Forthcoming in David Braddon-Mitchell & Robert Nola (eds), *Naturalism and Analysis* (MIT Press).

Is Semantics in the Plan?

Peter Menzies and Huw Price

1. Introduction

The so-called Canberra Plan is a grandchild of the Ramsey-Carnap treatment of theoretical terms. In its original form, the Ramsey-Carnap approach provided a method for analysing the meaning of scientific terms, such as “electron”, “gene” and “quark”—terms whose meanings could plausibly be delineated by their roles within scientific theories. But in the hands of David Lewis (1970, 1972), the original approach begat a more ambitious descendant, generalised and extended in two distinct ways: first, Lewis applied the technique to analyse the meaning of terms introduced not just by explicit scientific theories, but also by implicit folk theories such as folk psychology; second, he supplemented the theory to provide an account of the way in which the referents of the analysed terms might be *identified* on the basis of empirical investigation.

In the hands of the Canberra Planners, the Ramsey-Carnap-Lewis technique has been generalised further still. As Lewis had applied the model, theoretical terms were defined in terms of *causal* roles, and this played a crucial part in determining the kinds of entities that could be identified as the referents of the terms. But Frank Jackson (1998) and others (Tooley 1987; Menzies 1996) have extended the model to apply to terms for entities—such as moral properties and the causal relation itself—that appear to lack causal roles. Instead, practitioners of the Canberra Plan have framed definitions using a more amorphous notion of a functional role within a theory (where “functional” need not mean “causal-functional”, apparently). And Jackson has put this generalised model in the service of an ambitious physicalist metaphysics that attempts to show how all properties, including moral properties, can be identified with physical properties.

In this paper, our interest is in the role and place of semantic notions such as reference, satisfaction and truth in the Canberra program, and in its distinguished Lewisian parent. These questions are of considerable significance in their own right. For one thing, as we shall see, notions such as reference and satisfaction appear to play crucial roles in Lewis’s development of the Ramsey-Carnap technique. However, it is arguable that these uses are eliminable, in a sense familiar from discussions of the pros and cons of deflationary approaches to truth and reference: if so, then Lewis's own program is compatible with semantic deflationism, at least at this point. In view of the importance of Lewis's program, and the popularity of deflationism, this is a significant conclusion.

The same question can be raised with respect to the more ambitious program of the Canberra Plan, but in this case, as we want to explain, it acquires new bite. For one thing, it is arguable that the 'globalisation' of Lewis's technique envisaged by Jackson and his co-workers requires that semantic notions play the role played by causation in Lewis's original. If so, it seems to make the global program incompatible with semantic deflationism—which, canonically, wants to deny that semantic notions play such a substantial theoretical role. Evidently, that conclusion, too, is interesting in its own right.

Perhaps more importantly, however, the conclusion seems to leave the Canberra program vulnerable to two kinds of objections. The first is that resting on semantic foundations has the effect of placing the desired metaphysical conclusions out of reach, because there is no prospect of our achieving the kind of knowledge of the relevant semantic relations on which such conclusions would therefore have to rely. The second raises a spectre of circularity, or perhaps incompleteness, in virtue of the global ambitions of the Canberra program—if the program does depend on semantic foundations in this way, can it consistently be applied to the metaphysics of the semantic notions themselves? (Note that Jackson himself takes it that the program is applicable to these notions, and indeed offers them as the first of his examples of its intended application, in his 1998 (p. 2); so there are *ad hominem* grounds for raising the issue, as well as the theoretical grounds already mentioned.)

Our plan of attack is as follows. In §2 we outline Lewis's program for theoretical definition and identification, with two issues particularly in view. The first is the question as to whether the program is compatible with semantic deflationism—we argue that it is. The second concerns the precise role of causal notions in the program, and their sensitivity to issues about the precise objectives of the program.

In §3 we outline the Canberra Plan's proposed generalisation of Lewis's program. We make good the claim that at least on the most obvious understanding of the goals of the program in question, it requires that semantic notions take over a substantial role played by causal notions in Lewis's original program; and is hence incompatible with semantic deflationism. This leads us, in §4, to a discussion of the difficulties just mentioned: the issue as to whether the Canberra program can consistently be applied to the semantic notions themselves, and the question as to whether semantic notions provide a useful route to the investigation of the metaphysics of other topics.

