Functionalism and Self-Consciousness

MARK McCULLAGH

Abstract: I offer a philosophically well-motivated work-around for a problem that George Bealer ('Self-consciousness', *Philosophical Review* v. 106, 1997) has identified, which he claims is fatal to functionalism. The problem concerns how to generate a satisfactory Ramsey sentence of a psychological theory in which mental predicates occur within the scopes of other mental predicates. My central claim is that the functional roles in terms of which a creature capable of self-consciousness identifies her own mental states must be roles that items could play within creatures whose psychology is less complex than her own.

1. Introduction

Most philosophers would be surprised by the claim that Frege's distinction between sense and reference is fatal to functionalism. Yet George Bealer (1997) has ingeniously argued that it is. His argument opens up an issue not yet given serious attention by functionalists. In this paper I shall propose one way—indeed, it seems to me the only way—for the functionalist to respond to Bealer's argument. That response involves the claim that a creature capable of self-conscious awareness must identify her own mental states as playing roles that could be played within creatures whose psychology is less complex than her own. This is further testament to the startling nature of Bealer's argument, for we should not have thought that functionalism's prospects turned on such a claim. When we reflect on the ontological point of functionalism, though, we will see that that claim is one a functionalist should make.

2. Ramsification

Bealer's target is the functionalist's claim that mental properties—of being in pain, of thinking that Paris is in France—are ontologically second-order with respect to physical properties, meaning that they may be defined in terms either of physical properties or quantification over physical properties.

Justifying that claim requires either supplying the requisite definitions or

For helpful comments on conference presentations of predecessors of this paper I am very grateful to Victoria Berdon (Mid-South Philosophy Conference, February 1999) and Joe Moore (Central APA, May 1999), and to audiences at those talks.

Address for correspondence: Department of Philosophy, Southern Methodist University, Dallas, TX 75275-0142, USA.
Email: mcculla@mail.smu.edu

Mind & Language, Vol. 15 No. 5 November 2000, pp. 481–499.

specifying a recipe for doing so. Lacking an agreed-upon psychological theory on the basis of which to do the former, functionalists have done the latter. The recipe is in its essentials due to Ramsey (1929); it was later elaborated upon by Martin (1966), Lewis (1970) and others. Simplifying considerably, it may be broken down into four steps (but see the just-cited works for more details; I follow Bealer's treatment especially). First, take the psychological theory with which one is working, and reformulate it from English into one long sentence in the symbolism of first-order logic. Second, replace the distinct mental predicate letters—the letters used to replace 'believes', 'hears', 'is in pain' and so on in the theory as stated in English—with distinct predicate variables. Third, prefix the resulting open sentence with existential quantifiers binding the predicate variables. Finally, interpret those quantifiers as ranging over the class of physical properties (somehow conceived). The result is a theory, stated in the symbolism of second-order logic, that captures in physical terms the 'functional roles' of the mental states specified by the original English psychological theory.[1] That theory is thereby shown to be a theory of states and relations that are second-order in just the sense the functionalist claims.

Truth-conditionally adequate definitions of the individual mental properties may be constructed from the Ramsified theory in a similarly mechanical way (Bealer 1997, p. 74, following Martin). For we may say of any mental property M specified in a psychological theory, that an agent X has M just in case (i) there are physical properties $P_1 \ldots P_n$ that satisfy the Ramsified version of the theory; and (ii) among those properties, X has the one that corresponds to M in the theory.

3. The Problem

Bealer's central point is that when we Ramsify a psychological theory that states facts characteristic of creatures capable of self-conscious awareness, we get definitions that systematically assign the wrong contents to certain mental properties. A creature who is self-consciously aware that she is thinking of Paris should have *that she is thinking of Paris* assigned as the content of that

[1] We can illustrate this with the following psychological statement.

For any x, if x is near George, then George sees x.

Assume that *being near* something is a physical relation $N(\ ,\ )$. We can then partially formalize as follows.

$(x)(N(x, \text{George}) \supset \text{George sees } x)$

Following the procedure sketched in the text gives the following Ramsification in which there is no use of a psychological predicate: the existential quantification is over *physical* properties.

$(\exists R_1)(x)(N(x, \text{George}) \supset R_1(\text{George}, x))$

Of course, the existence of the relation N makes this true, so for all this statement on its own tells us, being near something *is* seeing it. But the example is meant merely to illustrate the Ramsification procedure.

state; yet, Bealer argues, a Ramsified psychological theory is going to assign to it a different content, *that she is F*, where *being F* is the physical property that in her plays the role of the mental property of thinking of Paris.[2] As Bealer puts it:

> the envisaged functional definitions would require the wrong sorts of things to be the contents of self-consciousness: the contents would have to be propositions involving these 'realizations' rather than the mental properties themselves. (1997, p. 69)

How does Bealer arrive at this conclusion? Consider a psychological theory *A* that includes statements of facts characteristic of creatures capable of self-conscious awareness. Let us agree with Bealer that among the statements in *A* is the following, which he calls principle *P*:

> **Principle *P*:**
> If a person is in pain and engaging in introspection, that person will be self-consciously aware that he is in pain. (1997, p. 77)

Bealer claims that principle *P* reflects an essential fact about the nature of self-conscious awareness, namely that episodes of it are brought about by episodes of introspection. It is a claim about the genesis of such episodes; and it does seem to get at something that's special about them. I shall not dispute Bealer's claim that for a theory to be capable of adequately characterizing the mental lives of creatures capable of self-conscious awareness, it must contain principle *P* or some statement very like it.

