Solving the Caesar Problem — with Metaphysics

Gideon Rosen, Princeton University
Stephen Yablo, MIT

July 2006

1. Introduction

In the course of defining numbers Frege memorably digresses to discuss the definition of items he calls *directions* (Frege 1884, §§64 -68). His first thought (§64) is that directions are things that lines *a* and *b* share if and only if they are parallel:

(DE) dir(*a*) = dir(*b*) iff *a* || *b*.

But then a worry occurs to him (§66). The goal was to define a certain kind of *object*, a kind including, for instance, the direction of the Earth's axis. And on reflection it is not clear that DE does that.

Certainly DE imposes a strong constraint on the direction-of *function*: it must associate parallel lines with the same object and non-parallel lines with distinct objects. And it constrains the *number* of directions too: there must be as many as there are non-parallel lines. But it is hard to see how DE constrains the directions themselves. As far as DE is concerned, directions might be almost anything. The direction of the Earth's axis, to take Frege's example, might be England. "[O]f course," Frege says, "no one is going to confuse England with the direction of the Earth's axis" (§66). But the lack of confusion here is "no thanks to our definition."

This sounds like a psychological point: if anyone were inclined to confuse directions with countries, DE would not get in the way. But the psychological point has a logical basis. Frege's real concern is that it would not be a confusion to identify England with the direction of the Earth's axis if there were no more to directions than is set out in the proposed definition.

What does it mean to say that "there is no more to directions" than is set out in the definition? That depends on how one thinks of definitions. The usual view is that definitions are verbal in nature. They convey verbal understanding by conveying what a word means, otherwise known as its sense or the concept it expresses. Explicit definitions do this by aligning the word with a phrase assumed to be already understood. Implicit definitions do it by stipulating "the truth of a certain sentence … embedding the definiendum and composed of otherwise previously understood vocabulary" (Hale and Wright 2000: 286).

On this view, to say that "there is no more to directions than is set out in DE" is to say that DE exhausts the concept of direction (= the sense of "direction"). Frege's point would be that DE does not exhaust the concept, since DE allows us to identify the direction of the Earth's axis with England, whereas the full concept of direction does not allow that.

How might one attempt to answer Frege on this interpretation of his complaint? It would have to be shown that DE delivers a richer concept of direction than he supposes. Its overt content may not be up to the task, but perhaps it has latent content that he is overlooking. Anyone who thinks that it would be a misunderstanding of the proposed definition to suppose that England is (or might be) the direction of the Earth's axis must think that the definition imposes a constraint on the interpretation of dir that goes beyond the requirement that DE be true. A solution to our problem (the Caesar problem) must therefore involve two ingredients: (a) an explicit formulation of this further constraint, and (b) a demonstration that any function that satisfies this further constraint cannot possibly have ordinary concrete objects[1] like countries or Roman emperors in its range.

One approach to this problem is to insist that when a new function symbol like dir is introduced by means of an abstraction principle like DE, the real constraint imposed by the definition is that this new symbol express a concept or a sense such that

> The thought/content/proposition that dir (a) = dir (b) is identical to the thought/content/proposition that a || b.

[1] Or for that matter, ordinary, independently identifiable abstract objects like the key of E flat minor or the Declaration of Independence.

This approach takes on large burdens. It must first to explain the relevant notion of a thought/content/proposition, and second explain why concepts satisfying this sort of constraint cannot have ordinary objects in their ranges. Perhaps these challenges can be met, but it seems clear that they have not been yet (Hale 1997, 2001; Potter and Smiley 2001, 2002; Yablo ms).

Another proposal is that the latent content of the definition is that *dir* is to express a function, the items in whose range fall under a *sortal* *concept* for which parallelism constitutes a *criterion of identity* (Wright 1983, Hale 1988, Hale and Wright, 2001, Fine 2002). Here again the challenge is first to explain these technical notions, and then to show that ordinary objects cannot fall under a sortal concept for which parallelism constitutes such a criterion. Maybe this approach can be made to work and maybe not (Rosen 1993, Rosen 2003, Hale and Wright 2003); hopes of a simple, straightforward solution along these lines are now faded, however, and so we propose to try something different.[2]

2. Understanding a definition

Consider an ordinary explicit definition. Suppose for instance that we introduce the word "grue" by laying it down that

(GR) For all *x*, *x* is grue iff *x* is green and observed or blue and unobserved.

[2] *Apparently* different, anyway. As we note below (n.6), our proposal may represent a version of this more orthodox approach.

What does the competent recipient of such a stipulation learn about the word "grue"? At a minimum he learns that its extension coincides with the extension of the open sentence on the right hand side; but of course a competent recipient will learn much more. He will learn, for instance, that the word has a certain intension: that an object is grue in a world w iff it is green and observed or blue and unobserved in w. So part of the latent content of GR is given overtly by GR+:

(GR+) Necessarily, for all x, x is grue iff x is green and observed or blue and unobserved.

Suppose now that a one place predicate like "grue" stands for a property; then we can say that someone who receives the definition GR in the intended spirit learns the intension of the new word's associated property. Does he learn anything else about this property? It would seem that he does. Consider someone who receives the definition and then says,

> I see that a thing counts as grue (in a world) whenever it is either green and observed or blue and unobserved. But I am still not sure what it is to be grue. Perhaps for a thing to be grue just is for it to be green and observed or blue and unobserved. But it might also be that for a thing to be grue is for it to be known by God to be green and observed or blue and unobserved. Or perhaps to be grue is to be green and observed and such that $e^{i\pi} +1 = 0$ or blue and unobserved and such that $e^{i\pi} +1 = 0$.