In a brief concluding section (§5), finally, we reassess the claims of the Canberra Plan to be an heir to Lewis's program. We note that our discussion reveals that there are in fact two competing claimants to the Lewisian mantle, in the form of two distinct interpretations of the Canberra program. One version relies on substantial semantic notions, and is accordingly subject to the difficulties we have identified; the other does not rely on such notions, but is correspondingly less ambitious than Lewis's original program, in a sense our discussion in the earlier sections of the paper will have made clear. In different ways, both options offer us significantly less than the Canberra Plan might have seemed to promise. Our main conclusion is that the choice cannot be avoided: Canberra Planners cannot have their cake and eat it too.

2. Lewis's Model

Lewis's model for theoretical definition and identification involves two distinct techniques, or stages. It will be important in what follows that we be able to distinguish these stages, so we introduce them separately.

2.1 The first stage

The first stage or technique of Lewis's program is the Ramsey-Carnap-Lewis account of the meaning of theoretical terms. On this account, a theory is thought of as providing an implicit functional definition of the terms it introduces. Lewis (1970, 1972) gives an elegant schematic characterisation of how to make this definition explicit. Suppose we have a theory, T, that introduces some new terms $t_1,\ldots,t_n$. These are the T-terms—the theoretical terms. The other terms are the O-terms—the old or original terms whose meaning and reference are understood prior to the introduction of the theory. The theory T can be presented in the form of a single conjunctive sentence—the *postulate* of the theory. Lewis writes: "It says of the entities—states, magnitudes, species, or whatever—named by the T-terms that they occupy certain *causal roles*: that they stand in specified causal (and other) relations to entities named by O-terms, and to one another." (1999 [1972], p. 254)

Thus the postulate is written:

$$T[t]$$

By replacing all the terms in the postulate with variables $x_1,\ldots,x_n$ and prefixing this formula with existential quantifiers, we obtain the Ramsey sentence of T:

$$\exists x T[x].$$

This says that there is an n-tuple of entities satisfying the postulate, or in other words that there is a realisation of the theory T. We can also obtain the modified Ramsey sentence, which says that that T has a unique realisation:

$$\exists_1 x T[x].$$

Lewis suggests that if we want a meaning postulate for T, we should adopt what he calls the modified Carnap sentence:

$$\exists_1 x T[x] \supset T[t].$$

This says that if T is uniquely realised, the T-terms name the components of this realisation. This meaning postulate implies a sentence that explicitly defines the T-terms by means of the O-terms:

$$t = \text{the unique } x \text{ such that } T[x].$$

Lewis calls this a *functional* definition: the T-terms have been defined as the occupants of the causal role specified by the theory T—they are the entities, whatever those may be, that bear the specified causal relations to one another and to the referents of the O-terms.

2.2 Does the first stage require inflationary semantics?

We have just employed the phrase "the referents of the O-terms" and (quoting from Lewis) the phrase "entities named by the O-terms". Question: Can these uses of the semantic notions of reference and naming be understood in a deflationary or "disquotational" spirit, or do they indicate that the first stage of Lewis's program is already committed to employing more robust semantic notions? The answer is that the uses in question can indeed be taken in a deflationary spirit. They meet a need that arises, in effect, simply from the logical generality of the exposition given above.

In any actual case, we can replace an expression such as “the referents of the O-terms” by appropriate uses of the O-terms in question themselves.

Consider, for example, Lewis’s famous application of this account in the service of a functional definition of terms for mental states. He asks us to think of folk psychology as a term-introducing theory, consisting in platitudes regarding the causal relations of mental states, sensory stimuli, and behavioural responses. The general form of these platitudes will be:

> When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so behavioural response.

The postulate of the theory is the long conjunction of these platitudes. The T-terms are “belief”, “desire”, and the like, while the O-terms include “sensory stimulus” and “behavioural response”, and more specific terms in the same families, as well as the causal vocabulary itself. On the basis of the postulate, we can form a functional definition of mental states that defines the mental states collectively in terms of their causal relations to stimuli, responses, and each other. And because we can *use* the O-terms themselves in formulating this definition, we don’t need to employ semantic notions such as reference (at least at this first stage in the program).

