Ramsey, Reference and Reductionism

Huw Price

This is an unpublished piece from July 1998.[1] It discusses the use of semantic notions such as reference in the Canberra Plan, the question whether this use creates a problematic circularity if the Canberra Plan is applied to the semantic notions themselves, and the relation of this question to Putnam's model-theoretic argument. I used some of the ideas in later papers such as (Price 2004, 2009) and (Menzies & Price, 2009), but the bulk of discussion of the relation of my concern to Putnam's argument (and to responses to Putnam by others) never made it into print.

References

Menzies, Peter and Price, Huw, 2009. 'Is semantics in the plan?', in D. Braddon-Mitchell & R. Nola, eds., *Conceptual Analysis and Philosophical Naturalism* (MIT Press), 183–200.

Price, Huw, 2004. 'Naturalism without representationalism', in David Macarthur and Mario de Caro (eds), *Naturalism in Question* (Cambridge, Mass.: Harvard University Press), 71–88.

Price, Huw, 2009. 'The semantic foundations of metaphysics', in Ian Ravenscroft, ed., *Minds, Worlds and Conditionals: Essays in Honour of Frank Jackson* (Oxford University Press), 111–140.

[1] This version was delivered to the Australasian Association of Philosophy Conference at Macquarie University, North Ryde, NSW, in July 1998, under the title 'Ramsey's Ticket to Ryde: A Trip on the Circle Line?'

Ramsey, Reference and Reductionism

Huw Price

1. INTRODUCTION

A metaphysical reduction aims to identify the entities of one domain with a subclass of those of another. It is a response to a puzzle of the form "What is X" (or "What are Xs?"), where X is some notion of philosophical interest. Metaphysical reductionists read such a query in a particular fashion, however. They take what is called for to be not, or not merely, a deeper understanding of our *concept* of X, for which a so-called *conceptual* reduction would suffice, but rather a theoretical *identification* of X with some member of a reducing class—or at any rate an argument that such an identification is possible, at least in principle.

In contemporary metaphysics the reductionists' query often derives its urgency from a commitment to physicalism, or naturalism—i.e., the assumption that all there are are physical (or "natural") entities and properties.[1] Given this assumption, there is simply nothing else for X to be but something in the natural world. If people who speak of X are speaking of anything at all, they are speaking of something natural, and the task of reductive metaphysics is to tell us what that is; or again, to point us in the right direction—roughly, to tell us how it could be that the term "X" refers to something in the natural world.

In that last sentence, for the first time in the paper so far, I used the notion of reference. I take it that the use seemed unremarkable, a fact which illustrates the familiarity of certain semantic notions—that of "making true", for example, as well as reference itself—in contemporary metaphysics. How do these notions come to play such a central role, especially in an area of philosophy which often sees itself as opposed to the

---

[1]What "physical" or "natural" mean here is itself a nice issue, of course, but one that I can set aside for the purposes of this paper. I want to oppose a popular contemporary conception of the task of naturalism in philosophy, and I won't disadvantage my opponents if I assume that their view is well-defined.

linguistic turn? Largely, I think, because the problems addressed originate in linguistic practice. As philosophers, we find that we ourselves, and the members of our speech communities, use many terms and concepts whose significance seems problematic (especially, perhaps, in the light of naturalism). Thus the problem originates at the level of language, or theory, and the task of philosophy seems to be to "place" and (at least where appropriate) to legitimate these aspects of linguistic practice. We want to relate aspects of language to the world—and hence, apparently, it is to be expected that we shall use semantic terms in doing so.

In fact, semantic notions such as reference and truthmaking play at least two roles in contemporary metaphysics of this kind. First, they are at the core of the guiding conception of what it is to "place" a theory, or show it legitimate. Legitimation and placement is a matter of showing that, and to what, the terms of a theory refer; that, and how, its sentences have truthmakers. Second, these notions play a crucial role in a popular strategy for metaphysical reduction, that of showing that X = Y *by* showing that "X" and "Y" refer to same thing.

Given that the semantic notions play these roles, it is natural to wonder about their own legitimacy and "placement"—and, indeed, about the coherency of these questions, given the role that the semantic notions themselves play in framing them. In this paper I want to explore one aspect of this "wonder". For most of the paper, I want to concentrate on the consequences of the fact that the semantic notions play the second of the two roles just described, at the core of a popular program for metaphysical reduction. I think that the fact that they play this role hasn't been as clear as it might have been, in part because the program in question is easily confused with a closely related program, in which causal rather than semantic notions play the role in question. I have no argument here with the causal version, but I want to show that the semantic version is in trouble, if it tries to apply its own methods to the semantic notions themselves. Moreover, I think the effect of this result is to call into question the use of the semantic notions in the first of the two roles described above, at the heart of the guiding conception of what it is to

place or legitimate a theory. For it suggests that there is an incoherency involved in trying to apply this standard to the semantic notions themselves.

Thus I want to begin with a local difficulty, concerning the application of a popular reductive program to a small group of semantic notions. The difficulty stems from the fact that these notions often play a substantive role in the reductionist program itself. The result, I want to argue, is that the program gives rise to vicious circularities, if we seek to apply it to the semantic notions themselves.

This conclusion is interesting enough in its own right. Naturalistic accounts of reference and related semantic notions are high on the wish-lists of many contemporary philosophers. It would be an unwelcome discovery that there is a principled objection to an account of these notions along what many would see as the preferred lines. But the damage would not end there. In virtue of the role these notions play in a popular characterisation of the preferred methodology for metaphysical reduction, and of the project of metaphysics more generally, the local problem threatens to contaminate the entire subject, so conceived.

I emphasise that I am not arguing that there is a difficulty which afflicts any attempt at theoretical identification whatsoever. The difficulty stems from the role that semantic terms play in what has become a popular strategy for physicalism. However, the strategy in question is closely related to—indeed, descends from—a strategy in which such terms do not play an essential role. My objection applies to the semantic strategy but not its non-semantic relative. But the latter suffices for many purposes, I think, including, probably, those of science. Theoretical identifications in some domains are thus untouched by the present argument.