Now consider what happens to principle *P* when we Ramsify this theory so as to arrive at a definition of self-conscious awareness. Bealer claims that we get the following, which I shall refer to as definition *D*:

> **Definition *D*:**
> **x is self-consciously aware that** *p* iff_{def} there exist first-order properties **R** [= $<R_1 \ldots R_n>$] such that
> (i) they satisfy *B* [the overall theory *A* minus principle *P*];
> (ii) if x is R_1 [which plays the role of pain] and R_3 [introspection], then x will be related by R_4 [self-conscious awareness] to the proposition that he is R_1;
> (iii) x is related by R_4 to the proposition that *p*. (Bealer 1997, p. 78; my annotations)

Clause (ii) in definition *D* is the one that supports Bealer's central point. Its

---

[2] I'll often speak indifferently of 'being in a mental state' rather than 'having a mental property'; nothing turns on the difference.

presence in the definition is due to the presence of principle *P* in the psychological theory. Clause (ii) says that an introspecting subject in pain is self-consciously aware not of the proposition *that he is in pain* but of the proposition *that he is* R_1. But that's not what an introspecting subject in pain is self-consciously aware of. For as Bealer claims, and I shall agree, someone is self-consciously aware that they are *F* only if *being F* is a mental property (1997, p. 79). So clause (ii) assigns the wrong contents to states of self-conscious awareness. Bealer concludes that the functionalist claim about the psychological theory's ontological import is false *because* that theory specifies facts characteristic of self-conscious awareness.

This is a new sort of objection to functionalism. For it does not rule out the possibility that functionalism gives a correct account of the ontology of the mental states of those creatures, if such there be, whose mental lives can adequately be characterized by a psychological theory in which no mental predicate occurs in a 'that'-clause governed by a mental predicate. It is thus different in scope from e.g. qualia-based objections, since the philosophers, who press those objections believe that any creature with a mental life has qualia-laden experiences. (There is something it's like even to be a bat, as Nagel (1974) argued.) In Bealer's argument the functionalist faces an entirely new sort of challenge.

I wish to agree with Bealer on the following claim. *If we follow the Ramsification procedure*, we get definition *D* of self-conscious awareness, and definition *D* must be rejected since it assigns the wrong contents to episodes of self-conscious awareness. This is giving Bealer some rope, for it is not obvious that this is something to which we must agree. We can't be sure that our psychological theory will employ *de dicto* characterizations of mental states, such as principle *P* employs in its consequent. After all, that theory is presumably some polished product of mature psychological theorizing, and we are not entitled to assume that it will characterize mental properties as we usually do. Perhaps it will employ only *de re* characterizations, in which case rather than principle *P* we would have:

> If a person is in pain and engaging in introspection, that person will be self-consciously aware of a pain state.

If that is what we work with, then rather than clause (ii) as in definition *D*, we would have the following:

> if x is R_1 [which plays the role of pain] and R_3 [introspection], then x will be related by R_4 [self-conscious awareness] to an R_1 state.

And it is not at all clear that this is unacceptable. For it is not implausible to say that when one is aware *that* one is in pain, one's awareness is *of* the physical property that plays the functional role of being in pain. Nonetheless, to sum

up my concessions to Bealer thus far, I shall suppose that: our psychological theory does contain something like principle *P*; when it is Ramsified it does issue in definition *D*; and definition *D* does assign incorrect contents to episodes of self-conscious awareness.

4. There are Other Ways of Doing What Ramsification Does

What makes definition *D* unacceptable is that it results from replacing *all* occurrences of 'is in pain' with the same predicate variable, 'R_1': both those that are unembedded within 'that'-clauses governed by mental predicates, and those that are so embedded. Granted, Ramsification requires doing that. But is there any independent reason for doing it? One could, of course, say that the functionalist *must* employ Ramsification in order to justify her claim about the ontological import of psychological theories. But there are two reasons why that cannot be right.

The first is that English contains ambiguous predicates; so our theory, which is stated in English, might contain ambiguous predicates. And we would certainly not replace differently-disambiguating occurrences of such a predicate with the same predicate letter, when reformulating our theory in logical symbolism. We would use different predicate letters, and therefore, when replacing the predicate letters with predicate variables in the third step of Ramsification, different predicate variables. Mere lexical identity is not enough to require using the same predicate variable to replace different occurrences of the same predicate expression. This means that Ramsification could be mechanically applied only to a psychological theory none of whose predicates is ambiguous. So there is already a class of cases in which we should not mechanically follow the Ramsification procedure.

But even if we knew that our theory contains no ambiguous predicates, we still should not say that the functionalist is required to employ Ramsification. For Ramsification, recall, is simply one recipe for vindicating the functionalist's claim about the ontological status of the properties picked out by the mental-state predicates in the theory. There is no reason to suppose that Ramsification is the only means of performing that task. Ramsification is certainly a powerful and philosophically interesting technique, but it is certainly not the only possible technique for recasting a theory in such a way that it involves quantification over, or reference to, only physical properties. What the functionalist must show is that that can be done with a psychological theory; she is not committed at the outset to one particular recipe for doing it. So although she might agree that when Ramsified, the theory issues in definition *D*, she might propose that the theory's ontological commitments be made explicit by some technique other than Ramsification.