A simpler example, to reinforce this point: suppose we are interested in the theory of smoke detectors. Presumably, all smoke detectors have an on-state, which, in a properly functioning detector, is typically caused by the presence of smoke, and causes the emission of a loud noise (or other alarm signal). We could say:

> The referent of the term “on-state” is the state which is typically caused by the referent of the term “smoke” and which typically causes the referent of the term “loud noise”.

But clearly the reference to referents is otiose: we just told you what the on-state is, without using it.

In general, then, the first stage of Lewis’s model—the Ramsey-Carnap-Lewis technique for functionally defining theoretical terms—does not make essential use of non-deflationary semantic notions. (It might do so in particular cases, of course: non-deflationary semantic terms might be among the O-terms.)

2.3 The second stage

The second part of the model consists in Lewis's technique for identifying the referents of functionally defined terms on the basis of empirical information. Lewis observed that theoretical identifications such as the identifications of water with H_2O and of light with electromagnetic radiation had previously been thought of as "pieces of voluntary theorising": they were hypothesised as bridge laws identifying the entities of one theory with the entities of another theory. But the Ramsey-Carnap-Lewis account of theoretical definition made possible another model of theoretical identification, according to which they are logically implied by the functional definitions of the theoretical entities, taken in conjunction with other bodies of knowledge.

Again, Lewis's treatment of mental states provides a useful illustration of this technique (and of a further distinction that will play some role in what follows). According to Lewis, the identity of mental state-types with particular physical brain state-types could in principle be established by a simple argument from two premises:

> Mental state M = the occupant of the M-causal role R.
> Neural state N = the occupant of the M-causal role R.
> Therefore, mental state M = neural state N.

Lewis says that the first premise is an *a priori* truth, supplied by the functional definition of mental states in terms of their causal role. The second premise would be an *a posteriori* truth, supplied by the advance of neurophysiology. The core of the second stage of Lewis's program is that what the first stage provides, in effect, is a non-trivial target for empirical investigation: in this case, investigation of *what it is,* in fact, that plays the causal role R.

2.4 The role of causation

Note that there is always a trivial answer to the above question, viz., that it is precisely mental state M that plays causal role R. The second stage of Lewis's program can be construed as offering us a guarantee that there is a non-trivial answer, and prescription for finding it. We are interested in highlighting the role that causation plays in underwriting this guarantee. To this end, it will be helpful to step back a little, conceptually and historically, and consider the original incarnation of Lewis's proposal from "An Argument for the Identity Theory" (Lewis 1966). Here, as the title of Lewis's paper indicates, the emphasis is not so much on the particular identification of mental states with brain states, or even on the method for finding such an identification, but on the guarantee that there is some such identification to be found: in other words, on the general argument for physicalism about the mental.

The structure of Lewis's paper makes it explicit that his argument for materialism about mental states has two premises. And as he says, "[t]he first of my two premises for establishing the identity theory is the principle that the definitive characteristic of any experience as such is its causal role." (1966, p. 19) In other words, the first premise is simply what later comes to be formalised in terms of the Ramsey-Carnap technique. Lewis goes on to emphasise (1966, p. 17) that this premise is not in itself a materialist premise: on what experiences actually are, it is neutral, so long as they play the required causal roles.

As for the second premise, Lewis characterises it as "the explanatory adequacy of physics". (1966, p. 23) It is the principle that "there is some unified body of scientific theory, of the sort we now accept, which together provide a true and exhaustive account of all physical phenomena." (1966, p. 23) As Lewis goes on toexplain, to assume this principle is not to assume physicalism itself: "My second premise does not rule out the existence of nonphysical phenomena; it is not an ontological thesis in its own right. It only denies that we need ever explain physical phenomena by nonphysical ones." (pp. 23–24) The crucial point is that in the light of the first premise, "none of these nonphysical phenomena can be experiences." Why? Because "they must be entirely inefficacious with respect to all physical phenomena", whereas the first premise tells us that experiences are not inefficacious in this way: on the contrary, they have physical causes and effects, such as bodily stimuli and behaviour.