The transition between the two strategies does not seem to be clearly marked or well recognised, so I'll need to begin with a little history, in order to clarify my target.

2. TWO ARGUMENTS FOR PHYSICALISM

Some thirty years ago, Davids Lewis and Armstrong independently proposed an argument for physicalism about the mental, along these lines:[2]

Mental state M = the occupant of causal role R

Physical state P = the occupant of causal role R

Mental state M = physical state P

The first premise was held to follow from conceptual truth about mental states—i.e., roughly, that they are defined in terms of their causal roles, in a network connecting perceptual inputs, internal states, and behaviour. The second premise was held to follow (for some appropriate P) from the principle that every event has a physical cause. The conclusion follows by the transitivity of identity.

Later, Lewis proposed that all theoretical identification follows this pattern.[3] Theoretical entities are defined in terms of their causal roles. Theoretical identification is then a logical consequence of the discovery that a certain range of entities have those causal roles.

Lewis fills out this idea using the Ramsey-Carnap technique for defining theoretical terms.[4] Ramsey had seen that there is a formal way to finesse the problem of the meaning of theoretical terms in science. In principle, we can always rewrite our theoretical commitments without using the terms which are the source of the problem. Whenever we hold true a theory of the form T(A), in which "A" is a theoretical term, we also hold true the proposition

(1) There exists an *x* such that T(*x*),

---

[2]David Lewis, "An Argument for the Identity Theory", *Journal of Philosophy* 1966; David Armstrong, *A Materialist Theory of Mind,* London, Routledge, 1968.

[3]"Psychophysical and Theoretical Identifications", *The Australasian Journal of Philosophy* 50(1972), 249–58.

[4]F. P. Ramsey, "Theories", reprinted his *Philosophical Papers*, ed. D. H. Mellor (Cambridge: Cambridge University Press, 1990).

in which all occurrences of the term "A" are replaced by the variable *x*, bound by an existential quantifier. Moreover, if T(A) incorporates *all* we take to be true of As, then to be committed to T(A) is *nothing more than* to be committed to (1).[5] After all, if we claimed to know anything more about A than that it was whatever makes (1) true, that extra commitment would conflict with the requirement that T(A) already incorporate all that we take to be true of A.

Against this background, then, Lewis proposed that the general model for theoretical identification is something like this:[6]

A = The unique *x* such that T(*x*)

T(B)

A = B

The first premise is the definition of "A" in terms of its theoretical role, while the second is the result of an (imagined) empirical discovery that the entity we know as "B" actually fills that role.

Two notes. First, there is a subtle shift between the application of this technique in the service of an argument for the identity theory about the mental, and the general model just presented. In the general model, which Lewis takes to operate in science, theoretical identification follows the empirical discovery that B has the properties which define A. In the original case, the argument for physicalism doesn't ask us to wait for the empirical discovery that a particular physical state is the cause of a given mental state. Rather, physicalism gets the go-ahead as soon as we accept that because everything has a physical cause, there is *some* physical cause of the mental state in question.

This makes the argument for physicalism a little fragile. If it is to be an argument at all, rather than a bald assertion that the mental state M is identical to some physical state, then the middle term—"the occupant of causal role R"—is crucial. Without it, there

---

[5]Modulo the issue of uniqueness, at any rate.

[6]David Lewis, "How to Define Theoretical Terms", *Journal of Philosophy* LXVII(1970).

would be no distinction between the general principle which supports the second premise ("Everything has a physical cause") and the physicalist's desired conclusion ("Everything is physical"). The middle term provides the logical separation which is essential, if the argument is not to be question-begging. Without an appeal to causation, in other words, the physicalist would have no argument, but simply a bare statement of the desired conclusion.

Second, the general model does not involve causation in any essential way. In the paper in which Lewis connects this general model for the definition of theoretical terms to his earlier argument for the identity theory, he uses the terms "causal role" and "theoretical role" more-or-less interchangeably, I think. However, this is easily seen not to be required by the formal technique itself. In Ramsey-Carnap terms, there is nothing to prevent us treating "causation" (and related terms) as among the theoretical vocabulary. The general model goes through exactly as before. In other words, discovering that B fills the A-role need not be a matter of discovering that B occupies a *causal* role. It is a matter of discovering that B satisfies some description, but not necessarily a description in which causal notions occur.

More recently, the Ramsey-Carnap-Lewis technique has come to be seen as the basis of a generalised argument for physicalism, perhaps most explicitly and clearly by Frank Jackson. Following Jackson's lead, the "Canberra Plan" proposes to apply the technique wherever philosophy encounters a puzzle about the nature of entities of a certain kind. The essential idea remains the same. In particular, the program retains the two-stage structure of the Armstrong-Lewis argument. Wherever we encounter a puzzle of the form "What is X?", we collect together everything that we (or "the folk") take to be true of X, assemble this into our theory of X, and eliminate the term "X" by forming the corresponding Ramsey sentence. So far, this is just conceptual analysis. The second stage provides the metaphysical "oompf": as in the Armstrong-Lewis argument, we are entitled to *identify* X with whatever "occupies the role" defined by theory in question.

If this is to provide a global strategy for physicalism, it must rely on some sort of existence claim, analogous to the one which grounds the second premise in the Armstrong-Lewis argument. (The physicalist is not proposing that we keep an open mind, sitting on our hands until the empirical discoveries roll in.) What form does this existence claim take? Here the question as to whether "role" means "causal role" or "theoretical role" in the more general sense just noted becomes critical. If it means causal role, then everything can precede as before. The second premise we need is essentially the same as that invoked by Lewis and Armstrong—viz., roughly, that everything has physical causes; "the explanatory adequacy of physics", as Lewis puts it.

