Mental Representation, Naturalism, and Teleosemantics

Peter Godfrey-Smith

Philosophy Program, RSSS
Australian National University
and
Philosophy Department
Harvard University

To appear in Teleosemantics, edited by David Papineau and Graham Macdonald. (Final Version -- August 2004/Y)

1. Introduction

The "teleosemantic" program is part of the attempt to give a naturalistic explanation of the semantic properties of mental representations. The aim is to show how the internal states of a wholly physical agent could, as a matter of objective fact, represent the world beyond them. The most popular approach to solving this problem has been to use concepts of physical correlation with some kinship to those employed in information theory (Dretske 1981, 1988; Fodor 1987, 1990). Teleosemantics, which tries to solve the problem using a concept of biological function, arrived in the mid 1980s with ground-breaking works by Millikan (1984) and Papineau (1984, 1987).[1]

The decade or so from the early 1980s to the early 1990s was the heyday of the program of giving naturalistic theories of mental representation. The work was pervaded by a sense of optimism; here was a philosophical problem that seemed both fundamental and solvable. Its solution would be a major contribution to

cognitive science, and would also contribute to many other parts of philosophy, especially epistemology. The work was accompanied by skeptics and naysayers of various kinds (Stich 1983, Dennett 1987), but in many circles optimism prevailed. On some days it seemed that a full solution might be just around the corner.

This whole program seems to have lost momentum, at least for now. Fodor, who once had detailed solutions to offer on a regular basis, now seems to express only a vague hope that some form of informational semantics will succeed (1998, p. 12). Teleosemantics seems to have a fair number of people still working on it, with various degrees of faith, as can be seen in this volume. Millikan's enthusiasm about her initial proposals seems undiminished, in contrast to Fodor. But the teleosemantic program is not insulated from the general turn away from optimism. Sometimes an idea loses momentum in philosophy for no good reason -- because of a mixture of internal fatigue and a shift in professional fashion, for example. It is possible that this is what happened with naturalistic theories of representation. But I think that many people have been quietly wondering for a few years whether the naysayers might have been right all along.

More concretely, I think there is a growing suspicion that we have been looking for the wrong kind of theory, in some big sense. Naturalistic treatments of semantic properties have somehow lost proper contact with the phenomena, both
in philosophy of mind and in parts of philosophy of language. But this suspicion is not accompanied by any consensus on how to rectify the problem. In this paper, my response to this difficult situation is to re-examine some basic issues, put together a sketch of one possible alternative approach, and then work forward again

with the aid of this sketch.[2] So a lot of the paper is concerned with the idea of mental representation in general, and what philosophy can contribute to our understanding of this phenomenon. These foundational discussions take up the next two sections. Section 4 then looks at some empirical work that makes use of the idea of mental representation --- a different empirical literature from the ones that philosophers usually focus on. Then in Section 5 I look at teleosemantics from the perspective established in the preceding sections.

2. A model-based view of representationalism

According to the main stream of work in naturalistic philosophy of mind in the 1980s, inner states of organisms like us represent the world. "Representation" here is understood as a real, fairly unified natural relation that is picked out and understood in a very vague way by folk theory, and will eventually be described in much more detail by cognitive science and philosophy. One standard form of opposition to this picture is the "interpretivist" family of positions (Dennett 1987, Davidson 1984), according to which there are no semantic properties over and above those attributed by interpreters, where the role of interpreter is associated with a characteristic set of interests and point of view.

This mild caricature of a familiar clash provides a point from which to look for new alternatives. What we want, I suggest, is a view that says something like this: There are indeed various kinds of connectedness and specificity that link inner states with conditions in the external world. But we should not look so directly to the everyday concepts of representation, belief, meaning, and so on, in describing what these connections are. The

folk apparatus of everyday interpretation is primarily a social tool. It has genuine descriptive and explanatory uses, but these are mixed in with other features, and it is easy to be misled by socially-tuned quirks of the apparatus, when trying to use it describe real relations between inner states and the world.[3]

In some ways, this alternative shades into each of the more standard options mentioned above. But it is not supposed to be just a middle road. The idea here is that it is time to consider different possible accounts of what kinds of application semantic descriptions might have, both in principle and in practice, to inner states of physical systems. This paper explores one possibility of this type.

The main idea I will discuss is that we might see the idea of mental representation as the application of a particular model to mental phenomena. More precisely, we might see one kind of application of the idea of mental representation in these terms. The model in question is a schematized version of the pattern seen in one central kind of public representation use. That pattern is extracted and used in an attempt to understand mental processes. I see this attempted model-based understanding of the mind as available to the "folk," and available also to scientists and philosophers who treat the model in more serious and rigorous ways.

This model is one "route" to the semantic description of inner states. It is probably not the only one. A route that may be distinct from this one, at least in part, goes via a concept of computation -- via the idea of physical interactions that mirror logical relations among propositions. A third way may be via information theory in Shannon's (1948) sense. I will leave open whether or not one of the "routes" will turn out to be primary or fundamental. Certainly

there come to be connections between them (see the end of section 4 below). In addition, my aim here is not to offer a theory of how we acquire and use the most basic mentalistic concepts (thought, belief, pain, etc.). My focus is specifically on the idea of representation.