More bizarrely yet, consider someone who hears the definition and says

> Ah, I see what you're driving at: For a thing to be grue is for it to be either green and observed and such that that $e^{i\pi} +1 = 0$ or blue and unobserved and such that $e^{i\pi} +1 =0$.

Both of these characters have failed to understand the definition, even though both have interpreted "grue" so as to make both GR and and GR+ true. They have failed to appreciate that the latent content of the definition — the real constraint it imposes on the interpretation of the new word — is that

(GR++) For a thing to be grue just is for it to be green and observed or blue and unobserved.

More generally, we claim that in many cases (though not in all), when a new predicate $\underline{F}$ is introduced by means of an ordinary explicit definition of the form

> For all $\underline{x}$, $\underline{Fx}$ iff $\phi(\underline{x})$,

the real constraint imposed on the interpretation of $\underline{F}$ is the hyperintensional constraint that

To be F is to be ϕ.[3]

The new predicate is constrained to pick out a property with a certain definition or analyis given explicitly in the definiens, and is thus precluded from picking out other properties that may coincide with the intended referent in extension or intension. The bizarre misunderstandings listed above come from failing to appreciate the fact that the verbal definition has a latent content of this form.

Our hypothesis is that something analogous goes wrong if one hears the implicit definition of dir and imagines that England either does or might fall within the associated function's range. Just as hearing the explicit definition of "grue" teaches us everything there is to know about what it is for a thing to be grue[4], hearing the implicit definition of dir teaches us everything there is to know about what it is for a thing to be the direction of a line. We further submit that learning this puts us in a position to know that ordinary objects like England cannot possibly qualify as directions.

3. Formalities

[3] There are significant exceptions to this principle, such as mere "reference fixing" stipulations. The word "acid" might be introduced by the stipulation

A substance is an acid iff it tends to turn litmus paper red.

But this sort of stipulation does not even purport to tell us what it is to be an acid.

[4] More carefully: someone who hears the definition and knows what it is for a thing to be blue, green, observed, etc. is thereby in a position to know all there is to know about what it is for a thing to be grue.

The basic idea might be implemented in various ways, but the simplest seems to be this. Each entity $\underline{t}$ has a real definition, $\mathrm{DEF}_{\underline{t}}$. The real definition of $\underline{t}$ is a collection of structured propositions involving $\underline{t}$, which together say all of what there is to be said about what it is to be $\underline{t}$. Corresponding to $\underline{t}$ we have two sentential operators, $\Delta_{\underline{t}}$ and $\nabla_{\underline{t}}$. The first is truly prefixable of a formula $\phi(\underline{t})$ just if the proposition $\phi(\underline{t})$ expresses is a member of (or trivial logical consequence of – a qualification henceforth omitted) $\mathrm{DEF}_{\underline{t}}$. The second is truly prefixable of $\phi(\underline{t})$ just if the propositions in $\mathrm{DEF}_{\underline{t}}$ are conjuncts of (or trivial logical consequences of – another qualification henceforth omitted) the proposition expressed by $\phi(\underline{t})$. So, for instance, if *grue* is the color-like feature attributed by the use of "grue", then

$$\Delta_{\underline{\text{grue}}}\ \mathrm{GR}$$

says that it is definitive of *grue* that a thing is grue iff it is green and observed or blue and unobserved, while

$$\nabla_{\underline{\text{grue}}}\ \mathrm{GR}$$

says that nothing beyond GR is definitive of *grue*. The conjunction of these two claims can (and so might as well) be written

✡$_{\underline{\text{grue}}}$ GR

– in words, GR is exhaustively definitive of what it is to be grue. The reason that there is no room for the thought that another (unmentioned) part of what it is to be grue is to be known to the divine intellect, or such that $e^{i\pi} +1 =0$, is that anyone who appreciates the standard definition comes to know, not just that GR is true of grue, but that GR is exhaustively definitive of grue, and that there is nothing about *God* or $\underline{e}$ or $\underline{i}$ or $\underline{\pi}$ anywhere in it.

As the "grue" example suggests, the idea is meant to apply not just to definitions of terms but definitions of all kinds of expressions – predicates, function symbols, even perhaps connectives. The latent content of an explicit definition of the form

$\underline{a}$ = the α,

for all $\underline{x}_1, \ldots \underline{x}_n$, $\underline{P}\underline{x}_1, \ldots \underline{x}_n \leftrightarrow \psi(\underline{x}_1, \ldots \underline{x}_n)$,

for all $\underline{x}_1, \ldots \underline{x}_n$, $\underline{f}(\underline{x}_1, \ldots \underline{x}_m) = \phi(\underline{x}_1 \ldots \underline{x}_m)$,

for all $\underline{X}$, $\underline{Y}$, $\underline{X}*\underline{Y} \leftrightarrow \ldots.\ \underline{X}\ldots\underline{Y}\ldots..$

is given explicitly by the corresponding formula prefixed by $✡_{\underline{a}}$ or $✡_{\underline{P}}$ or $✡_{\underline{f}}$ or $✡_{\underline{*}}$ Definitions of this sort assign an object/property/function/operator with a determinate nature to a name/predicate/functor/connective by specifying what it is for a thing to be the object in question or the function in question or etc.