At least in the abstract, then, we have a clue as to how a reply to Bealer's argument would go. Such a reply would consist in showing how, otherwise than by Ramsification, a psychological theory can be reformulated so as to

involve quantification over, or reference to, only physical properties. Below I shall be suggesting one such non-standard recipe. But before moving on to that proposal, we must consider the reason Bealer adduces in support of the claim that unembedded and embedded occurrences must be formalized alike. (I'll call that the **Formalization Claim**.) For he does not rest it on the bare assumption that Ramsification is the method being used.

5. Bealer's Move: From Semantics to Formalization

This is where Frege's distinction between sense and reference comes in. Bealer claims that when we reformulate the psychological theory so as to make explicit the fact that the properties it's about are ontologically second-order with respect to physical properties, we must replace *all* occurrences of any mental predicate with the *same* predicate variable once we appreciate certain facts about the semantic functioning of the unembedded and the embedded occurrences of that predicate. Which facts are these? Here is Bealer:

> in principle *P* the proposition *denoted* by the 'that'-clause 'that he is in pain' is the same as the proposition *expressed* by the unembedded antecedent 'he is in pain'. In their Ramsification of *P*, our functionalists should therefore use one and the same predicate variable 'R1' to replace both the embedded and the unembedded occurrences of 'is in pain'. (1997, p. 82; also p. 78 n. 11)

In familiar Fregean terms, the claim is that we should replace all occurrences of the predicate with the same predicate variable because the unembedded occurrences directly denote what the embedded occurrences indirectly denote.

Let us agree with this Fregean claim that the unembedded occurrences denote directly what the embedded occurrences denote indirectly. Call that the **Semantic Claim**. Now we may ask: does the Formalization Claim follow from the Semantic Claim, as Bealer holds?

Far from it: the Semantic Claim justifies a *denial* of the Formalization Claim. For the Semantic Claim straightforwardly entails that the denotations of the embedded occurrences differ from the denotations of the unembedded occurrences. And it is a principle of good formalization from English into logical symbolism, that expressions with different denotations be symbolized using different letters. After all, to symbolize them with the same letters is to ensure that they receive the same interpretation; but the Semantic Claim says that the embedded occurrences should be interpreted differently from the unembedded occurrences.

Decisive as that consideration may seem, there are, nonetheless, some grounds on which to claim that this cannot be the right story about how to Ramsify theories in which mental predicates occur both unembedded and embedded. One might agree that using the same predicate variable to replace

differently-denoting occurrences of a predicate violates good logical practice. But one could claim that this violation is justified nonetheless in this case, once we appreciate what gets obscured when we use different predicate variables, as the previous paragraph argued we should. When we replace unembedded occurrences of 'is in pain' with 'R_1' and embedded occurrences with 'R_{26}', say, we do indeed end up with a formalization that reflects the fact that the former are doing different semantic work from the latter. But it is a formalization that *fails* to reflect the fact that the former are doing semantic work that is intimately related to the semantic work of the latter. After all, the fact that the semantic functioning of these occurrences is related in this way contributes essentially to the plausibility that principle *P* had when we first considered it. Principle *P* would have had no plausibility whatever, had it said something like this:

> If a person is in pain and engaging in introspection, that person will be self-consciously aware that they are in shmain,

where all we knew about 'is in shmain' is that it picks out a different property from 'is in pain'.

The Formalization Claim is best seen, I think, not as supported solely by the Semantic Claim but in addition by the claim that *there is no other way besides using the same predicate variables* to reflect the fact that the semantic work that the unembedded occurrences of 'is in pain' do is intimately related to the semantic work that the embedded occurrences do. The thought is that using different predicate variables would obliterate any trace of that relation. This, it seems to me, is what Bealer is getting at by invoking the Fregean distinction between sense and reference.[3] So there are some grounds on which Bealer might defend the idea that the Semantic Claim supports the Formalization Claim, even though on general logical grounds we would have thought them incompatible.

6. Constraining Interpretations Otherwise than by Choice of Variables

Where are we now? The only claim in Bealer's argument that I have taken issue with thus far is the Formalization Claim. I've followed Bealer in supposing the ancillary claims mentioned in §3, and I've agreed with the Fregean Semantic Claim. And, even though the Semantic Claim certainly appears to justify a denial of the Formalization Claim, I've allowed (§5) that there is at least one line of thought according to which it supports it. At this point, I

[3] Bealer simply presents the Semantic Claim as justifying the Formalization Claim, without offering an explanation of how it does; so I cannot be certain that I am correctly imputing to him this take on their relation.

can begin to lay out the strategy the abstract possibility of which we discerned in §4, and explain why it scuttles that line of thought.

The thought is that *only* by replacing *all* occurrences of a given mental predicate with the same predicate variable in our transformation of the psychological theory can we get a formalization that reflects the intimate relation between the semantic work done by the unembedded occurrences and that done by the embedded occurrences. Is this true?