In general, then, one thing that the second stage of Lewis's program can be taken to provide is an argument for physicalism about the domain in question, constructed by analogy with the mental case. We stress that this kind of application of the program does not *presuppose* physicalism. Rather, it argues *for* physicalism, by invoking two premises: first, the claim that the entities in question are characterisable in terms of their causal roles; and second, what amount to a physicalist principle about causation itself, to the effect that non-physical things do not have physical effects. If the program is to be applied in this way, then, causation plays an essential role: it is at the core of the crucial second premise of the argument;

which means, in turn, that it is vital that the theoretical roles identified at the first stage of the Lewis program be *causal* roles. (Later we want to ask whether the Canberra generalisation of Lewis's program claims an analogous second stage; and if so, what plays the role that causation plays in Lewis's program.)

The second stage of Lewis's program can also be taken in a less general way. As we noted above, the identification of a state M as the occupant of some role R does not in itself yield a non-trivial identification—after all, the best answer to the question "What plays the R role?" might be simply, "Why, M, of course!" The second stage of Lewis's program offers us something more than this trivial answer. Because the first stage of the program identifies the entities in question (originally, mental entities) as occupants of (physically-efficacious) *causal* roles, and the second premise assures us that the study of such causal roles lies within the scope of the physical sciences, their combination gives us a route to a non-trivial identification, at least in principle: roughly, it tells us that physics will get us there.

Thus, to summarise, we have distinguished two versions of the goal of the second stage of Lewis's program. The first and more general version offers an argument for physicalism about the domain in question. The second version offers what we might call a particular technique for theoretical identifications: it instructs us to investigate the nature of the occupants of those causal roles delineated in the first stage of the program. The second relies on the first, in the sense that it is the general argument for physicalism, or at least the causal closure principle on which that argument relies, that guarantees that the methods required are essentially those of the physical sciences in general.

We have noted that causal notions play an absolutely central role in both versions. Without an appeal to causation, there would be no general second stage to Lewis's program. The first stage could stand alone, of course, as an account of the meaning of the T-terms in question, but it would not yield an argument for physicalism, and it would not provide any general method for making theoretical identifications. (At least, it would not yield a method beyond the obvious one, viz., the recommendation to look for the x such that T[x], where ∃xT[x] is the Ramsey sentence in question.)

It will be helpful to have a name for the role that causation plays here. Since it is providing a hook, or tag, in terms of which these issues of identity can be addressed by the second stage of Lewis's program, we shall say that it is an *ID tag*. Later, we will be interested in the question as to whether semantic notions are required as ID tags, in the globalised version of Lewis's program.

3. To the Canberra Station: the Semantic Route From Lewis to Jackson

As we have already noted, the essence of the Canberra Plan consists in an extension of Lewis's program to a range of cases not envisaged by Lewis himself. For example, Frank Jackson (1998) invokes Lewis's model to provide an account of the nature of moral properties. Again, the program has two stages. First, Jackson asks us to consider how current folk morality, as reflected in our intuitions about how descriptive and moral terms are interconnected, might develop into a mature folk morality in the limit of critical reflection. The key postulate of mature folk morality will then consist of a conjunction of all the platitudes that describe the relationships between non-moral descriptions of situation and moral descriptions, the interconnections between moral descriptions, and the relationship of moral judgements to motivation and behaviour. If one introduces variables for the names of all moral properties and binds them with quantifiers, one obtains the modified Ramsey-sentence. Then, by adopting the modified Carnap-sentence as a meaning postulate for mature folk morality, one can formulate definitions of moral terms by reference to their functional role in this theory.

Jackson's ambitions are not limited to this first-stage application of Lewis's model, however. On the contrary, he argues that his doctrine of "moral functionalism"lends itself to theoretical *identification* of ethical properties with descriptive ones. He considers, for example, the possibility that the best solution to the equations of mature folk morality might be ones that yield *a posteriori* identifications of rightness with maximising expected hedonic value and goodness with positive expected hedonic value (1998, p. 142).

However, as Jackson himself notes, the functional roles of moral properties are not typically *causal* roles. The platitudes of mature folk morality are rarely couched in causal terms, apparently. For example, Jackson writes, one platitude might be that a fair division of some good is, other things being equal, morally better than an unfair division—but that does not mean that being fair *causes* things to be morally better. (1998, p. 131)

In the light of this fact—i.e., the fact that in this case the relevant functional roles are not *causal* roles—we want to ask the following question: What plays the general methodological role in this case that causation plays in Lewis's less ambitious program, viz., as we put it, the role of an ID tag?