But is this what the Canberra Plan has in mind? I don't know anywhere where the issue is addressed directly, and I have some sociological evidence for thinking that it hasn't been addressed. On the one hand, David Chalmers—an author apparently sympathetic to Jackson's program for reductive explanation—notes that the program needs to take causality and laws to be basic, not themselves amenable to causal-explanatory reduction. Rather, they need to be primitive, in order to provide the framework the program needs elsewhere. On the other hand, Peter Menzies—another author seemingly sympathetic to Jackson's program—proposes to apply the Canberra technique to the case of causation itself; saying, in effect, that causation is whatever occupies the appropriate role in folk causal theory. These authors cannot be singing precisely the same song. But so far as I am aware, neither observes that there are really two different tunes here, or at least two incompatible keys for the same tune, and says explicitly which of the two he proposes to follow.

However, Chalmers' instinct seems to me to be right. If the Canberra Plan is to provide a *global* argument for physicalism—an argument for physicalism about all truths—the existence claim in the crucial second premise cannot itself refer to causation. (After all, the theoretical vocabulary in question is not supposed to occur in this premise.) In other words, unlike the Lewis-Armstrong argument in the psychophysical case, it cannot rest on the claim that everything has physical causes.

So what takes the place of causation? The answer, I think, is one of a bundle a semantic notions, such as *reference, satisfaction* and *truthmaking*. Here's a rough example of the kind of argument which results:

J's belief that he is a beetle = The truthmaker of "J believes that he is a beetle"

Some state of J's brain = The truthmaker of "J believes that he is a beetle"

J's belief that he is a beetle = Some state of J's brain

In this case, the second premise might be taken to follow, not from any principle about causation, but from the claim that since mental states of affairs supervene on neuropysiological states of affairs, they have have physical (in fact, neuropysiological) truthmakers.

The general model goes something like this:

A = The (unique) occupant of the semantic role T(*x*)

For some physical B, B = The (unique) occupant of the semantic role T(*x*)

For some physical B, A = B

Here, "occupant of semantic role" is a kind of place-holder, which might be filled out in terms of reference, satisfaction or truthmaking, or perhaps some combination.

Obviously, the great appeal of this semantic reading is that it does hold out the promise of a global strategy. Not everything we talk about participates in causal networks, perhaps, but surely it all participates in semantic relations—it provides the referents and truthmakers for what we say, and thus fills a theoretical role in that semantic sense, if not always a causal sense.

As in the causal case, however, the middle term in this argument is not eliminable. Otherwise, the assumption which grounds the physicalist's second premise collapses into the desired conclusion. If reference is the key semantic notion, for example, then the

assumption concerned is that all (successful) referring terms have a physical referent. For this to be distinct from the physicalist's intended conclusion, the notion of reference must be substantive. Otherwise, "For all X, the term 'X' has a physical referent" is not distinct from "For all X, X is physical". In other words, "The referent of the term 'X' cannot be simply another way of saying "X", as a deflationary theory of reference would imply.

Thus the semantic notions are doing substantive work here. If someone proposed to use this style of theoretical identification in support of global physicalism in metaphysics, it would be a problem for them if the technique couldn't be applied without vicious circularity to the relevant semantic notions themselves. And yet the prima facie problem is obvious: if we try to apply the technique to the semantic notions themselves, we'll be left with nothing to do the work. That, in effect, is the problem I want to focus on in this paper.[7]

Thus my opponents are those philosophers who, explicitly or implicitly, are committed to using semantic notions in the service of a general physicalist program of reductive identification. I'll call this a "semantically-grounded" (as opposed to "causally-grounded") program for physicalism. Exactly who falls into which camp may be a matter for dispute. As a rough guide, I suggest that anyone who thinks that the notion of a truthmaker has important work to do in support of physicalism belongs in the latter camp, while anyone who appeals solely to the intuition that everything event has a physical cause or explanation may well belong in the former camp. But these lines haven't been drawn particularly sharply in recent philosophy, and I don't know if there is anyone who could fairly be described as committed to the semantically-grounded approach. However, the global ambitions of the Canberra Plan seem to set it on that course, and my aim is to fire a warning shot across its bows. If I am right, a semantically-grounded

---

[7]What if someone proposed to take the semantically-grounded argument in the second of the two senses I distinguished earlier—i.e., so that it rested on the discovery that for a particular B, B occupies the theoretical role definitional of A? Then A = B would simply be a product of conceptual analysis, I think, for with causation out of the picture, there is nowhere for empirical discovery to get a grip. So this can't be what the Canberra Plan has in mind.

version of global reductive naturalism has an embarrassing counterexample, in the worst possible place. Reference and related semantic notions are central to this version of the global program, yet inaccessible to its own methods.[8]

Again, I emphasise that this conclusion need not concern a reductionist who is less ambitious in one of two ways: either (i) in restricting her attention to cases in which the semantic notions are not needed (e.g. because a causally-grounded identification will suffice); or (ii) in being prepared to allow that despite their role in reductive naturalism, the semantic notions themselves properly lie outside the scope of a such a program. For the remainder of the paper I'll set aside causally-grounded reductionism, and hence option (i).

Someone who takes option (ii) owes us some other account of reference. What does the notion do, if not refer to some natural relation? There might seem to be two main possibilities: to opt for some kind of nonnaturalistic realism about reference, or to offer some sort of pragmatist or expressivist account of reference—perhaps a deflationist account, for example. However, the latter approach seems unable to supply the kind of substantive semantic notions the semantically-grounded program requires, and the former will be unappealing to anyone attracted to reductive naturalism in the first place. So there isn't much room to manouver here.

I'll return to these issues briefly later in the paper. Until then, my main aim is to show that semantically-grounded global reductionists do need to confront them—that the project of a naturalistic reduction of reference is much more problematic than has been appreciated.