No. I shall start explaining why with a simple example. Suppose that we wished our formalization to reflect the fact that in a given natural-language sentence, one occurrence of an expression does the same semantic work as another. For example, suppose we were symbolizing the English sentence

Simon is friendly and Simon is honest,

in which the name 'Simon' denotes the same individual in each occurrence. Our training is to do this by using the same individual constant to replace each occurrence, giving us

$Fs \cdot Hs$

as the symbolization. But there is another way to reflect the fact we want to reflect. We could append clauses to the formalization that ensure that the two occurrences receive the same interpretation. For example, we could symbolize the sentence thus:

$(Fs \cdot Ht) \cdot s = t.$

Any model of the former symbolization is a model of the latter, and *vice versa*; in this sense, semantic content is preserved across the two formalizations. The point is simple. Choosing terms judiciously is one way to constrain the range of interpretations of the formalization of a natural-language statement; appending clauses judiciously is another.

In the example just given, we wanted to ensure that '*s*' and '*t*' receive the same interpretation, and we did this by appending a simple identity statement. When symbolizing principle *P*, however, we do not want to do that with the two occurrences of 'is in pain'. For we are supposing the truth of the Semantic Claim, which is that those two occurrences do different, though intimately related, semantic work. So it is not by appending any sort of identity statement to our formalization that we could have it constrain interpretations as we want.

But all this means is that we should try to formulate a different clause that does capture that intimate relation. In the next section I'll be suggesting how to do that. My point in this section has simply been that one may constrain interpretations otherwise than by choice of terms. So there is a false assumption built in to the line of thought I discussed at the end of §5, according to

which the Formalization Claim specifies the *only* way for a formalization of a psychological theory to reflect the fact that the semantic work that mental predicates do in their unembedded occurrences is intimately related to the semantic work they do in their embedded occurrences.

7. An Assumption about the Contents of Introspective Mental States

What the functionalist wants, then, is a modified Ramsification that avoids the unacceptable result of standard Ramsification. In this section I describe the thought behind my proposal.

Consider again principle *P*, and cognate principles having to do with thought, desire, and so on:

> If a person is thinking that *p* and engaging in introspection, that person will be self-consciously aware that he is thinking that *p*;
>
> If a person desires *x* and engages in introspection, that person will be self-consciously aware that he desires *x*;

or even, as Bealer notes (p. 85 n. 18):

> If a person is self-consciously aware that *p* and engages in introspection, that person will be self-consciously aware that he is self-consciously aware that *p*.

What should the functionalist want to make of such principles? According to the functionalist, to be in pain is to have a physical property playing the pain role; similarly for thinking, desiring, being self-consciously aware, and so on. So I shall assume that the functionalist would agree to the following.

> **Content Assumption:**
> If one believes (thinks, desires, etc.) that one has mental property *M*, one's belief (thought, desire, etc.) is *that one has a physical property that in one plays the M role*.

This assumption is tendentious. There are two objections it invites. The first is that it involves attributing too much theory to ordinary persons: it credits ordinary thinkers with the concept of a functional role, but that is a theoretical concept that only philosophers use. The second is that even those thinkers who possess that concept might not employ it as the Content Assumption specifies. For there are many ordinary thinkers who do not think about mental states as the functionalist would recommend. Many of us think of the

mental properties as somehow on an ontological par with physical properties. Those thinkers are ones whose thoughts about mental properties do not appear to have the contents the Content Assumption says they have.

There are two ways in which the functionalist could reply to these objections. The first would be to claim that each thinker has implicit commitments that support the Content Assumption. No doubt the functionalist would argue that each thinker is implicitly committed to the obtaining of causal relations between physical properties and mental properties. That is a plausible claim, and perhaps it is enough to vindicate the Content Assumption: perhaps these implicitly functionalist commitments can be shown somehow to prevail over whatever dualist commitments ordinary thinkers might have.[4]

But a second sort of reply is more relevant in our context. The functionalist could agree that it's not *clear* that self-directed mental properties have the contents the Content Assumption says they have; but he could claim that nonetheless it is *much more plausible* that they have those contents than that they have the contents that clause (ii) of definition *D* (and its cognates) assigns to them. For the former have the right truth conditions (if functionalism is true) while the latter do not. The content *that one has a physical property playing the M role* is true of any creature that has mental property *M* and false of any creature that does not, if functionalism is right about the ontology of mental properties; but the content *that one has the physical property R* is not true of thinking creatures whose mental properties are realized otherwise than by a set of physical properties of which *R* is a member.

I shall assume, then, that it would *not* be a refutation of functionalism, to show that the functionalist must assign contents to mental properties as the Content Assumption says. The functionalist might be happy to assign contents in that way, since they are, at least, equivalent in truth-condition to whatever the correct contents are (if functionalism is true).

8. A Regress Problem

How then might we specify a recipe for recasting in second-order logical symbolism a psychological theory containing embedded occurrences of mental predicates, one that respects the Content Assumption? Bealer has shown that for the functionalist, standard Ramsification will not work. So things are going to be a bit hairy. Indeed, they will be even hairier than we should expect: for

[4] For a defence of the idea that it is 'plausible to credit us, as we use psychological terms, with understanding' functional concepts, see Pettit (forthcoming). Pettit claims that there is a way of having discriminatory knowledge of an item that is neither observational nor descriptive: it consists in one's dispositions to behave as if one had discriminatory knowledge of either of those sorts. But Pettit does not undertake to show that in thinking of mental states we actually *employ* those concepts rather than, say, ones that answer to dualist commitments.

a regress problem emerges when we try to pursue the formalization strategy that the Content Assumption naturally motivates.

Let us try to do that. Recall principle *P*, occurring as part of our overall psychological theory *A*:

> If a person is in pain and engaging in introspection, that person will be self-consciously aware that he is in pain.