4. The proposal

So far we have been talking about explicit definitions, but the proposal is that implicit definitions like the direction equivalence and (to come finally to Frege's main concern) Hume's Principle may likewise be seen as real definitions of the functions they introduce. Hume's Principle

(HP) $\#\underline{F} = \#\underline{G} \leftrightarrow \underline{F} \approx \underline{G}$[5]

is normally taken as a verbal definition, and of course it is a verbal definition in part. But its ambitions are higher. Just as GR, advanced in the right definitional spirit, tells us not only that "grue" stands for a property that makes GR true, but also that

(GR++) ✡$_{\underline{\text{grue}}}$ $\forall \underline{x}$ ($\underline{x}$ is grue iff $\underline{x}$ is green and observed or blue and unobserved),

so HP advanced in the same spirit tells us not only that # stands for a function that makes HP true, but also that

(HP++) ✡$_{\#}$ $(\#\underline{F} = \#\underline{G} \leftrightarrow \underline{F} \approx \underline{G)}$

[5] The right hand side, F ≈ G, is shorthand for the second-order formula that asserts the existence of a one-one correspondence between F and G.

If we understand the definition correctly, we come away knowing that "#" can only stand for a function whose real definition is exhausted by the fact that it satisfies Hume's Principle. Of course, one can't be sure, to begin with, that there are functions like this. But if there are — call them *essential numerators* — then there is no question of what their natures are, since their natures flow from their definitions and their definitions are settled. To put the point in epistemic terms, if a function is an essential numerator then anyone who knows that it is an essential numerator and knows what it is for two bunches of things to be equinumerous is thereby in a position to know all there is to know about the function's nature.

Now it is one thing to define the function-symbol "#", another (and prima facie easier) thing to define the predicate "is a Number." Assuming we understand "#," the predicate can be defined explicitly:

(N) For all *x*, *x* is a Number iff for some *F*, *x* = #*F*.

If this definition is understood along the lines sketched above, then anyone who understands it comes to know that "Number" picks out a property satisfying the following condition:

(N++) ✡_{Number} (For all *x*, *x* is a Number iff for some *F*, *x* = #*F)*

Putting this all together, we conclude that when the neo-Fregean's definitions are properly understood, their recipient (assuming she knows what it is for two bunches of things to be equinumerous) comes away knowing everything there is to know about what it is to be a number.[6]

5. Three questions

Suppose we are right that the neo-Fregean has it open to him to offer his stipulations in the spirit indicated.[7] Three questions then arise. First, how can we convince ourselves

[6] Any property whose definition has this form will be a sortal property in one sense of the term: To be an instance of the property is to be the value of a function, whose nature is given by reference to an equivalence relation on items of a more basic kind. Grasp of (the nature of) any such property will then require a capacity to understand claims of identity and difference among the items in question, when those items are given as values of the relevant function. This is probably not an adequate account of the general notion of a sortal concept. Person, for example, is supposed to be a paradigmatic sortal concept, and yet it is implausible that the real definition of person makes reference in this way to an equivalence relation on more basic items (person stages?). We do not possess any adequate general analysis of the notion of a sortal. But our proposal seems clearly to entail that the properties associated with neo-Fregean predicates — i.e., predicates defined in the manner illustrated here for Number — will be sortal. Hence our suggestion (n. 2 above) that our proposal may amount to an unorthodox version of Hale and Wright's approach to the Casear problem.

[7] It is worth stressing that stipulations of a neo-Fregean sort need not be offered in this ambitious spirit. A neo-Hilbertian might introduce a function symbol ϕ by laying down the axiom

$$\phi(\underline{a}) = \phi(\underline{b}) \text{ iff } \underline{Rab}$$

with the express intention that any interpretation of the new symbol that satisfies the overt axiom is as good as any other. ("It must always be possible to substitute in all geometric statements the words table, chair, beer mug for point, line, plane.") The challenge for neo-Fregean Platonism as we see it is to describe an understanding of the latent content of the abstraction principles that would suffice to introduce concepts of the that there is at least one essential numerator, and hence that the stipulations serve to assign a function to the new function word? Second, how can we convince ourselves that there is at most one essential numerator, and hence that the stipulations confer determinate reference on the new function word and on complex functional terms constructed from it? Third, how can we show that the items in the range of the essential numerator are abstract mathematical objects of the sort with which arithmetic is intuitively concerned, and in particular, that Julius Caesar is not among them?

The first of these challenges is basically to answer the skeptic who doubts that any ontological ground is gained by the neo-Fregean stipulations. This is obviously a serious challenge, but it has nothing to do with the Caesar problem, so we propose to ignore it.[8] The second challenge is to answer a character we call the libertine, who thinks that so far as the neo-Fregean definitions are concerned, numbers and directions could be almost anything because the stipulations are multiply satisfiable.[9] The third challenge is to

sort we in fact possess: concepts that apply more or less determinately to abstract objects of a certain sort. It is no objection if these same stipulations might also have been used for a different purpose.