One more example, before we return to this issue. Once one allows that the technique for giving functional definitions need not appeal to causal roles, one must allow that the technique may be applied to the causal relation itself. Indeed, several writers, including one of the present authors, have explored this possibility. (Menzies 1996; Tooley 1987) Thus, if one thinks that we have a folk theory of causation that implicitly treats causal terms as theoretical terms, one might apply the Ramsey-Carnap-Lewis technique in the familiar way to define them explicitly. One such application might yield the definition that causation is the intrinsic relation that typically accompanies counterfactual dependence between distinct events. (Menzies 1996) It might be hoped that such a definition could also be harnessed to provide an *a posteriori* identification of causation with some other physically specifiable relation, say energy-momentum. But again, the question arises as to what ID tag could mediate this identification.

3.1 The search for ID tags

Once again, it is important to emphasise that like its Lewisian parent, the Canberra version of the Ramsey-Carnap-Lewis technique can be deployed in ways which are more or less ambitious, metaphysically speaking. The metaphysically unambitious deployment utilises only the first stage of the technique, in the service of a neutral account of the meaning of theoretical terms. As in the case of the Lewisian version, it is evident that this stage does not require substantial semantic notions—semantic references can be understood in a deflationary spirit.

Our concern is with the metaphysically ambitious deployment of the generalised program—the deployment that counts theoretical identification among its ambitions (in either the more or the less general senses, as we distinguished them in §2.3—the general argument for physicalism concerning the area in question, or the particular prescription for *a posteriori* identifications). We saw that in the Lewisian case, this deployment relies on causation as an ID tag. Hence our question—What, if anything, can play this role in the ambitious version of the generalised program?

It seems to us that there is only possible answer to this question. Causation gets replaced by one or more semantic notions, such as reference or satisfaction: semantic notions come to provide the ID tags, providing the basis for a general argument for physicalism and the mediating links for particular *a posteriori* identifications. For example, let us suppose that we wish to produce a completely general argument for physicalism, without recourse to causal ID tags. Then it seems that the only candidate for a completely general argument would be one like the following, which makes essential use of semantic ID tags:

> t is the referent of the term "the unique x such that T[x]"
> All referents are physical entities
> Therefore, t is a physical entity.

Here the first premise uses the Ramsey-Carnap-Lewis technique to define the meaning of the term 't' by way of its role in a relevant theory T, with the explicit addition of a semantic tag. The second premise is a general premise that plays the role in this argument that the physical causal closure principle played in Lewis's original argument for physicalism. As we saw in §2.4, the causal closure principle is able to act as a premise in an argument for physicalism because it does not presuppose physicalism: it leaves open the possibility, for example, that there are non-physical entities that do not have any effect on the physical. Correspondingly, the second premise in the argument above does not itself presuppose the physicalist conclusion it purports to establish: it leaves open the possibility that there are non-physical entities that are not the referents of any theoretical terms. Nonetheless, the principle is sufficiently general that it can support the conclusion that anything referred to by a theoretical term is a physical entity.

As we say, this seems to be the only kind of argument available to an advocate of the Canberra Plan who wishes to generalise Lewis's original argument for physicalism. Unless semantic ID tags are to replace causal ID tags, there seems to be no prospect of a general analogue of the crucial second premise of Lewis's argument. (Canberra Planners might well invoke some other argument for physicalism, of course. Our point is simply that without a suitable ID tag to replace causation, such an argument cannot be an analogue of Lewis's argument for materialism.)

Thus if semantic notions did not play the role of ID tags in the Canberra Plan's extended Lewisian program, the Plan could not provide an analogue of what we called the general version of the second stage of Lewis's own program: a schema for an argument for physicalism, applicable wherever the program itself is applicable. This leaves the question as to whether the program could provide an interesting analogue of the second version of the second stage of Lewis's own program: viz., a schema for a method of *a posteriori* identifications. Here, Lewis's principle of the causal closure of the physical world offered us the advice, in effect, that we should always look to physics to find the occupants of the relevant causal roles—and again, the generality of the prescription rests on the fact that causation is playing the role of an ID tag. If there is to be no general substitute for causation as an ID tag in the extended version of the Lewisian program, there can be no general prescription of this kind. Instead, what we are left with is something like this:

> To find out what t is, write down what you know about t in the form "t is the unique x such that T[x]"; and then ask yourself what is the thing x such that T[x].