Ramsifying this as we did in §3, we get the following open sentence (the predicate variables being bound by the predicate quantifiers prefixing the Ramsification of *PT*).

> If x is R_1 and R_3, x will be related by R_4 to the proposition that x is R_1.

As a start on pursuing our strategy, though, we should rather want something like the following.

> If x is R_1 and R_3, x will be related by R_4 to the proposition that: x has some physical property that in x plays the pain role.

Since our predicate quantifiers are ranging over physical properties, we may change this to:

> If x is R_1 and R_3, x will be related by R_4 to the proposition that: $(\exists S)$(x is S and S plays the pain role in x).

But we still need to spell out 'plays the pain role'.

The natural thought at this point is that we should avail ourselves of our overall theory *A*, since it spells out the roles of all the mental properties. That would give the following.

> If x is R_1 and R_3, x will be related by R_4 to the proposition that: $(\exists S_1 \ldots S_n)$ (x is S_1, and $S_1 \ldots S_n$ satisfies A).

Unfortunately, this cannot work. For the only way to spell out '$S_1 \ldots S_n$ satisfies A' is by stating A. But that is something we cannot do within one of A's own clauses; and we are treating principle *P*, which is a clause in *A*.

This shows that the Content Assumption cannot be pursued in what would seem the most straightforward way. For we cannot, on pain of an infinite regress, state our entire psychological theory within one of its own clauses.[5]

---

[5] Bealer (1997, 99) discusses a functionalist proposal that is subject to the same problem as the present one, and rejects it for essentially that reason. The proposal is that we think of agents as having mental states in virtue of tokening (in various 'attitude boxes') sentences in a Language of Thought, the contents of which are determined by their conceptual roles. The thought behind the proposal is that we can get around Bealer's argument by employing a theory in which no mental predicate occurs within a 'that'-clause, but in which occur clauses that *mention* mental predicate expressions in the subject's Language of Thought. The contents

9. Identifying Mental States in Terms of Incomplete Specifications of Their Roles

One is tempted at this point just to give up on principle *P* and its fellows. But that is something the functionalist cannot do. The problem Bealer has identified concerns not so much the details of this or that particular psychological principle, but the general possibility that our psychological theory will include statements in which mental predicates occur within the scope of mental predicates. Bealer correctly claims that 'what is needed is a *systematic* response to the problem of multiply embedded psychological predicates' (1997, p. 86). The functionalist cannot rest with anything less, if she is to defend her ontological claim. The Content Assumption seemed to motivate a systematic response, but we have just found that that response is fatally flawed.

Luckily, there is a way out of this problem. It rests on the fact that in order to *identify* something for the purpose of having thoughts about it, it is not necessary to have comprehensive knowledge of it.[6] The Content Assumption says that thinkers identify their mental properties as properties that play certain roles in their mental lives. They must, then, be able to pick out those roles in thought; they must be able to identify them. But this does not entail that they must be able comprehensively to specify those roles. (Indeed, as we will see in §12, the functionalist has ontological grounds on which to deny that that is something any thinker could do.)

Applied to the present case, this means that thinkers don't need to identify their thoughts by specifying their roles as comprehensively as the overall psychological theory does. That theory is designed to supply explanations and predictions of mental states and events. We should therefore expect it to be far more comprehensive than a theory that aims at nothing more than *distinguishing* thinking that *p* from desiring that *p* or thinking that *q*, and supplying thereby a means of identifying items that play those roles. The Content Assumption entails only that agents possess such an identifying ability. The crucial point behind the proposal I shall make, then, is that agents can identify the roles their mental states play without being able to fully specify those roles (as the overall psychological theory should do).

Consider now the suggestion that we use for the role-specifying task called for by the Content Assumption, a trimmed-down sub-theory of our overall

---

of these Language of Thought expressions are assigned by other clauses in the theory; but again, these clauses only *mention* those expressions and do not contain 'that'-clauses with mental predicate in them. Thus we would have, e.g.:

> 'T' means *thinking* $\equiv_{\text{def}}$ 'T' behaves in the subject's attitude boxes in the way the predicate 'thinks' behaves in our overall psychological theory.

Here again the problem is that we can't state the overall theory within one of its own clauses.

[6] For items in different ontological categories, of course, the knowledge that is required for identification in thought will differ; and in some cases it may be knowledge-how rather than knowledge-that. P.F. Strawson's *Individuals* (1959) is the seminal discussion of this issue.

psychological theory rather than the theory itself.[7] Since our problem is being caused by the occurrence of mental predicates within the scope of mental predicates, the natural candidate to consider is this: the overall theory *A* minus all statements in which mental predicates occur within the scope of mental predicates. Call this stripped-down theory *Unembedded-A*. This theory could *fully* characterize the mental lives only of creatures that had no thoughts about thoughts, no beliefs about thoughts, no beliefs about beliefs, and so on. Call these **unreflective** creatures.

This solves our regress problem, at least. For in dealing with principle *P*, for example, we now get the following:

> If x is R_1 and x is R_3, then x will be related by R_4 to the proposition that: $(\exists S_1 \ldots S_n)$(x is S_1, and $S_1 \ldots S_n$ satisfies *Unembedded-A*).

And in spelling this out we don't get a regress of restatements of *A*, since principle *P* does not occur in *Unembedded-A*. Indeed, we eliminate the general problem altogether since, as Bealer himself points out, it is caused by embedded occurrences of mental predicates, and there are no such occurrences in *Unembedded-A*. So *Unembedded-A* is a theory we can use to spell out (within our formalization of *A*) how agents *specify* mental properties' roles well enough to *identify* those roles.