[8] A complete theory of these matters would tell us when in general a formula of the form ✡$_{\underline{f}}$[…$\underline{f}$…] is satisfiable. A maximally generous position would suppose that whenever … $\underline{f}$… has the form of an explicit definition there is always automatically an item with the requisite nature. But whatever the merits of generosity, the "bad company" objection to neo-Fregean Platonism (Boolos 1990) shows that this cannot be maintained when … $\underline{f}$… is an abstraction principle. A solution to the bad company objection will identify a class of kosher abstraction principles (presumably including the direction equivalence and HP) whose truth may be freely stipulated for the purpose of introducing a new function symbol. Conjecture: On any suitable account of this form, ✡$_{\underline{f}}$[…$\underline{f}$…] will be satisfiable whenever …$\underline{f}$… is a kosher abstraction.

[9] The most debauched libertine maintains that the values of the function introduced by means of abstraction principles can be anything at all — or more precisely, that the admissible interpretations of the new function symbol include every function that satisfies the formal constraint on the right hand side, including those whose ranges

answer a character we call the *pervert*, who grants that the stipulations are uniquely satisfied but insists that numbers and directions could still (in an epistemic sense) be almost anything. What does it gain us to know that there is such an entity as *the* number of Martian moons if that entity might, for all we know, be Julius Caesar?

6. Simple perversity

Consider the perverse hypothesis that the number of Martian moons = Julius Caesar. We instantly reject the notion. Why? There may be many reasons, but the following seems intuitively most fundamental. When it is proposed that JC might be the number of Martian Moons, we ask: Why him? Why not Caligula? Why not the key of E-flat minor? There is something *absurd* in the suggestion that JC might be the number of Martian moons precisely because the suggestion is impossibly *arbitrary*. And we are inclined to conclude that because the hypothesis would be absurd in this way, it cannot be true.

Our implicit reasoning seems to be: If JC were the number of Martian moons — if JC were 2 — then there would have to be some account of why *he*, rather than some other thing, is 2 — and, if this is different, some account of why *2*, rather than some other number, is JC. But it seems perfectly clear that there can be no such account. And so we conclude that the perverse hypothesis cannot be true.

include ordinary concrete objects. A more restrained sort of libertine maintains that the values can be anything at all so long as they are abstract, or so long as they exist necessarily (Rosen 2003). We ignore this distinction in what follows.

The principle we seem to be relying on here is that facts of the form [#F = a] cannot be brute facts. (We write [P] for the fact that P, and suppose that facts are structured complexes involving objects, relations, functions and the like as real constituents.) It is because this principle would be violated if 2 were JC that we reject the hypothesis as absurd. So we have two premises and a conclusion.

(P1) Facts of the form [#F=a] cannot be brute facts. When they obtain there must be an account of why they obtain.

(P2) On the perverse hypothesis some facts of that form would be brute and unaccountable.

(C) So, the perverse hypothesis is false.

Before we consider the status of the premises we should say a bit more about the sort of explanation or account we have in mind. Obviously enough, we are not looking for causal or historical explanations. Our guiding assumption is that some facts are grounded in others, or hold in virtue of others. The clearest examples of this are disjunctive facts, which typically hold in virtue of their true disjuncts, and existentially general facts that typically hold in virtue of their true instances. But there are other examples, some potentially controversial. Thus it might be said that facts involving determinable properties hold in virtue of the corresponding determinate facts —The ball is blue in virtue of being (say) cobalt blue; that facts involving definable properties hold in virtue of facts involving the real definitions of those properties — The figure is a square in virtue of being an equilateral right quadrilateral[10]; that certain thin moral facts hold in virtue of some more concrete *wrong-making* feature — The act was wrong in virtue of the fact that it was an unjustified violation of the victim's right to privacy; and so on.

(P1) asserts that facts of the form [$\#F = a$] must be grounded in more fundamental facts in this sense. There must be some other facts about a and about F and about # in virtue of which such identities obtain when they do.

Of course there is no point in pining after this sort of explanation if it is not to be had. So it may be helpful to show how there *could be* an explanation for why $\#F = a$. Suppose that there exist *Frege numbers*. A Frege number is a thing with a distinctive sort of essence. For each Frege number n there is a cardinality quantifier, $\exists_n$, such that

$$\text{✡}_n \quad \text{For all } F,\ (n = \#F \leftrightarrow \exists_n x\ Fx).$$

So, for instance, the Frege number 2 is an item that is, by definition, the number that numbers the Fs iff there is an F, x, and another F, y, and every F is either x or y. Also, though, and just as important, *all* essential facts about the Frege number 2 are implicit in the one just mentioned. This means that it is not essential to this item that, say, there are or could be human beings.

[10] This assumes that the fact that s is a square is distinct from the fact that s is an equilateral right quadrilateral, and hence that the properties in question are distinct.

If there are Frege numbers, then facts of the form [$\#F = a$] have a straightforward explanation. Why is it that the number of eggs in the bowl = 6? Because there are six eggs in the bowl and 6 is, by its very Frege-numberish nature, the number that belongs to a concept iff six things fall under it. Once again, we do not insist for present purposes that there are Frege numbers. The claim is rather that *if* there are such things, then facts of the form [$\#F = a$] will be explicable.

This suggestion about the nature of #'s values illustrates a larger claim about (we borrow the term from Goodman) "generating functions".[11] A generating function g is a function whose value on a given input x is *essentially* a value of g for arguments like x, and whose value on x has no more to its essence than that. Take for instance the *class of* function C – the function that takes plural arguments X and yields as output the class (if any) to which all and only the Xs belong. Few would dispute that

For all X, for all y, $y = C(X) \rightarrow \text{✡}_y\, y = C(X)$.

