this it is merely a paradigm instance of what we have suggested is the prototypical form of a legitimate implicit definition: its effect is that one kind of context free of the definiendum—a statement of one–one correlation between suitable concepts—is stipulated as sufficient for the truth of one kind of context embedding the definiendum: that identifying the numbers belonging to those respective concepts. That is its introductory component. And conversely, the latter type of context is stipulated as sufficient for the former. That is the principle's eliminative component. All thus seems squarely in keeping with the constraint that in order to avoid arrogance, legitimate implicit definitions must have an essentially conditional character. If the additional conditionalization in HP** is proposed in the interests of avoiding arrogance, it thus merely involves a condition too many.

This point will bear some emphasis. It derives entirely from a purely structural feature common to all Fregean abstractions. Whenever the meaning of a functional expression '§' is fixed by means of an abstraction principle '$\forall\alpha\forall\beta(\S(\alpha) = \S(\beta) \leftrightarrow \alpha \approx \beta)$', what is stipulated as true is always a (universally quantified) *biconditional*, so that what is done is to fix the truth-*conditions* for identities linking §-terms. The truth-*value* of instances of the abstraction's left-hand side is never itself a matter of direct stipulation—if any identities of the form '$\S(\alpha) = \S(\beta)$' are true, that is always the product of two factors: their truth-conditions, as

[48] It is, of course, true—and essential to the case for number-theoretic logicism—that the truth of certain instances of the right-hand side of Hume's Principle is a matter of logic. In particular, it is vital that the existence of a one–one correlation of the non-self-identicals with themselves should be—as it is—a theorem of second-order logic. More generally, if F is any finitely instantiated sortal concept, then, as a matter of logic alone, the identity relation correlates F's instances one–one with themselves, so that, applying Hume's Principle, we have it that $\mathrm{N}x\mathrm{F}x = \mathrm{N}x\mathrm{F}x$, whence $\exists y\ y = \mathrm{N}x\mathrm{F}x$.

given by the stipulation, together with the independently constituted and, in the best case, independently ascertainable truth of corresponding instances of the abstraction's right-hand side. The existence of referents for §-terms is therefore never part of what is stipulated—and implicit definition through Fregean abstraction is accordingly never arrogant *per se*.

It is sometimes complained that introduction of (terms for) abstract objects by Fregean abstraction somehow makes their existence a matter of mere linguistic convention, internal to language in a way which sacrifices objectivity or mind-independence.[49] The inappropriateness of this charge may be appreciated by contrasting explanation of terms by Fregean abstraction with more questionable forms of implicit definition which, if sanctioned, really would make the existence of their referents a matter of stipulation. We might, for instance, essay to fix the meanings of the primitive terms of arithmetic—'0', 'natural number', and 'successor'—by simply laying down the Dedekind–Peano axioms directly. Or we might fix the meaning of the sole non-logical primitive term of set theory—'ε'—by stipulating the truth of some suitable collection of set-theoretic axioms, such as those of Zermelo–Fraenkel. In either case, the stipulation of the axioms would directly call for the existence of an appropriately large range of objects—infinitely many finite cardinal numbers in the former case, and (prescinding from Löwenheim–Skolem problems) a much larger collection of sets in the latter—and would therefore be arrogant. Whether that must constitute a decisive objection to them is, again, not a question we shall take a view on here. However it is answered, it appears to us that the relative modesty of an explanation of *cardinal number* via Hume's Principle puts it at a definite advantage over any purported implicit definition which proceeds through arrogant axiomatic stipulation.

This is, to be sure, a point of some delicacy, to which it is easy to be tempted into misguided objection. For instance, it may be charged that since Hume's Principle has only infinite models, the stipulation that the principle is true can stand in no significant contrast with a direct stipulation of the Peano axioms—that there is no 'relative modesty', really. In both cases a stipulation is made which cannot hold good unless the domain of its first-order quantifiers is infinite, and in neither case would any antecedent justification be offered for this presupposition.[50] So what's the difference?

Well, anyone disposed to lodge this objection should ask themselves: how in principle might the infinity of the series of natural numbers *ever* be recognized? Of course there's the option of simply denying that it can be—of simply denying that the widespread belief that there are infinitely many natural numbers is anything more than a fiction, or groundless assumption. But anyone sympathetic

[49] See, e.g. Michael Dummett, *Frege Philosophy of Language* (London: Duckworth, 1973): 498 ff.

[50] Michael Dummett lodges essentially this objection in his original *Encyclopaedia of Philosophy* article on Frege, at 236. See Paul Edwards (ed.), *Encyclopaedia of Philosophy* (New York: Macmillan, 1965): vol. iii, 225-37.

to the opposing thought, viz. that the infinity of the natural numbers—and indeed the truth of the Dedekind–Peano axioms—is part of our most basic knowledge, should be receptive to the idea that it is *inferential* knowledge, grounded ultimately in deeper principles of some kind determining the nature of cardinal number. For the only alternative which takes it seriously—the idea that the truth of the usual axioms is somehow apprehended primitively and *immediately*—is not only epistemologically utterly unilluminating but flies in the face of the historical fact that the grasp and practice of the theory of the finite cardinals did not originate with the Dedekind–Peano axiomatization but antedated and informed it.