There is no doubt that that this is a general prescription *of some kind*. What is questionable is whether it provides us with anything new or non-trivial; or at any rate, anything whose novel or non-trivial elements amount to more than the application of the formal Ramsey-Carnap technique itself—which, as we have stressed, belongs to the uncontroversial first stage of the Lewisian and Canberra programs.

Thus we conclude that the unambitious, one-stage version of the Canberra program does not require robust, nondeflationary semantic notions; but that the second stage does require such notions, to play the role of ID tags, if the program is to offer what Lewis's program offers at this second stage: a general schematic argument for physicalism, and a non-trivial methodology for theoretical identification, on a case-by-case basis.

With these conclusions in hand, we now turn to their consequences. As we have already noted, there are two main questions to consider. The first is whether the Canberra program can be applied to the metaphysics of the semantic relations themselves (or whether, on the contrary, the role that semantic notions are required to play as ID tags stands in the way of the application of the Canberra technique to these notions themselves). The second is whether there is some general difficulty in relying on semantic notions in this way—a difficulty which would be problematic in general, and not merely in the case of any proposed application to the case of the semantic notions themselves.

4. Semantics on the Canberra Plan?

4.1 The local case

We have seen in the last section that the application of the Ramsey-Carnap-Lewis technique to a theory presupposes that there is a division between the T-terms introduced by the theory and the O-terms understood before the theory's introduction. The technique relies on such a division because the T-terms are defined in terms of O-terms, taken in conjunction with the logical vocabulary. We have also seen that the set of T-terms and O-terms vary from one theory to another. The term for the causal relation may be an O-term in the theory of folk psychology, but it is a T-term in the folk theory of causation.

We have also seen how the practitioners of the Canberra Plan have progressively extended the technique to encompass a broader range of terms beyond those mentioned in explicit scientific theories. Following Lewis's application of the technique to folk psychology, they have applied it to folk theories of colours, causation, and ethics. Jackson also suggests applying it to the case of the semantic properties themselves: indeed, this is his very first example of the kind of "location problem" to which he takes the methodology to be applicable, in Jackson (1998). We are now interested in the status of such an application, in the light of our conclusions in the previous section. (For simplicity we assume that the semantic notions stand or fall together in this respect—in practice, clearly, there is scope for defining some of them in terms of others, but at some point, our issue will arise for whichever notion or notions are treated as basic.)

The key issue now is whether semantic notions can function as ID tags, in a generalised Lewis approach to the identification of semantic properties and relations. And the answer, fairly obviously, is that they cannot. Why? Simply because the method requires the elimination of the target T-vocabulary, which would not be possible if the target semantic terms are themselves part of the O-vocabulary.

This point has perhaps been obscured by the naturalness of a harmless (because eliminable) use of disquotational semantic vocabulary,.Thus if ∃xSat[x] is the Ramsey-sentence for the T-term "satisfaction" we may say harmlessly, that satisfaction is the unique x (if such there be) such that x *satisfies* Sat[x]. The reason this is harmless is that the italicised use of “satisfies” is entirely eliminable. We may say, equivalently (if less elegantly), that satisfaction is the unique x (if such there be) such Sat[x]. Once the first stage of the application of the Ramsey-Carnap-Lewis technique to “satisfaction” is characterised in this latter way, it is apparent that there is no scope for the notion of satisfaction itself to function as an ID tag—to provide the necessary “hook” for a further *a posteriori* identification of the satisfaction relation.

Note that this doesn’t imply that there can be no metaphysics of satisfaction; only that such a metaphysics cannot avail itself of the generalised Lewis-Jackson methodology, if this is conceived as invoking semantic relations for the job for which Lewis’s program invoked causal relations. We still have the “thin” version of the program, the kind we mentioned above. This doesn’t offer a new technique for the metaphysics of semantic properties (apart from the first-stage Ramsey-Carnap technique itself); but it isn’t incompatible with old techniques.