The recipe is this. First, Ramsify *Unembedded-A*; this is unproblematic since there are no embedded occurrences of mental predicates in it. Then Ramsify *A*, except for the embedded occurrences of mental predicates. These predicates are of two sorts: either they occur in *Unembedded-A*, or they do not. Here I describe the treatment of the former (among which are, surely, such predicates as 'believes', 'thinks' and 'desires'); in §11 I discuss cases of the latter sort. Assume then that our embedded mental predicate, 'is *M*', occurs in *Unembedded-A*. We then replace 'x is *M*' with

$(\exists S_1 \ldots S_n)$(x is S_1, and $S_1 \ldots S_n$ satisfies *Unembedded-A*).[8]

Here, 'S_1' is the variable used to replace 'is *M*' in the Ramsification of *Unem-*

---

[7] Of course we could use another theory entirely, even a false one, for it might be that thinkers are wrong about their own mental states and those of others. But it would be best for functionalism if it could deal with the psychologies of creatures who identify their own mental states correctly. That is the reason for using a subtheory of *A*, which is being presumed a correct theory.

[8] Or for relations, replace 'x is related by M to y', say, with '$(\exists S_1 \ldots S_n)$(x is related by S_1 to y, and $S_1 \ldots S_n$ satisfies *Unembedded-A*)'.

bedded-A. Treating all the clauses of *A* in this manner gives what I shall call the **U-theory** of *A*.[9]

This strategy applies straightforwardly to multiply embedded occurrences of mental predicates: we get multiply embedded quantifications over property-sets that realize *Unembedded-A*. Suppose that theory *A* includes the following principle.

> If a person is thinking that he is thinking that *p*, and is engaging in introspection, then that person will be self-consciously aware that he is thinking that he is thinking that *p*.

Applying our strategy to this sentence gives the following as part of the U-theory of *A*. (Here 'R_2' is the two-place predicate variable that replaces 'is thinking' in unembedded occurrences in the Ramsification of *A*; its two arguments are a subject term and a proposition term. 'S_2,' 'T_2' and 'U_2' are the variables used to replace that predicate in its occurrences in merely notationally variant Ramsifications of *Unembedded-A*.)

> If *x* is related by R_2 to the proposition that: $(\exists S_1 \ldots S_n)$(*x* is related by S_2 to the proposition that *p*, and $S_1 \ldots S_n$ satisfies *Unembedded-A*), and *x* is R_3, then *x* will be related by R_4 to the proposition that: $(\exists T_1 \ldots T_n)$(*x* is related by T_2 to the proposition that: $(\exists U_1 \ldots U_n)$(*x* is related by U_2 to the proposition that *p*, and $U_1 \ldots U_n$ satisfies *Unembedded-A*), and $T_1 \ldots T_n$ satisfies *Unembedded-A*).

This is unwieldy, to be sure, but perfectly sensible, as the following English paraphrasing of the quantifier-involving parts makes clear. Here we use 'is

---

[9] This proposal is unlike one that Bealer (1997, pp. 100–1) discusses and correctly rejects, even though their respective predecessors are similar. I noted above (n. 5) that Bealer considers a proposal with the same flaw as the one discussed in §8. The successor to that proposal that Bealer considers is the idea that the overall theory be stated in a logically typed vocabulary, giving definitions of hierarchies of psychological predicates. For example, there would be thinking$_0$ (which is of propositions about non-mental facts), thinking$_1$ (which is of propositions about either non-mental facts or facts about thinkings$_0$), thinking$_2$ (etc.), and so on. Bealer rightly rejects that proposal on the grounds that psychological relations are untyped. One *thinks* that Paris is in France; one does not select whether to *think*$_0$ that Paris is in France, or *think*$_{20}$ that Paris is in France (or even to *think*$_n$ that Paris is in France, for all *n* less than some number). But this proposal is unlike ours for two reasons. First, it has no motivation whatsoever other than the need to get around Bealer's point about Ramsification whereas our proposal is based on the philosophical point that identifying something (well enough to have thoughts about it) doesn't require knowing everything about it. Second, our proposal does not involve a type hierarchy. What *Unembedded-A* supplies, I claim, is a specification of the role of *thinking that Paris is in France* that can be used to *identify* that role—even as one that an item could play in a creature whose psychology is *more* complex than *Unembedded-A* captures. What gets identified is *thinking that Paris is in France*, simpliciter, not some typed variant of that.

related by a thinking property to' to abbreviate 'is related by a physical binary relation playing the role of "thinks that" as spelled out in *Unembedded-A* to':

> If x is related by R_2 to the proposition that x **is related by a thinking property to the proposition that** p, and x is R_3, x will be related by R_4 to the proposition that: x **is related by a thinking property to the proposition that:** x **is related by a thinking property to the proposition that** p**.**

We have arrived, then, at what Bealer himself says is required to meet his argument: 'a *systematic* response to the problem of multiply embedded psychological predicates'.

10. Evaluating the Technique

We have got around the regress problem; but is this strategy philosophically acceptable? Will the U-theory of *A* correctly assign contents to all of a creature's thoughts about its thoughts, beliefs about its thoughts, and so on? Given the role that *Unembedded-A* is being called upon to play, the question comes to this. Can thoughts, beliefs, desires, and so on be *identified in terms of*[10] roles that items could play in unreflective creatures? We should consider some cases.