3. The Role of Reference

The claim that the notion of reference is problematic for much contemporary metaphysics is not new, of course. In particular, it is the core of Hilary Putnam's model

[8]My focus will be on the semantic notions, but it is worth noting that the problem really stems from the global nature of the program's ambitions. A global version of the causally-grounded program would run into the same kind of problem, I think. The basic problem might be put like this: Without substantive notions of some kind, the second stage of the Canberra Plan loses its metaphysical "oompf". The question is whether it can be applied to the notions which supply that "oompf".

theoretic argument against what he calls metaphysical realism.[9] As will be evident, my argument is related to Putnam's. Though Putnam's target is different from mine, and I think that his argument does not present the difficulty concerning reference and reductionism in its most pressing form, it seems to me that he was right to identify reference as the weak point of much contemporary work in metaphysics. In particular, I want to try to show that in a new form, the point is invulnerable to widely-accepted responses to Putnam's argument by writers such as David Lewis, Michael Devitt and Mark Heller.[10]

The rest of the paper goes like this. In the next section (§4) I turn to the issue as to whether a semantically-grounded version of the Lewis-Jackson argument for reductive physicalism can be applied to provide a satisfactory account of reference itself. This involves a discussion of Putnam's argument, as it applies to the reductionist's project, and of the Lewis-Devitt-Heller response. I ask in what sense this response leaves intact the project of a reductive naturalistic account of reference.

Not surprisingly, the answer turns out to depend on what one expects from a reductive theory. In §5 I argue that while some kinds of reductive accounts of reference would be vitiated by circularity, the interpretation of the reductive project which is closest in spirit to the Lewis-Devitt-Heller response to Putnam seems to escape. The issue of the acceptability of what would thereby survive is rather subtle. As we'll see, it seems unattractive compared to what we normally expect from reduction, but arguably this is because our expectations have simply been too high. All the same, if there is a kind of reduction which remains viable in the light of Putnam-style objections—an interpretation of the project which renders the circularity involved in applying it to reference itself non-

---

[9]See Putnam, *Meaning and the Moral Sciences,* Boston, 1978; *Reason, Truth and History,* New York, Cambridge University Press, 1981.

[10]David Lewis, "Putnam's Paradox", *The Australasian Journal of Philosophy* 62(1984), 221–236. Michael Devitt, "Realism and the Renegade Putnam: A Critical Study of Meaning and the Moral Sciences", *Nous* 17(1983), 291–301; and *Realism and Truth,* 2nd. ed., Princeton: Princeton University Press, 1991, 225–230. Mark Heller, "Putnam, Reference, and Realism", in Peter French, Theodore Uehling and Howard Wettstein, eds., *Realism and Antirealism (Midwest Studies in Philosophy, Vol. XII),* Minneapolis: University of Minnesota Press, 1988, 113–127.

vicious—it is thinner and less nourishing version than reductionists seem to have appreciated.

Morevover, even this spare interpretation of the reductionist's project seems to be blocked, if the problem with applying the project to reference itself is couched in different terms. In §§6–7 I present two new forms of the objection. The first of these remains in the spirit of Putnam's objection, I think, while the second turns out to be interestingly related to an argument with a different target by Paul Boghossian. A discussion of the latter comparison, and of responses to Boghossian's argument by Michael Devitt, will help to clarify the problem, as I see it.

Up to this point, my main concern in the paper will have been to show that the semantic notions are problematic for a semantically-grounded approach to metaphysical reduction—that no such approach will give us the kind of illumination with respect to these notions that it attempts to offer in general. But this role in a popular program for metaphysical reduction is only one of the two roles I described at the beginning of the paper, which semantic notions play in contemporary metaphysics. The other role is in grounding the dominant conception of what it is for philosophy to legitimate or "place" a theory in common use (viz., that it is a matter of showing that the terms of the theory in question do refer, that its sentences do have truthmakers). I close by noting that the argument of the paper also casts doubt on the capacity of the semantic notions to serve in this role. At the very least, the upshot seems to be that they themselves are beyond the reach of such legitimation.

4. Reference Ramsified?

What does the Ramsey-Lewis-Jackson technique tell us about reference itself? In principle, the answer is straightforward. To decide what reference actually *is*, we muster the core commitments of our implicit theory of reference, and Ramsify the result. Thus the reference relation is whatever natural relation best realises the role of the concept *reference*—in other words, whatever makes true

(2) There exists an *x* such that RT(*x*),

where RT(reference) is our core theory of reference. Reference is whatever the theory RT refers to—whatever occupies the "reference" role.

This last way of putting the matter highlights the difficulty to which I want to draw attention. In applying the Canberra Plan to the case of reference, the notion of reference appears (explicitly or implicitly, depending on which formulation we choose) in the analysans. What we are told, in effect, is that reference is whatever we refer to when we use the notion of reference. Isn't this circularity vicious?

In support of the claim that the circularity is vicious, we might appeal to Putnam's model theoretic argument. In effect, Putnam points out that the formulation leaves the solution under-constrained. There are many possible interpretations of the term "reference", each of which makes true the sentence:

(3) "Reference" refers to reference.

To apply Putnam's point in this way is to appeal to an aspect of his much more powerful result concerning the existence of isomorphisms on sufficiently rich models. Putnam's own target is the thesis that even an ideal theory might be false, which is a central tenet of what Putnam calls metaphysical realism. Against this thesis, Putnam argues that there can be no fact of the matter, from the metaphysical realist's own point of view. Any theory has multiple interpretations, of which some render it true and some false. If truth is understood in the metaphysical realist's terms, Putnam concludes, then even an ideal theory lacks a determinate truth value.

In the context of this wider argument, Putnam responds to the imagined objection that a naturalistic reference relation (such as that canvassed by Hartry Field[11]) might pick out a preferred interpretation of the theory in question. Putnam says that the theory of reference would be in the same boat as everything else—it is "just more theory", and

[11] Hartry Field, "Tarski's Theory of Truth", *The Journal of Philosophy,* 1972.

hence as much subject to the problem of multiple interpretations as the theories we started with.

Putnam's model theoretic argument is controversial, however. Many commentators regard as satisfactory a response proposed in related terms by writers such as Heller, Devitt and Lewis. Roughly, the HDL argument (as I'll call it) is that what matters is not that we have a determinate *theory* of reference, but simply that there be a basis for a preferred reference relation *in the world*, to do the job of giving our other theories determinate interpretations. What does this job is not a *theory* of reference, but the reference relation itself.