It strikes us as wholly definitive of a class that it be the class to which all and only *these* items belong. In more complex cases, however, this simple formula ($y = g(x) \rightarrow \text{✡}_y\, y = g(x)$) will not hold. It does not lie in the nature of a given Frege number to be the number that belongs to some *particular* concept. Rather it lies in the nature of such a number to

[11] Goodman 1956.

be the number that belongs to *any* concept with a certain higher-order feature (corresponding to the cardinality quantifier).

How should the notion be defined in general? A generating function g is associated with a partition, $\phi_1, \ldots \phi_i \ldots$ of its domain, in such a way that for each value y of g, there is a cell ϕ_y in this partition such that it lies in the nature of y to be the value of g for any member of ϕ_y. The cardinality quantifiers partition the domain of concepts, and each Frege number is associated, by definition, with a particular cell in this partition. Similarly, the various *orientation properties* — properties shared in common by classes of parallel lines — partition the domain of lines. And it lies in the nature of (what might be called) *Frege directions* to be the directions that belong to any line that possesses a certain such property. (The simple case mentioned above, where the values of the function are definitionally linked to some particular input, is the special case in which the relevant partition is the maximally fine grained unit partition.) The proper definition, then, seems to be this:

> $g{:}X \rightarrow Y$ is a generating function iff there exist properties $\phi_1 \ldots \phi_i \ldots$ such that
>
> (a) Every item in X instantiates exactly one ϕ_i,
>
> (b) For all x, y, if $\phi_i x$ & $\phi_i y$ then $g(x) = g(y))$
>
> (c) For each y in Y there is a ϕ_i such that ✡_y(for all x ($y = g(x) \leftrightarrow \phi_i x$)).

The point that matters is that facts of the form $[g(x) = a]$ will be explicable whenever g is a generating function. In each such case we will be able to say: The reason that $g(x) = a$

is that *x* is ϕ, and *g* by nature maps all ϕs to the same thing, and *a* is defined to be *that very thing*, the thing to which g by nature maps all ϕs.

7. Polymorphous perversity

Now it is clear that on the perverse hypothesis, facts of the form [#*F* = *a*] cannot be explained in the manner indicated. For it is characteristic of the perverse hypothesis to suppose that the values of abstraction functions are not definitively values of such functions, much less exhaustively definitively values of such functions. They are objects like Julius Caesar or the key of E-flat minor whose natures make no reference to the number-of function or equinumerosity. So on the perverse hypothesis, this style of explanation is cut off.

Another style of explanation is, however, conceivable. For the pervert might reply as follows: I agree that facts of the form [#*F* = *a*] cannot be brute facts, and that they cannot be explained by reference to the nature of the object *a*. But perhaps they can be explained by reference to the nature of the abstraction function, in this case #. Perhaps the answer to the question, "Why is Caesar the number of Martian moons?" is just this: There are two Martian moons, and it lies in the nature of # that Caesar = #*F* iff there are two *F*s.

How can we respond to this sort of pervert? By reminding her that the # function is *exhaustively* defined by HP and that HP makes no mention of Caesar or of anything that might implicate him. To put it epistemically, anyone who appreciates the definition and knows what equinumerosity and identity are is in a position to know all there is to know about the nature of #. But someone who appreciates the definition and knows what equinumerosity and identity are is *not* thereby in a position to know that Julius Caesar is the number of Martian moons iff there are two martian moons. So, our objector notwithstanding, it is not definitive of # that Julius Caesar is #*F* iff $\exists_n x$ F*x*. The same goes for the key of E-flat minor and every other independently identifiable object, abstract or concrete. It would therefore seem that if the perverse hypothesis is true, facts of the form [#*F* = *a*] cannot be explained by appeal to the definition of *a* or by appeal to the definition of #.

It is not out of the question that these facts might be explicable in some other way — either by appeal to the definition of some third thing or without appeal to definitions at all. We cannot say with confidence that every constitutive explanation — every truth of the form *P obtains in virtue of Q* — must be mediated either implicitly or explicitly by a claim about the definitions or essences of the items that figure in *P*. So we shall only say this. It is hard to think what such an explanation would look like in the present case. If this is right, then it is reasonable to accept (P2). Facts of the form [#*F* = *a*] would be brute if the perverse hypothesis were true.[12]

[12] We leave aside the possibility of explanations that reduce functional facts to *relational* facts with a uniqueness clause. It may be that [#*F* = *a*] obtains in virtue of the fact that *a* is the unique item that bears the number-of *relation* to *F*. But then the question will arise why *this* fact obtains. And it seems no more satisfactory that the relational fact should be brute than that the corresponding functional fact should be brute.

8. Brute perversity

So the pervert must say the following: I cannot tell you why Caesar is the number two. I cannot tell you what it is about him that suits him to the role, or what it is about the role that suits him to it, or what it is about the two together that suits them to each other. But why shouldn't facts of the form [#*F* = *a*] be brute facts? There is nothing wrong with the supposition that *some* facts are brute facts. Indeed it is natural (why?) to suppose that *all* facts are somehow determined by a foundational array of brute facts — facts that do not obtain in virtue of anything more fundamental.[13] Some of these brute facts will presumably be relational, e.g., the fact that one thing is at such and such a distance from another. If relational facts can be fundamental, why shouldn't some *functional* facts be fundamental? And if functional facts can be fundamental, why shouldn't facts of the form [#*F* = *a*] be among them?