First, consider pains, tickles and the like. These are mental states that unreflective creatures can have. There is no debate nowadays over whether the lower animals can feel pain; and among these animals are many whose mental lives are unpunctuated by episodes of introspection, or thoughts about beliefs. There should be nothing to trouble the functionalist, then, in the proposal that we characterize thinkers as *identifying*—not fully specifying, to be sure—the functional roles of their pains, tickles and so on, in terms of roles that items could play in, say, a gerbil.[11]

What about thoughts, beliefs and desires? Here there is room for debate. Consider the higher mammals. Perhaps they have states somewhat like thoughts but which do not, say, obey the Generality Constraint that Gareth

---

[10] I say 'identified in terms of' rather than 'identified as' because the treatment of multiply-embedded occurrences requires the former. When (as in the example in §9) an agent is characterized as identifying their *thought that they are thinking that p*, they are not characterized as identifying it as a state that an unreflective creature could be in; but they are characterized as identifying it *in terms of* states that unreflective creatures could be in, *via* embedded re-applications of the strategy described in §9.

[11] Indeed, they could be identified *as* items playing such roles, since none of these predicates takes 'that'-clause complements within which mental predicates can occur. Thus the qualification explained in n. 10 is not needed here; any pain is one whose role can be identified *as* one that an item could play in an unreflective creature.

Evans (1982) has claimed is distinctive of thoughts.[12] Perhaps Fido is able to think that his master is giving him water, or that his water is hot, yet is unable to entertain the proposition that his master is hot. But this limitation on Fido's abilities wouldn't seem to be due to his being unreflective; it is much more plausible to say that it is due to his limited inferential capacities. There doesn't seem to be anything incoherent in the idea of a very clever, inferentially agile, creature that is unreflective, yet whose thoughts obey the Generality Constraint. Similarly for beliefs and desires: the claim is that my *belief that the sun is shining*, for example, is an item whose functional role in me I can identify as the role that that belief plays in an unreflective creature. This is a claim that many philosophers would endorse. I take it, then, that it would not be a refutation of functionalism, to point out that it requires that beliefs and desires play roles that can be identified in terms of roles that items could play in the lives of unreflective creatures.

When treated in the manner of §9, then, our theory characterizes agents as *identifying* a great many of their own mental states—pains, thoughts, beliefs, desires—by specifying them in terms of roles that states could play in the mental lives of unreflective creatures. And it does this in a perfectly general way, a way that applies straightforwardly to multiple embeddings.

11. Should We Worry About the Possibility of Counterexamples?

Any objection to the proposal of §9 is going to have to come to the claim that agents cannot identify the roles of their mental properties in terms of the roles that are spelled out in *Unembedded-A*.

What suggest themselves as problematic in this regard, are the roles of episodes of self-conscious awareness. It is not immediately clear how those roles could be spelled out in terms of roles specified in *Unembedded-A*, since the predicate 'is self-consciously aware that' probably does not even occur in that theory: it would probably drop out along with embedding-involving principles, such as *P*, in which its role is spelled out. We need some such spelling-out, though, in order to replace embedded occurrences of that predicate in our treatment of the overall theory *A*. (The unembedded occurrences are of course no problem; they just get replaced by a certain predicate variable.)

This particular problem vanishes, however, when we realize that the states picked out by the predicate 'is self-consciously aware that' can just as well be picked out as *beliefs* that are caused in a certain way, and whose contents are propositions about mental states. For neither the former fact about them, nor the latter (given the fact that §9's strategy handles multiple embeddings) poses

[12] The constraint is this: 'if a subject can be credited with the thought that *a* is *F*, then he must have the conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception' (p. 104).

any problem for our forming the *U-theory* of a theory that contains, say, the following principles:

> If *x* is in pain and is introspecting *x* will believe that *x* is in pain.
>
> If *x* believes that *p* and is introspecting, *x* will believe that *x* believes that *p*.

For predicates that don't appear in *Unembedded-A*, then, the recipe of §9 cannot be mechanically applied. But this is no problem if, as with 'is self-consciously aware that', we can spell out the relevant role *in terms of* roles spelled out in *Unembedded-A*, if not *as* some role spelled out in *Unembedded-A*.

Should we worry about there being other mental predicates for which we cannot do this? Since we do not have our final overall psychological theory in hand, it is probably pointless to debate whether our strategy can handle all possible psychological theories. For one thing, we already know that that claim is false for reasons that make no trouble for functionalism. As Bealer himself points out (1997, p. 75 n. 8), functionalism would be trivially false if understood as a claim about the ontological status of properties picked out by a theory that states that some mental property is not physically realizable. But that is a problem for functionalism only if we have reason to believe that our final theory will be of that sort, and we have no such reason.[13]

More generally, though, there is the following consideration against taking this abstract worry seriously. If a predicate picks out a property whose role in a mental life cannot be specified—not even well enough for purposes of *identifying* that role in thought—in terms of beliefs, desires, pains, tickles and so on, we would have good reason to deny that it picks out a *mental* property. So there does not appear to be anything worth worrying about, in the merely abstract possibility that there exist mental states or events whose roles may not be specifiable in terms of roles that some items could play in the mental lives of unreflective creatures. We would only be justified in worrying about this, if we had reason to believe that the final theory picks out properties of which that is true. But again, we do not.[14]

---

[13] Similarly, we have no reason to insist that our final psychological theory must support characterizations of people as having mental states whose contents are Liar paradoxes (or that are infinitely self-involving in an analogous way). Indeed, to take them as paradoxical is to have one good reason to doubt whether anyone actually does have mental states with those contents. (Another argument that Bealer presents in his paper—the §3 argument against an 'ideological' functionalism that drops the ontological claim distinctive of the functionalism I am defending in this paper—presupposes the possible truth of such characterizations.)