In the present context, it is important to appreciate that the HDL response is a reply to Putnam's objection to realism in general. Heller compares the dialectic to that of debates between internalists and externalists in epistemology. In that case, externalists hold that a belief may be justified by objective facts concerning its reliability, even if we are unaware of these facts. Similarly, Heller suggests, our theories may stand in objective referential relations to the world, even if we are unaware of those relations (except, presumably, in trivial ways—we know that "electron" refers to electrons, but we may have no non trivial way to characterise this fact).

For the moment, let's grant that HDL have a valid objection to Putnam's general claim about theories. At any rate, let's grant that Putnam's argument does not close the door on metaphysical realism. Closing the door would require an argument that there isn't a reference relation in *re,* and Putnam hasn't given us that. As HDL point out, the mere existence of an objective reference relation would solve the general problem of multiple interpretations. In saying this we *use* the notion of reference, but do not *theorise* about it.

Nevertheless, Putnam's opponents can hardly decline to theorise about reference. After all, what could more important to metaphysical realism than a relation on whose existence the viability of the whole program depends? So for Putnam's opponents, as for

mine, the following issue remains pressing. Can they intelligibly theorise about reference, by their own lights?

Heller addresses this issue. Again he appeals to the epistemological analogy, and says that in both the espistemological and the metaphysical cases, there is no bar to theorising about the grounding facts concerned; but that if and when we do so, the same kinds of consideration obtain at a new level, in each case. In epistemology, our beliefs about first-order reliability may themselves be objectively reliable, without our having grounds for believing them to be so. And in the case of reference, a theory of reference may itself have determinate reference, without our being in a position to judge that this is so.

Heller says that it is not his goal to offer a theory of reference. However, in a remark which echoes the Ramsey-Lewis approach to theoretical terms, he suggests that a promising approach would go like this:

> The question can be broken into two: "What is the relevant role that a relation would have to play in order to satisfy the externalist's needs?" and "What relation, in fact, plays that role?" These questions do not seem unanswerable. (p. 125)

My point is that the second question is not answerable in terms which do not use the very notions at issue—reference, role-occupying, and truth-making. When we do try to theorise about these notions in Ramsey-Lewis terms, in other words, we find that the notions we are trying to theorise about play crucial roles in the theory itself (at the second, "metaphysical" stage). And how could this be informative? Being told that reference is whatever the term "reference" refers to is like being given an equation with too many variables—it doesn't specify a unique solution. (Suppose I tell you that the number I am thinking of is the solution of the equation $x = 2x/2$. Obviously I haven't said enough to make it a determinate matter what number I have in mind.) This seems to be a local version of Putnam's point, applied not against metaphysical realism but against the view that one could coherently theorise about reference in the reductive naturalist's preferred terms.

Thus to grant that HDL have a valid objection to Putnam's general point, is not to grant that there is any prospect of a satisfactory theory of reference. HDL do not tell us how there might be such a theory. They use the notion of reference—in their terms, it is the boot that keeps the door to realism ajar—but they don't establish that one can meaningfully theorise about it. As it stands, then, the dialectic of HDL's exchange with Putnam threatens to degenerate into something like this:

HDL: "The door to realism isn't closed, and it won't close so long as there's a boot in it—in other words, an appropriate reference relation in *re*."

HP: "This 'boot' is obviously crucial. Do tell me more about it."

HDL: "Why, it's the object wedged in the door."

In other words, the threat seems to be that even if we grant the HDL objection to Putnam's argument against metaphysical realism, Putnam's opponents are left unable to theorise about reference in any useful way.

In this debate, the pressure to theorise about reference is internal—Putnam's opponents surely owe us some account of whatever it is which is supposed to prevent the door to realism from closing. In the broader context of this paper, no such pressure is needed. So long as the ambitions of semantically-grounded reductive naturalism are global, it itself is committed to theorising about reference—to applying its general methods to that case. In what sense, if any, can those methods hope to produce something nontrivial, and nonvicious?

5. The Aims of Reduction

In order to clarify this issue, let's consider the claim that reference is what is picked out by the theory RT—in other words, in effect, this claim:

(4) Reference is what the theory RT refers to.

(As before, we are treating the relevant semantic notions as a package deal—assuming, for example, that casting this claim in terms of truthmaking, or predicate satisfaction, would not represent substantial progress.) Is the circularity here necessarily vicious? The

answer seems to depend on our conception of the intended role of this kind of reductive theory. Let's consider some possibilities:

(i) Instructing a novice in the use of a term.

If (4) were offered in this sense, the circularity certainly would be vicious. Someone who did not already understand the term "refers" would not achieve such an understanding in the light of (4). However, it is independently doubtful that (4) could sensibly be intended in this sense. To instruct the novice in the meaning of a theoretical term, we simply need to offer him the Ramsey sentence—we don't need the metaphysical second stage of the reductivist's project, to which (4) is supposed to belong.

(ii) Eliminating theoretical terms.

The original purpose of the Ramsey technique was to eliminate theoretical terms—to say what we want to say about X without using the term "X". This is possible for "reference", as (2) indicates.[12] In effect, (2) simply says that there is something which has exactly the properties we take reference to have. As yet, however, there is nothing metaphysical about this—it doesn't identify reference with anything else, naturalistic or otherwise. But as soon as the reductive program moves to the second stage, the theoretical term is reintroduced—it, or some equivalent, appears as in (4). In the context of a semantically-grounded version the Canberra program, therefore, there is a vicious circularity if we construe the goal of the reductionist's project in this way.

(iii) "Ghost-busting"—eliminating metaphysical spooks.

We noted earlier that the argument for physicalism needs a substantial middle term, to prevent it from collapsing into a question-begging statement of its intended conclusion. Thus a semantically-grounded version requires that semantic notions be substantive—and so, prima facie, matters with respect to which metaphysical puzzlement may arise.