One response is to remind the pervert that her hypothesis strikes us absurd, and that this reflects how natural we find it to think that there would have to be some explanation of why Caesar was the number two, if in fact he were. It is no answer to say that brute facts are not *intrinsically* objectionable. For they are not intrinsically unobjectionable either. We need to consider our reaction to the supposed brutality of this supposed fact in particular. And it seems highly relevant that we react to the suggestion with incredulity. One might rest the case for premise (P1) on this intuition alone.

[13] Of course the facts in this foundational array need not be altogether inexplicable. They may admit of causal explanation, for example.

Alternatively, one might make a broadly theoretical case. Suppose we are comparing two hypotheses: the perverse hypothesis, on the one hand, and the hypothesis that there exist Frege-numbers on the other. We note that the former must count facts of the form [#F =a] inexplicable, while the latter offers natural, intuitively compelling explanations for them. The latter package, then, exhibits greater explanatory coherence. And this, one might think, constitutes a reason to believe that it is true.

There is much to be said for responses like this; but they are not in the spirit of neo-Fregean Platonism. It is more in the spirit of the neo-Fregean view to seek a general principle that would tell us when facts of a certain kind cannot be brute facts.[14] We mention one such principle that seems to do the trick in the present context while conceding that it is not as luminously self-evident as one might wish.

A fact A is non-basic or derivative, we said, if it holds in virtue of another fact (or other facts). Non-basic facts can be explained by pointing to the facts in virtue of which they obtain; and if they can be so explained, they should be, meaning just that they should not be treated as inexplicable.

Now one great source, perhaps the main source, of non-basic facts is definitions; for facts about a definiendum will in general hold in virtue of facts about its definiens. The definition of grue, for example, in telling us that

[14] For a discussion of principles of roughly this sort, though in a different context, see Hale and Wright 1992.

(GR++) ✡$_{\text{grue}}$ (for all *x*, *x* is grue iff *x* is green and observed or blue and unobserved)

seems also to tell us that facts of the form [… grue …] are always grounded in facts about blue, green and observation. The definition of the Tri-State Area, in telling us that

(TS++) ✡$_{\text{TS}}$ (for all *x*, *x* is the Tri-State Area iff New York, New Jersey and Connecticut are parts of *x*, and every part of *x* overlaps, New York, New Jersey or Connecticut.)

seems to be telling us too that facts about the Tri-State area invariably derive from facts about New York, New Jersey and Connecticut. How, though, do GR++ and TS++ have these results? To answer this, we need a notion of a non-basic *thing* to put alongside our earlier notion of a non-basic fact.

A thing is non-basic if it is *reductively definable*, in the sense that its real definition has the form of an equation (typically a universally quantified biconditional or identity) in which the item in question is totally absent from the right-hand side. So for example, a reductively definable *relation* *R* will be one whose real definition takes the form

$$\text{✡}_{\underline{R}}\ (\forall \underline{x}_1, \ldots \underline{x}_{\underline{n}})(\underline{R}\ (\underline{x}_1, \ldots \underline{x}_{\underline{n}}) \leftrightarrow \phi[\underline{x}_1, \ldots \underline{x}_n]),$$

where $\underline{R}$ is totally absent from ϕ.[15] A reductively definable *object* $\underline{a}$ might be one whose real definition takes the form

$$✡_{\underline{a}}\ (\forall \underline{x})(\underline{x} = \underline{a} \leftrightarrow \phi[\underline{x}]),$$

where $\underline{a}$ is totally absent from ϕ, etc.[16] Perhaps the principle we need is just this:

(NB) If $\underline{a}$ is a non-basic entity, then facts involving $\underline{a}$ are non-basic facts.

In assessing this principle it is important to note that not all items that admit of real definition are non-basic. At one extreme, we have items $\underline{a}$ whose real definition consists in non-biconditional propositions involve $\underline{a}$, or biconditional propositions with $\underline{a}$ on both sides:

$$DEF_{\underline{a}} = \{\underline{Ca},\ \sim\underline{Da},\ \underline{Fa} \leftrightarrow \underline{Ga},\ \underline{Rab}, \text{etc. }\}$$

Here there is no suggestion that the facts about $\underline{a}$ should hold in virtue of $\underline{a}$-free facts. By contrast, and at the other extreme, when the real definition takes the form of an *explicit definition*, as with GR++ and TS++ above, then there does seem to be the implication that facts involving the defined item should be explicable in terms of more basic facts.

[15] Absent in the strong sense that $\underline{R}$ figures neither in ϕ nor in the definition of any of the constituents of ϕ, nor in the definitions of any of *their* constituents, etc.

[16] We could also allow for the case in which the definition takes the form of an identity, e.g., $✡_{\underline{a}}\ \underline{a} = \underline{f}(\underline{b})$.

HP lies between these two extremes. It is a reductive definition of #, but not an explicit definition of #. Brute facts about reductively definable items seem to us about as objectionable as brute facts about explicitly definable items. NB thus strikes us as plausible, and we know of no compelling counterexamples. But we concede that NB is not self-evident. A more searching account would look for a derivation of NB, or some other rule for determining which facts demand explanation. Our main point is that for reasons we may or may not have identified, facts of the form [#F = a] clearly seem to require explanation, and that such explanations are unlikely to be forthcoming on the perverse hypothesis. If this is right, then numbers might well be Frege-numbers. But they cannot be Roman emperors.