4.2 The global case

The second question we raised at the end of §3 is whether there is any *general* difficulty in relying on semantic notions as ID tags in the generalised -Lewis program—any difficulty which might be problematic in general, and not merely for the proposed application to the case of the semantic notions themselves. We want to raise two points of this kind.

4.2.1 STICH’S PROBLEM: For the first problem, we turn to an excellent discussion of a closely related issue by Stephen Stich, in the first chapter of his *Deconstructing the Mind* (Stich 1996). Stich is concerned with eliminativism about folk psychological notions such as belief and desire. He notes that many philosophers take the eliminativist thesis to be the view that the terms “belief””and “desire” *do not refer*. But if that is how eliminativism is to be characterised, Stich argues, then in order to assess it we need a theory of reference—a theory capable of guiding our judgement about whether these terms do succeed in referring.

Stich argues that this leaves metaphysics hostage to the inevitable indeterminacies in a scientific theory of reference. In other words, it means that we can’t decide whether eliminativism is true until we sort out the issue between competing theories of reference—and that's likely to mean 'never', given the nature of scientific theory. (The threat of deflationism also lurks in the background here, of course, but we leave that aside.) Even worse, it would seem that in crucial cases, the metaphysics needs to *precede* the theory of reference. In order to decide what relation reference is, presumably, we need to be able to examine typical cases. In other words, we need to be able to study the various relationships that obtain between words or thoughts on the one side, and the items to which they (supposedly) refer on the other. But how can we do this in the case of 'belief' and 'desire', while it is in doubt whether these terms refer to anything? In order to know where to look, we'd have to know not only *that* they refer, but also *to what*.

Thus we have two problems for eliminativism, if it is to rely on semantic relations such as reference: the *referential indeterminacy problem* and the *precedence problem,* as we might call them. Clearly, both are problems not simply for eliminativism, but for any metaphysical view which relies on reference in this way. Both problems apply just as much if the question is "What is belief?", if this is to be understood as "To what does the term 'belief' refer?", as they do to the question "Are there beliefs?", understood as "Does the term 'belief' refer to anything?" Thus they apply in particular to any version of the generalised Lewis program which seeks to employ reference as an ID tag.

Stich's own response to the problem is to abandon semantics, and ask the relevant metaphysical questions in material form: "Are there beliefs?", in place of "Does the term 'belief' refer to anything?" for example. This certainly seems the appropriate response in some cases (folk psychology might be more controversial than Stich thinks, perhaps, but chemistry isn't, for example). But a Canberra Planner could only follow Stich down this path at the cost of abandoning the ambitious program in which (as we explained in §3) semantic notions are required to play the role that causal notions play in Lewis's original program.

In summary, then, the first general problem with relying on semantic notions for as ID tags, is that these notions are not capable of bearing the weight, in at least two senses:

1. *Referential indeterminacy.* There is no realistic prospect that a theory of semantic notions will ever be sufficiently well-founded, or sufficiently uncontroversial, to provide what causal notions arguably do provide in

the original Lewisian cases, viz., a practical basis for *a posteriori* investigation of identity questions.

2. *Priority*. One source of the indeterminacy problem is that in the cases that matter, our knowledge of the relevant facts about semantic matters is inevitably subordinate to our knowledge of the relevant metaphysical matters. It is crying for the moon to suppose that we might reverse things, and use semantics as our guide to metaphysics.

4.2.2 THE CIRCULARITY PROBLEMS: The second class of problems for relying on semantic notions in the role of ID tags in the generalised Lewis program turn on the fact that (as we have already observed), these notions cannot play such a role in their own case. In other words, a version of the program founded on these semantic notions cannot turn its own spotlight on the semantic notions themselves. (This was a simple logical point: if the semantic terms are among the T-terms, they are not available as ID tags.) In our view, there are at least two ways to argue that this consequence is unsatisfactory.

The first is a methodological point. Interpreted in such a way as to rely on semantic notions as ID tags, the Canberra program is offering us a conception of the task of metaphysics: the task, in effect, is to investigate the referents of our words and thoughts, after due regimentation by Ramsey-Carnapmethods. However, we have seen that this task is incoherent in the case of the semantic notions themselves. Either, therefore, the conception of the task of metaphysics is flawed, or metaphysics is essentially uncompletable—inapplicable just where it matters most, in fact, given the role of the semantic notions in grounding the entire program.