---

[12]It wouldn't be possible for terms which play an internal role in the Ramsey sentence—this was Paul Horwich's point—but "reference" only gets used at the second ("metaphysical") stage of the Lewis-Jackson program, when we identify the referent or truth-maker of the Ramsey sentence.

Unlike deflationists, Canberra Planners thus have an obligation to show that reference and its semantic relatives are not spooky.

Yet the proposed method for eliminating spooks relies on an assumption formulated in terms of reference—e.g., that all terms in good standing have physical referents. It is hard to see how the program could be entitled to this assumption, while the "spookiness" of reference itself is at issue.

(iv) Metaphysical identification.

The least vicious interpretation of (4) seems to be one which owns up to the fact that the reductionist's project *presupposes* the notion of reference. If the general project of a reductive theory simply *is* that of specifying the reference of a theoretical term "X", then the case in which "X" is "reference" is not especially problematic. It assumes no more than other cases assume—namely, the notion of reference itself, in terms of which the general project is characterised.

Reading (iv) seems to be in the spirit of the HDL objection to Putnam's argument. In effect, the HDL point is that so long as there is a reference relation in *re,* reading (iv) is in good order. Does this mean that one can theorise about reference in Ramsey-Carnap-Lewis terms after all, provided one interprets the project in sense (iv)? I think that even if it did mean this, it is important to appreciate what a poor refuge this would be—how much would be involved in giving up theory in senses (i)–(iii)—*by the reductionist's own lights*.

After all, a large part of the point of the excercise is supposed to be "demystification"—in effect, finding a place for problematic notions, within a naturalistic world view. The Canberra Plan offers an attractive solution to metaphysical puzzles about all matters other than reference itself (and notions in the same semantic bundle). But it now appears that if it is semantically-grounded, it does so by placing reference beyond the reach of the same salvation. Whatever problem the Canberra Plan solves elsewhere—"to show how X can be something physical", for example—it cannot

solve in the case of reference itself. It cannot do for reference what it takes to be appropriate for everything else.

Supporters of HDL may respond at this point that the fact that no demystifying account of reference is possible does not show that there is no such relation. After all, they may say, the key point of the HDL objection to Putnam's argument is that it is the existence of the real relation that matters, not the existence of a satisfactory theory of that relation.

Even if this response is acceptable—and I challenge this in §§6–7 below—I think its cost has not been properly appreciated. After all, the point is not simply that there might be a real relation of which, as a matter of fact, we have no adequate theory. It is that we *can* have no theory of reference which achieves what the reductionist takes to be required in general. True, we can't have a philosophical account of anything in which the term "reference" doesn't occur. But elsewhere we can make progress, in the sense that we can eliminate other terms. With "reference" itself, the spade turns.

It might be objected that it turns for everything, eventually, and at the same point. Why hold this against the case of reference, when the only difference is that in this case, our starting point is bedrock itself? This answer is not incoherent, I think, but we should appreciate what a back-down it involves for reductive naturalism: it involves the concession, not just that *some* interesting notions are beyond the reach of the kind of demystification the program seemed to promise, but that the very notions at the core of the program itself are in this invidious position. And if everything else is ultimately in the same boat as reference, so much the worse, surely, for the program as a whole?

Where does this leave us? Reductive realism of this kind has not been shown to be incoherent, but it has been shown to rely on notions whose nature is profoundly inaccessible, by the program's own lights. It is important not to underestimate the point by appealing, say, to the possibility of a causal or teleological theory of reference. By the reductionist's lights, such an account is just more "first stage"—to complete a naturalistic reduction, we need to say that reference is what this first stage theory refers to. It might

seem that we can reduce the cost of the second stage by adding more to the first stage—by constraining ever more tightly what the relation underpinning the second stage could be. But if Putnam is right no amount of extra information makes a difference. It is all just "more theory".

Moreover, consider the difficulties involved in identifying a particular natural relation Q as the reference relation. In order to know that Q was the right candidate, we would already have to know what the reference relation is, so that we could establish that it is Q that stands in that relation to occurrences of the term "reference". In other cases, the Ramsey-Lewis technique tells us something of the form:

X is whatever key fits the lock defined by T(*x*)

In the case of "reference" itself, however, the standards for what constitutes a fit varies from key to key—for it is the notion of "fit" itself which is at issue.

It seems to me that although this isn't a refutation of the reductive program—reading (iv) remains open—it ought to give us cause to consider alternatives. A program which puts illumination so far out of reach might not be demonstrably false, but it is surely unattractive. Is the fate of philosophy really so bleak? Might the problem not lie with our assumed conception of the task of philosophy in this area?

What are the alternatives? To me, as I've said elsewhere, the most attractive option seems to be a nonreductive, explanatory naturalism, for which the crucial question is not "What is X?", but "How are we to explain ordinary talk of X?" However, this kind of approach will never solve the problem in its original form. At best it can offer a new life after therapy. Someone who remains in the grip of the reductive approach will certainly find it unsatisfactory.

Can we do any more to challenge the old life? I think we can. I think we can show not merely that the spade must turn in the sense that the reductionist cannot demonstrate success, but that the idea of success is in an important sense incoherent, by the reductionist's own lights. For what does success amount to, according to the

reductionist? In the case of reference, it is supposed to amount to the existence of a relation R such that "reference" refers to R. But the question whether there is such a relation cannot be sensibly asked, I think, for by the reductionist's lights a negative answer is incoherent.

When HDL say that the realist is undefeated so long as there is a reference relation in *re,* and that Putnam has not shown that this is not the case, we are being distracted from the main game. By dressing the issue up as an epistemological one—a matter of our inability to *know* that there is such a relation—the HDL argument encourages its opponents to concede the possibility that "the truth is out there". But this move is comparable to that of a theist, who attempts to fall back to agnosticism, in the face of claims that theism is unintelligible or incoherent. For a would-be theist, agnosticism is clearly more attractive than the view that theism is incoherent. But while the coherence of theism is in question, so too is that of agnosticism. An epistemological concession cannot address the original objection. At best, it diverts the theist's opponents. I think that a similar feint is at work in the present case.