Very simply: if the question is raised, how do we know that the natural numbers constitute an infinite series of which the Dedekind–Peano axioms hold good, the available answers would seem to be, crudely, of just three broad kinds: that we don't actually know any such thing—it's a fiction or a groundless stipulation; or that we just do, primitively and immediately, know it; or that we know it in a manner informed by deeper principles of some sort. Our proposal is an answer of the third kind: the infinity of the number series may known by knowing that it follows from the constitutive principle for the identity of cardinal numbers. It does not compromise the insight of this answer if the 'deeper principle' is itself stipulative, and it does not make the infinity of the number series a matter of sheer stipulation in its own right. To think otherwise is to overlook the essentially conditional character of the original stipulation. It is also and independently to make a mistake akin to that of someone who supposes that, because the rules of chess have an ultimately stipulative character, we might as well just have stipulated directly that it is impossible to mate with just bishop and king.

In sum: the stipulation that the Peano axioms are true would be a stipulation that there is an infinite population of objects behaving as they require. We are urging, by contrast, that the stipulation of Hume's Principle should be seen first and foremost as a meaning-conferring stipulation—one providing for the introduction and elimination of contexts of numerical identity—of which it is a relatively un-immediate, interesting, and welcome consequence that there is a population of objects of which the Peano axioms are true. But there is another objection to this way of looking at the matter, a little subtler than that just considered. This is that, although Hume's Principle seems to fit squarely with the pattern of non-arrogant stipulation which we suggested suffices for an implicit definition to be safe, it bears a disanalogy to other examples of the same pattern—one which effectively brings out that it does not suffice, in order for a stipulation to be unarrogant in the intuitive sense intended, that it break down into conditional, introductory and eliminative components after the fashion described. Compare any instance of the introductory component of Hume's Principle:

$$F \text{ 1-1 } G \rightarrow NxFx = NxGx$$

with, once again, the presumably quite innocent introductory stipulation for 'Jack the Ripper':

If any single person perpetrated all these crimes, Jack the Ripper did.

Then the salient difference is simply this: that whereas the truth of the introductory content—the antecedent—in the Ripper stipulation implicates exactly the ontology implicated by the consequent, viz. the existence of a human being who is responsible for the crimes in question, the introductory context in an instance of Hume's Principle is *ontologically leaner* than the consequent; for it does not, whereas the consequent does imply the existence of a referent for the terms, 'NxFx' and 'NxGx'. So, an objector may contend, all the emphasis upon the conditionality of Hume's Principle, and other abstraction principles, is really a charade. There may be a difference between merely *directly* stipulating that certain objects exist on the one hand and, on the other, stipulating that, in order for certain contexts which require their existence to be true, it is to suffice that certain other contexts are to be true which don't!—but it is surely a difference of no philosophical significance whatever.

This would be a sharp objection if it did not misrepresent the situation. The misrepresentation consists in the suggestion that the introductory contexts in Hume's Principle are innocent of commitment to an ontology of numbers. For that precisely overlooks the fact that, as we've expressed the matter elsewhere,[51] the stipulation of Hume's Principle, and other abstraction principles, is tantamount to a resolution to *reconceive* the subject matter of their introductory components in a fashion determined by the overall syntax of and antecedently understood components in the type of identity statement introduced. The objection takes it that in stipulating an abstraction principle is to hold—for instance, that for directions:

$$Da = Db \leftrightarrow a \,//\, b,$$

—we would somehow be attempting to make it the case that contexts concerning just lines and their relations which are *wholly innocent* of the distinctive ontology of the left-hand sides—directions—are nevertheless both necessary and sufficient for the truth of statements which are *not* so innocent. That, of course, would not even be a *possible* stipulation. It would not be possible even if we somehow possessed a collateral assurance that every line had a direction. (For if that assurance did not extend to the point that these 'directions' were identified and distinguished in the manner described by the abstraction, we could not *make* that so just by stipulating it; and if the assurance *did* extend to that point, then the 'stipulation' would be altogether pre-empted and otiose—for the abstraction would hold good without it.) The response which we are proposing, then, is that in order to understand how an abstraction principle *can* be a proper object of stipulation, it is precisely necessary to receive it as so determining the *concept* of the objects to

[51] Wright, 'On the Philosophical Significance of Frege's Theorem', Sect. 1; Hale, '*Grundlagen* §64', *Proceedings of the Aristotelian Society* 1996–7: 243–61 and 'Arithmetic Reflection without Intuition', *Aristotelian Society*, Suppl. Vol. 73: 75–98.

which it serves to introduce means of reference that its introductory—right-hand side—contexts would precisely *not* be innocent of commitment to those objects. Directions, for instance, are precisely to be conceived as entities of such a kind that it is conceptually sufficient for a pair of lines to share their direction that they be parallel. That is the whole point of Frege's own initial metaphorical gloss, that in an abstraction principle we 'carve up' a single content in a new way.[52] The objection is good only if this way of looking at the matter is illicit. But the neo-logicist contention is precisely that it is not illicit—that it is, indeed, the key to an understanding of a great deal of our thought and talk about abstracta.

We are under no illusions but that this matter will stand much more critical discussion. Our aim in this concluding section has been merely to outline a case that if the traditional connection between implicit definition and non-inferential a priori knowledge can indeed be sustained along the lines offered earlier, then the prospects are also good for applying that idea so as to achieve viable neo-Fregean foundations for arithmetic, and indeed for more extended classical mathematical theories.[53]

[52] Frege, *Die Grundlagen der Arithmetik* (Breslau: Wilhelm Koebner, 1884): §64.

[53] For work towards the formulation of a suitable abstractionist basis for classical analysis, see Bob Hale, 'Reals by Abstraction', *Philosophia Mathematica* (3) vol 8, 2000.