The second argument attempts to give the first a little more bite. It points out that in effect, this version of the generalised program offers us a semantic criterion for realism about any metaphysical matter: to be a realist about a domain is to believe that the terms characteristic of the domain in question do succeed in referring. A little more generally, the approach gives us a semantically-characterised account of *what the issues are* for what Jackson calls 'serious' metaphysics. Concerning any target subject matter, the key issues are *whether* the terms characteristic of that subject matter succeed in referring, and if so, *to what*. The problem is that this account is not applicable in the case of the semantic notions themselves. We can adopt some other account of the metaphysical issue in the case of the semantic notions, of course—as we noted above, we can simply ask whether such relations exist, and if so, what they are. But this involves turning away from the semantically-characterised criterion we have been offered for other cases.

One manifestation of this difficulty is the incoherence of irrealism about semantic properties, if irrealism is understood in the "failure of reference" way. This problem is noted by Boghossian (1990), for example. *Pace* Boghossian, however, it does not provide a transcendental argument for realism about semantic properties. For it doesn't exclude other kinds of irrealism about semantic properties, such as the materially-constituted view that there are no such things, or simply the deflationist view that there are no substantial semantic relations. (What is incoherent, in each case, is going on to try to express these forms of irrealism in the semantic fashion.)

4.3 The best-case scenario?

These arguments leave space for one "strong" version of the Canberra program. This version would accept that the program relies on the semantic notions as ID tags, and therefore that it isn't applicable to the semantic notions themselves; but would claim that it is nevertheless an advantage to be able to put all our other metaphysical eggs in the semantic basket, as it were—to conduct the metaphysics of all other topics in semantically-mediated vocabulary. This view is not incoherent, in our view, and would have a legitimate claim to be a descendant of the original Lewisian program, in its strong form. However, there seems little prospect that it could yield dividends comparable to those of the original Lewisian program, for the reasons we have noted. (In essence, they are the reasons identified by Stich: whatever the merits of Lewis's injunction to investigate causes via physics, investigating reference relations is a much less promising project.) And it could not claim the attractions of being a *global* program for metaphysics, because it cannot apply to its own foundations.

5. Conclusion

We have argued that the alternative to this strong version of the Canberra program is a version significantly weaker than Lewis's original: weaker precisely in the sense that in lacking any general alternative to causation to play the role of an ID tag, it is necessarily less ambitious at the second stage of Lewis's program—at the stage at which we move from conceptual analysis to metaphysics, in effect. Here, Lewis's causal closure principle offers a general schematic argument for physicalism about any domain to which his version of the program is applicable, and a general technique for a posteriori identifications. Without an alternative ID tag to call its own, the Canberra program is inevitably weaker at this point.

Thus we conclude that the lineage from Lewis's program to its Canberra descendants actually leads to a fork. Follow one branch, and we reach a robust but somewhat unappealing metaphysical descendant, irredeemably dependent on semantic foundations which are inaccessible both to its own methods and to useful *a posteriori* investigation. Follow the other branch, and we reach a different descendant—healthier than its sibling, to be sure, in not being dependent on something that lies forever out of reach, but weaker and less ambitious than its famous parent. We will not express an opinion as to which has the better claim to be Lewis's true heir, but we want to insist that they cannot share the mantle.

References

Boghossian, Paul, 1990: "The Status of Content", *Philosophical Review* 99, 157–184.

Jackson, Frank, 1998: *From Metaphysics to Ethics,* Oxford: Clarendon Press.

Lewis, David, 1966: "An Argument for the Identity Theory", *Journal of Philosophy* 63, 17–25.

———1970: "How to Define Theoretical Terms" *Journal of Philosophy* 67, 427–446.

———1972: "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy* 50, 249–258. Reprinted in his *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press, 1999.

Menzies, Peter, 1996: "Probabilistic Causation and the Pre-emption Problem", *Mind*, 104, 85–117.

Stich, Stephen, 1996: *Deconstructing the Mind*, New York: Oxford University Press.

Tooley, Michael, 1987: *Causation: A Realist Approach*, Oxford: Clarendon Press.