Like theism itself, agnosticism requires that the question "Is there a god?" have a determinate meaning, and intelligible answers, both negative and positive. In the present case, the corresponding issues are these: Is the HDL position really well-defined, in the appropriate sense; does it have a determinate meaning? And can the reductionist really make sense of the possibility that "reference" might not refer?

The reductionist view is vulnerable on both points, I think. Putnam's argument suggests that the proposition "There is a reference relation" is simply not determinate, in the way that his opponents require. And the role of the notion of reference in reductionism seems entail that the question "Is there a reference relation?" simply isn't intelligible, in the way that the reductionist program requires. I'll take up these points in turn, in the next two sections.

6. JUST MORE METATHEORY?

The argument for the first of the above points is straightforward. HDL tell us that what matters in meeting Putnam's model theoretic argument is that there be an appropriate reference relation in the world. Let us grant them everything they want with respect to this relation. In other words, let us grant that the issue concerns the truth of:

(5) There is a relation r such that $\mathrm{Th}_{\mathrm{HDL}}(r)$,

where $\mathrm{Th}_{\mathrm{HDL}}$ summarises everything that HDL take to be true of the reference relation. In other words, (5) is the analogue of the proposition "There is a god"—it is the proposition the reductionist wants us to concede *might* be true.

The trouble is, Putnam's point applies to (5) as much as to anything. No matter how much the realist *says* about what in the world would defeat Putnam's objection, Putnam's objection will apply to what the realist *says*. An Archimedean referee could say what the realist wants to say, but *we* can't. For we don't have the right sort of access to the world to *say* what it would take to give us the right sort of access of the world. Whatever we attempt to say about it is "just more metatheory", and doesn't escape Putnam's strictures.

Thus the HDL objection seems to rely on a kind of sleight of hand. When it tells us that what matters is that there be an appropriate reference relation in *re,* not that we have a theory of such a relation, we tend to overlook the fact that the "telling" itself is in the same boat as everything else. It is not unnatural that we should overlook it, for the alternative—the detached perspective on our own sayings and judgements—is always an extra step. But a problem which is evident whenever we do take the detached perspective, does not go away when our attention is elsewhere.[13]

I noted earlier that Heller compares the HDL response to Putnam to an externalist position in epistemology. It now seems that he sets up the analogy in the wrong way, and

---

[13]It is like taking a defective purchase back to the store, and being given the same item, repackaged, as a replacement. If we don't notice the trick we may feel better, but objectively we have the same problem as before.

that the proper comparison works in Putnam's favour. It is as if Putnam had shown that it follows from the externalist's own assumptions that no belief is reliable, in the externalist's sense. By the externalist's lights, this would amount to an argument that all our beliefs are unjustified, whatever we might think about the matter. What Putnam's argument actually suggests is that by the realist's lights, no statement has determinate truth value. If the argument works, then, by the realist's lights, its conclusion is independent of whether it occurs to us to raise the issue in any particular case. So it applies to (5), which is the realist's own attempt to specify the condition in the world under which our utterances would have determinate truth value. The reference and truth-value of this condition are as indeterminate as anything else.

7. The Self-examination Paradox

I now turn to the second point from the end of §5. Think of reductive naturalism as asking two questions:

(6) Does "X" refer?

(7) If "X" does refer, then to what does it refer?

Naturalism consists in the assumption that if the answer to question (6) is yes, then the answer to (7) must be something in the natural realm.

The difficulty arises when we try to apply these questions to the term "reference" itself. Thus:

(6*) Does "reference" refer?

(7*) If "reference" does refer, then to what does it do so?

The difficulty is that (6*) does not seem to admit of a negative answer, by the naturalists own lights. If "reference" does not refer, there is no genuine property that the term lacks. But the question presupposes that there is such a property.

Let's illustrate the apparent problem by means of schematic version of reductive naturalism. According to such a theory, the term "X" refers to Y just in case occurrences of "X" are related to Y by some natural relation R. Thus:

(8) "Rabbit" refers to rabbits iff R[Oc(Rabbit), rabbits],

(where "Oc(X)" abbreviates "actual occurrences of the term 'X'", or something similar).

What can such a theory say about the referents of reference ascriptions themselves? If we are to talk about the natural relation R—in particular, if we are to make the claim about R intended to be made by (8) itself—then occurences of the term "R" must stand in relation R to R itself. As (8) itself tells us,

(9) "R" refers to R iff R[Oc(R), R].

In order for the RHS of (8) to have its intended content, then, the RHS of (9) must hold. But of course the same is true of the RHS of (9) itself—it doesn't say what the proponent of this account wants it to say, unless it itself actually holds. The expression

(10) R[Oc(R), R]

thus has the peculiar characteristic that under its intended interpretation, it cannot be conceived to be false. If it were false, "R" would not refer to R, and so (10) would not say what the proponent of this view wants it to say.

Is this circularity vicious? In context, I think it is. The reductive naturalist's goal is to find natural referents for our metaphysical terms and beliefs. Any version of this approach embodies a crucial standard of legitimacy for the use of metaphysical terms. Such uses are legitimate only if the terms in question do have *some* referent. This amounts to the requirement that the terms in question stand in the relation R to *something*. Otherwise, they simply don't succeed in referring; they don't have any content, and in using them we are guilty of a peculiar sort of error—a systematic failure to connect.

With this in mind, consider the reductive question with respect to the notion of R itself. By the reductionist's own lights, this amounts to asking whether her own use of the notion R is legitimate, in the implied sense—in other words, asking whether (10) holds. But can this question sensibly be asked, given that its very formulation presupposes a positive answer? In other words, isn't the reductionist's test of legitimacy impossible to apply in this crucial case? The problem isn't that the notion R actually fails the test. It is rather that because failure is incoherent, it doesn't seem appropriate to say that it passes it, either—which seems to be a kind of failure, given that by the reductionist's own lights, legitimacy requires that a notion does pass this test.