9. Uniqueness

So far we have been assuming that definitions like HP have at most one solution. Let us now ask whether this assumption can be justified.

Note that it would not be a disaster if the assumption were mistaken. For we have already seen reason to think that any suitable candidate for # must be a generating function, and this means that its values must be essentially among its values, and that there is not much more to their nature than that. This already rules out functions with "ordinary" objects as values, since on the one hand, ordinary objects are not essentially values of abstraction functions, and on the other hand, even if they were essentially values of such functions, that would fall far short of exhausting their essence.

Suppose we are right that the only admissible solutions to HP are essential numerators ν — functions whose real definitions are exhausted by HP — whose values are exhaustively definable as ν-applied-to-so-and-so-many-things. If there are many such essential numerators, then singular terms formed with # are to some extent indeterminate in reference. But they are not wildly indeterminate. They do not divide their reference over absolutely everything, or over every necessarily existing object, or over every necessarily existing abstract object. The most radical and implausible versions of libertinism are thus refuted. We are left, at worst, with the sort of moderate indeterminacy implied by Benacerraf's discussion of set theoretic reductions of arithmetic (Benacerraf 1965).[17] And on reflection it would be neither surprising nor disturbing if the language of arithmetic turned out to exhibit this sort of indeterminacy.

That said, it seems possible to argue that there is at most one solution in the offing. We know that any candidate for the referent of '#' must have a certain real definition: it must be an item such that everything there is to be said about its nature is in some sense a consequence of the fact that, by its nature, it satisfies HP. The only way there could be two such functions is for there to be two functions with precisely the same

[17] In fact the indeterminacy that threatens here is significantly milder than the indeterminacy that concerned Benacerraf. In Benacerraf's case the trouble was that there are countless ways of identifying the numbers with items of a very different sort, namely sets: items whose natures to not involve the relation of equinumerosity. In the worst case scenario for the present view, the numerals would divide their reference over items whose essences involve an essential numerator, and hence the notion of equinumerosity. So in our case, while there may be many candidates for the referent of '2', the candidates are all in some sense clearly <u>numbers</u>.

real definition. One way to “solve” the uniqueness problem is therefore to appeal to a general principle that functions, or perhaps items in general, are individuated by their real definitions.[18]

Now this general principle may be too strong. Consider the positive and negative square roots of -1, i and $-i$. It may be that the only thing to said about the natures of these items is that each is defined by the condition $x^2 + 1 = 0$. And yet the theory requires that these two items be distinct. And so we should be open to the possibility that there might be two *objects* with same essence or real definition. But now consider the corresponding claim about (say) *properties*. Above we saw that

(GR++) ✡$_{grue}$ (for all x, x is grue iff x is green and observed or blue and unobserved)

leaves no room for the thought that another part of what it is to be grue, for some reason unmentioned, is to be such that God exists, or such that $e^{i\pi} + 1 = 0$. But GR++ would also seem to leave no room for the thought that there are two grue-type features both of which satisfy the formula (and are therefore alike in real definition). Consider how bizarre it would be to say: “I know exactly what it is to be grue: it is to be green and observed or blue and unobserved. And I know exactly what it is to be *groo*: it is to satisfy exactly the same condition. And yet grue and groo are distinct properties.”

[18] This principle needs to be stated carefully. As we have defined the notion, the real definition DEF_a of an item a is a set of propositions involving a itself, and given this, it may be trivial that distinct items never have the same real definition. The principle we wants holds that when $a \neq b$, DEF_a is distinct from the result of substituting a for b and b for a in DEF_b.

This suggests the following principle: While objects (or perhaps substances) may be the exception, in every other case things are individuated by their real definitions. In particular, if f and g are functions whose real definitions run exactly the same way, they are one and the same function. If this is right, there can be at most one essential numerator, and hence at most one solution to Hume's Principle properly understood.

10. Recap

As noted earlier, any solution to the Caesar problem must provide two things: (a) an explicit account the latent content of HP and hence of the constraint HP imposes on the interpretation of the function symbol #, and (b) a demonstration that the values of the function(s) that satisfy this constraint do not include "ordinary" objects like Julius Caesar.

As to (a), we have suggested that when a function symbol is introduced by a neo-Fregean abstraction principle, the symbol is constrained to denote a function that is exhaustively defined by that principle. This rules out the vast majority of functions that simply satisfy the principle. It is one thing to be a function that (as a matter of fact, or even as a matter of necessity) takes parallel lines into the same object and non-parallel lines into distinct objects. It is another thing to be a function whose nature is exhausted by the fact that it has this feature. We might call any function with this sort of nature an abstraction function.

As to (b), we have suggested, first, that when *f* is an abstraction function, facts of the form [*f*(*b*) = *a*] must admit of a certain sort of explanation — there must be some account of why *f*(*b*) is *a* rather than some other thing — and, second, that such facts are only explicable if the abstraction function is also a *generating function*: roughly, a function whose values are essentially among its values. In one good (though perhaps somewhat narrow) sense of the word, an *abstract object* is an object that is, by its very nature, the value of a certain abstraction function for a certain range of arguments. So if our principles are sound, it follows that the objects "introduced" by a neo-Fregean abstraction principle must be abstract objects in this sense. And this means that ordinary objects like Julius Caesar are excluded, since it is perfectly clear that such things are not abstract objects in this sense.[19]

Finally we have argued that given a plausible principle concerning the individuation of functions — no two functions with the same real definition — there is *at most one* function that satisfies the real content of any given neo-Fregean abstraction principle. And this means that functional terms of the form #*F* and dir(*a*) refer determinately if they refer at all.