[14] The following is a related point. We chose theory *Unembedded-A* to figure in our spelling out how thinkers identify the roles of their own mental properties. We did so because that theory contains none of the troublesome occurrences of mental predicates embedded within mental predicates; it can be straightforwardly Ramsified *à la* §2, and inserted where needed in our treatment of embedded mental predicates in *A*. But all that really matters about this theory is that it's finitely stateable in terms of physical properties. So if, under the

12. Physical Thinkers' Specifications of Their Own Thoughts' Functional Roles

My proposal rests on the claim that a creature can identify her own mental states in terms of roles that items could play in the mental lives of unreflective creatures. I claimed that this point has some solid philosophical support, namely in the claim that identifying an item (for the purpose of having thoughts about it) does not require comprehensively knowing facts about that item. But it also has some support in the ontological claim that Bealer takes as his target, and is thus especially well-motivated as supporting a reply to his argument. Let me explain.

Imagine what would be required of a thinker, were she to be capable of identifying one of her thoughts by spelling out the entire role that that thought plays in her mental life. Part of the entire role it has in her mental life is its role as the object of such an identification. So the identification would include a spelling-out of that role. But that is infinitely regressive. For spelling out the thought's entire role requires saying that it is the one which: has features $f_1 \ldots f_n$ and whose role is being specified right now—as the one that has features $f_1 \ldots f_n$ and whose role is being specified right now—as the one that has features $f_1 \ldots f_n$ and whose role is being specified right now—as ... etc. Spelling out the role of the thought requires spelling out the specification of its role; but its being specified in that fashion is *part* of its role. To specify a thought partly in terms of its being specified is to embark on an infinite regress of specifications of how it's being specified. Such an identifying process could come to an end only in a thinker that could spell out infinitely many elements of a thing's functional role in a finite period of time. But we have good reason to think that no physical thinker could do that. Every example of a physical thinker we know of—any animal, any computer—can perform only finitely many mental acts in a finite period of time (these acts being characterized in the terms of our overall psychological theory).

The conclusion is that no physically realized thinker could specify the entire functional role of any of her mental states or events. That is why the functionalist should be happy to maintain that any thinker *must* identify one of her mental states by an incomplete specification of its role in her mental life. For no physical thinker could do so with a complete specification, and the functionalist's ontological claim is that all thinkers are physical thinkers. This consideration undercuts the proposal considered in §8 just as surely as did the logical considerations; and it motivates a proposal that is along the lines of the one made in §9, according to which even reflective creatures identify their

---

pressure of worries along the lines just discussed, we decide that *Unembedded-A* is too impoverished to allow us to spell out how thinkers identify functional roles, we can choose to use a more comprehensive theory instead, as long as we can finitely state it in physical vocabulary. But again, for the reasons just discussed, there is no reason to be worried about that sort of situation as a merely abstract possibility.

own mental states in terms of the roles played by mental states in unreflective creatures. The proposal we made in §9, then, is one that a functionalist should be motivated towards not only by technical considerations about the Ramsification of her theory, but by ontological considerations having to do with the physicality of thinkers.

13. Conclusion

Bealer has shown that when we look at the details of how the functionalist is to vindicate her ontological claim about mental properties, we see a serious obstacle when it comes to the mental states of creatures that are capable of self-conscious awareness. I have argued that that obstacle can be overcome once we make two assumptions. The first is that thinkers *identify* their mental states *as* physical items playing certain functional roles. The second is that for this identifying purpose, incomplete specifications of those roles must suffice. This latter assumption, though, is one that a functionalist should accept; for no physical thinker could employ a complete, and therefore infinitely regressive, specification of the role of one of her mental states. Nonetheless it is remarkable that we had to uncover this commitment in order to vindicate functionalism in the face of Bealer's powerful and original argument.

Department of Philosophy
Southern Methodist University

References

Bealer, George. 1997: Self-consciousness. *Philosophical Review*, 106, 69–117.

Evans, Gareth. 1982: *The Varieties of Reference*. Edited by John McDowell. Oxford: Oxford University Press.

Lewis, David. 1970: How to define theoretical terms. *Journal of Philosophy*, 65, 113–26. Reprinted in *Philosophical papers, v.* 1, 78–95. Oxford: Oxford University Press.

Martin, R.M. 1966: On theoretical constants and Ramsey constants. *Philosophy of Science*, 31, 1–13.

Nagel, Thomas. 1974: What is it like to be a bat? *Philosophical Review*, 83, 435–50.

Pettit, Philip. Forthcoming. How the folk understand folk psychology. *Protosoziologie*.

Ramsey, F.P. 1929: Theories. In *Philosophical Papers*, D.H. Mellor (ed.), 112–36. Cambridge: Cambridge University Press.

Strawson, P.F. 1959: *Individuals*. London: Methuen.