Why doesn't it pass it? Because the legitimacy of the exam paper itself presupposes a positive answer. Compare this case:

The University of Sydney

POLICY 22

A statement of policy shall be official if and only if it bears the above seal

Is this an official statement of policy? It passes the test, but the test is only as good as the paper it is written on. In other words, the test itself presupposes that Policy 22 is official. Hence it cannot usefully be applied to decide its own case. The standard it sets for other would-be policy statements isn't self-applicable, in any useful sense. And the reductionist's notion of the reference relation R seems to be in the same boat: It provides a criterion of legitimacy for other notions, but a criterion inapplicable to its own case.

A reductionist might be tempted to argue that this is a feature, not a fault. Perhaps we have a kind of transcendental proof of the legitimacy of talk of external reference relations, based on the idea that the referential status of such talk cannot meaningfully be called into question. One person who takes this line in response to a similar problem is Paul Boghossian. Boghossian argues that irrealism about content is incoherent, because irrealists need content and related notions to frame their position—for example, to say that statements about meaning have no truth conditions. Boghossian opts instead for a kind of transcendental realism about content.[14]

It may be that because Boghossian himself attacks irrealism and is prepared to embrace the transcendental realism, more naturalistic realists haven't appreciated the threat the point poses to their own metaphysical program. Clearly, transcendental realism about reference would be quite out of character for reductive naturalists. It would involve the concession that the task of legitimating our use of semantic notions is quite different from that of legitimating other metaphysically interesting notions. This would be sufficient for the main conclusion I want to draw: the program of reductive naturalism cannot be global, and comes to grief on the very semantic notions on which it itself relies.

One contemporary naturalist who does appreciate the threat that the point poses to a reductive naturalism is Michael Devitt. Devitt realises that the general metaphysical program requires that reference be in the same boat as everything else. We need to be able to make sense of the possibility that there is no such thing as reference—nothing in the world which "answers to" this folk concept. In a response to Boghossian, Devitt proposes that we might discover that the semantic notions are empty, but suggests that we could consistently couch this conclusion in terms of scientifically respectable replacements for these notions. (The notion of "answering to" will need to be cashed in terms of this replacement, for example.) In other words, we might discover that the term "reference" doesn't refer* to anything, where *refer** is the natural relation which comes to replace reference itself.[15]

---

[14]"The Status of Content", *Philosophical Review,* 99(1990) 157–184.

[15]Michael Devitt, "Transcendentalism About Content", *Pacific Philosophical Quarterly* 71(1990) 247–266.

Does this proposal address the problem in its present form? I want to make two points. First, it would involve a complete rejigging of the reductive project, as construed by my opponents. Indeed, it would amount to the concession that the reference-specifying approach *is* mistaken—though this concession would go along with the claim that there is an alternative, in the form of the reference*-specifying approach. (Don't underestimate the size of this concession by thinking of reference* as the product of a naturalistic analysis of reference. Remember, the view we are being asked to imagine is that there is no such thing as reference: nothing refers, and so a semantically-grounded reductive program could hardly be more in error.)

Second, the problem seems to arise all over again in this new program. According to the new program, concepts are legitimate only if they refer* to something, and this is an empirical matter. So what about the concept of reference* itself. Can we sensibly ask whether "reference*" refers* to the relation reference*? It may seem that there is no problem. Isn't it just like asking whether the term or concept "love" loves the relation *loves?* No—the difference is that when we ask the latter question we do so from the orthodox perspective, according to which the use of a term is "defective" if it fails to *refer.* In the imagined standpoint, however, this had been replaced by the criterion that the use of a term is defective if it fails to *refer**. From this standpoint, then, the problem looms as large as ever.

Again, it is easy to miss this point by thinking that what is at issue is whether reference* is a good naturalistic candidate to be identified with reference. If that were the issue, we could certainly imagine discovering that it was not a good candidate because "reference*" does not refer* to the relation reference* itself. But this would be a judgement made from the standpoint which Devitt wants to persuade us we might coherently abandon. The issue is whether the possibility of error would be coherent from the new standpoint he envisages.

Thus it seems to me that Devitt's response to Boghossian simply reinforces the point I want to make. As Devitt sees, it is vital to reductive naturalism in metaphysics to keep

alive the possibility of error. Given that program's conception of the task of philosophy, error must be a coherent possibility. Devitt sees that in attacking the coherence of error about reference itself, Boghossian attacks this naturalist program at its roots. But Devitt's own efforts to plug the gap simply indicate how deep the problem goes.

8. Conclusions

A semantically-grounded version of the Canberra Plan is unable to deal with the semantic notions themselves. The options seem to be: (i) nonnaturalist realism; or (ii) irrealism about semantic notions. Option (i) conflicts with the core commitments of the kind of naturalist who finds this kind of reductionist program attractive, while option (ii) fails to provide the substantive notion required at the second stage of the Canberra program (the stage which supposedly distinguishes it from mere conceptual analysis).

A causally-grounded physicalism is untouched by these arguments, so Lewis's account of theoretical identification in science remains unscathed. But the problems would be likely to re-emerge if an attempt were made to "go global" with the causally-grounded program—especially, if it sought to apply its methods to the case of causation itself. So what is philosophy to say?

This question connects to the second implication of the paper. I noted at the beginning that semantic notions play two roles in contemporary metaphysics. Apart from their role in semantically-grounded programs for metaphysical reduction, they ground the dominant conception of what philosophy has to do, in order to "place" or legitimate linguistic practice, within a naturalistic world view. Placement and legitimation is seen as a matter of showing that the statements of the practice in question do have truthmakers, that its terms do refer.

The above argument suggests that this conception of placement and legitimation is inapplicable to the very semantic notions on which it depends. If so, then this conception of the task of philosophy is ultimately self-defeating.[16]

[16]At various stages I have been much assisted by discussions on some of these issues with Richard Holton, Tim Crane and Paul Horwich, and by comments on talks on related material at Sydney, Toronto, Simon Fraser, Oxford and MIT.