What we have not done is to show that there are in fact functions and objects satisfying these stringent conditions in the central cases. The initial appeal of neo-

[19] It also means the objects introduced by one abstraction principle are always distinct from the objects introduced by a distinct abstraction principle involving a distinct equivalence relation. So numbers, directions, shapes, etc. are all distinct. This principle has a number of surprising consequences. See Fine 2002: 46-54 for discussion.

Fregean Platonism lay in large part in its claim to vindicate the thought that reference to abstract objects like numbers and directions is a relatively unproblematic business. All it takes is the free stipulation of an abstraction principle — a definition, of sorts — and then presto, a new thoroughly meaningful function symbol is introduced, and therewith a capacity to refer to its (abstract) values. The bad company objection and its descendants throw cold water on that prospect. On pain of contradiction we must admit that the neo-Fregean procedure for introducing a new function symbol sometimes fails. On our account, the real latent content of such principles is both richer and stranger than one might at first have been inclined to suppose, and so it may seem that our proposal "increases the risk" of failure associated with the procedure. We have conjectured (see n. 8) that this apparent aggravation of the problem is an illusion. This will be true if in the end there is a metaphysical guarantee that every good abstraction principle defines a function, in the sense that there exists a function whose nature is exhausted by the fact that satisfies the principle in question. If that could be shown (and if the other principles upon which we have relied can be defended) we would possess a complete vindication of the central claims of neo-Fregean Platonism as we understand it.

REFERENCES

Benacerraf, P. (1965). "What Numbers Could Not Be", Philosophical Review 74: 47-73.

Boolos, G. (1990). "The Standard of Equality of Numbers", in Boolos (ed.), Meaning and Method: Essays in Honor of Hilary Putnam, Cambridge University Press, Cambridge, pp. 261- 77. Reprinted in his Logic, Logic and Logic, Harvard University Press, Cambridge, Mass., 1998, pp. 202-19.

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

Fine, K. (1994a). "Senses of Essence." Modality, Morality, and Belief, Sinnott Armstrong, Walter (ed). New York, Cambridge Univ Pr.

Fine, K. (1994b). "Essence and Modality." Nous Supplement, Volume 8: Logic and Language

Fine, K. (1995) "Ontological Dependence". Proceedings of the Aristotelian Society 95:269-90.

Frege, G. and M. Beaney (1997). The Frege Reader. Oxford ; Malden, MA, Blackwell Publishers.

Goodman, N. (1956). "A World of Individuals." The Problem of Universals: A Symposium. Notre Dame, Ind.: U of Notre Dame P, 1956. 13-31. Rpt. in Readings in Philosophy of Mathematics. Eds. Paul Benacerraf and Hilary Putnam. New York: Prentice Hall, 1964. 197-210.

Hale, B. (1997). "Grundlagen Paragraph 64." Proceedings of the Aristotelian Society **97**: 243-261.

Hale, B. (1994). "Dummett's Critique of Wright's Attempt to Resuscitate Frege." Philosophia Mathematica **3(2)**: 122-147.

Hale, B. (2001). "A Response to Potter and Smiley: Abstraction by Recarving." Proceedings of the Aristotelian Society **101**: 339-358.

Hale, B. and C. Wright (1992), "Nominalism and the Contingency of Abstract Objects", Journal of Philosophy 89, pp. 111--135

Hale, B. and C. Wright (2000). "Implicit Definition and the A Priori", in P. Boghossian and C. Peacocke, eds, New Essays on the A Priori, (Clarendon: Oxford).

Hale, B. and C. Wright (2001). The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics. Oxford, Clarendon Oxford Pr.

Hale, B. and C. Wright (2003). “Responses to Commentators”, Philosophical Books 44: 245-63.

Hale, B. (1988). Abstract Objects. New York, Blackwell

Heck Jr, R. G. (2000). "Syntactic Reductionism." Philosophia Mathematica 8 (2): 124-149.

Levine, J. (1996). "Logic and Truth in Frege (II)." Proceedings of the Aristotelian Society, Supplementary Volume 70: 141-175.

Potter, M. and T. Smiley (2001). "Abstraction by Recarving." Proceedings of the Aristotelian Society 101: 327-338.

Potter, M. and T. Smiley (2002). "Recarving Content: Hale's Final Proposal." Proceedings of the Aristotelian Society 102: 351-354.

Rosen, G. (1993). "The Refutation of Nominalism?" Philosophical Topics 21: 149-86.

Rosen, G. (2003). “Platonism, Semi-Platonism and the Caesar Problem," Philosophical Books 44: 229-244.

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

Wright, C. (1999). "Is Hume's Principle Analytic?" Notre Dame Journal of Formal Logic.

Wright, C. (1997). "On the Philosophical Significance of Frege's Theorem." Language, Thought, and Logic: Essays in Honour of Michael Dummett (ed. Heck Jr, Richard G.) New York, Oxford Univ Pr**:** 201-244.

Yablo, S. ms. "Content Carving, Some Ways"