The emphasis on models in this paper is influenced by some ideas in philosophy of science, where the distinctive properties of model-based understanding have been much discussed in recent years (see especially Giere 1988). The sense of "model" I use in this paper is as follows: a model is a hypothetical structure that is supposed to bear some relevant resemblance relation to a "target" system. The hypothetical structure may in many cases be derived from another familiar, well-understood system, though that is not essential to the strategy.

I think that many philosophers, and possibly more scientists, might accept that in some sense the idea of mental representation involves the application to the mind of a model derived from public symbol use. But this fact might usually be seen as not very informative. "Yes, sure it's a model; now let me get back to what I was doing." In this paper I will keep the idea in center stage. Interestingly, Wilfred Sellars' famous 1956 discussion of the "theoretical" nature of folk psychological concepts used a very sophisticated account of theorizing that gave an important role to models, in a sense of "model" fairly close to mine. But subsequent developments of Sellars' idea have not followed suit.

So let us now look at what I will call the "basic representationalist model." This is a structure -- a sort of schema or scenario -- that furnishes a way of describing agents and their use of symbols to deal with the world. Our starting point here is one familiar everyday sense of the term "representation," as applied to

public, external objects. In this sense, a representation is one thing that is taken to stand for another, in a way relevant to the control of behavior or some other decision. More specifically, I take the paradigm case here to be that when a person decides to control their behavior towards one domain, Y, by attending to the state of something else, X. The state of X is "consulted" in working out how to behave in relation to Y. This can take the form of a conscious behavioral strategy, and it is also the topic of a familiar kind of third-person interpretation. You might decide to consult a street map to negotiate your way around a new neighborhood. Someone looking on at you can specify both the map and the mapped domain; they say you are using the map as a guide to a particular territory.

This paper will look at both this very general sense of representation and a more specific subcategory. The general class of cases is those where some X is consulted as a guide to behavior directed on Y. The more specific category is the class of cases where this strategy involves the use of a resemblance relation (perhaps an abstract and limited one) between X and Y. When we consult street maps, we usually do so because we hope to make use of a resemblance relation between map and mapped domain. But the idea of consulting the state of one thing as a guide to another does not always involve a resemblance relation. (Here I mean that we need not always hope for or rely on a representation relation, not merely that we might sometimes hope for one that is not present.) In the simplest possible case, what is consulted as a guide to behavior could be something as simple as the value of a single binary variable. (One if by land, two if by sea.)

So far I have talked about a familiar public phenomenon. But this way of thinking about representation seems to lend itself

readily to the case of mental states or brain structures. In this paper I treat this as a kind of modelling exercise; we take a familiar pattern seen in social phenomena, and apply it to the case of thought. The "we" here includes both ordinary people and cognitive scientists looking for a more scientific handle on mental processes.

The view developed here recalls, in several respects, Sellars' account of the operation of our ordinary mentalist concepts -- roughly, what philosophers now call folk psychology (1956/1997). Sellars imagined that folk psychology might first appear as a theory in which inner processes were hypothesized to resemble outward verbal discourse. The present account is similar to Sellars' in form, but is not the same view or even necessarily linked closely to it. I am supposing that public representation use furnishes a model for inner processes, but speech itself is probably not an especially relevant kind of public representation, in this context. In addition, the special features of propositional attitudes and their ascriptions are not the focus here. The best way to develop the present story in detail might be to tie it to Sellars' account, but that possibility will not be discussed much below.[4]

One problem with writing about this set of ideas is confusion resulting from the profusion of things that can be called "models." When we consult the state of X in order to determine our behavior towards Y, it can be natural to say that we are using X as a <u>model</u> of Y. This is often a useful way of talking about the phenomenon in question. But what I am concerned with in this paper is the idea of taking <u>that</u> familiar pattern or situation -- where one thing is consulted as a guide to another -- and using that as a model for understanding some features of thought. So I will, purely for practical reasons, never use the term "model" in this paper for an

internal or external object (my schematic "X") that is being treated as a representation of something else; I will only use the term "model" when talking about how the public phenomenon of representation use can be used as a source of hypotheses about inner processes.

The basic representationalist model is a very natural (in the sense of appealing) way of thinking about some aspects of the mind. I see the model as something that ordinary folk readily turn to in describing some mental processes. Above I used the case of consulting the value of a binary variable as the simplest possible example of the kind of phenomenon seen in the basic representationalist model. Once we say this, it seems obvious that the variable consulted could be either internal or external to the brain, as long as the variable's value can be read.

For those in some intellectual traditions, however, alarm bells are now ringing. The application of representational talk of this kind to internal states is a trap, raising the prospect of regresses, private-language problems, and more. The representationalist adopts an innocent look: "Surely you can't object to the internalization of the value of a binary variable? Would it help if I etched it on my teeth, rather than in my brain?" And once the basic point about the possible internalization of a simple representation has been accepted, it seems reasonable to conjecture that the complex structures of the brain could contain components that function as more elaborate maps of external things, perhaps exploiting abstract resemblance relations to coordinate behavior with the world. Indeed, some authors are led to formulate very strong hypotheses along these lines. Here is Robert Cummins: "what makes sophisticated cognition possible is the fact that the mind can operate on something that has the same

structure as the domain it is said to cognize" (1994 pp. 297-298). I will return to this claim by Cummins below. But for now let us continue discussing the basic model itself, as opposed to versions of the view that include a role for resemblance relations.

My aim is to treat the basic representationalist model in a way that avoids philosophical excesses of several kinds. It is a mistake to think that there are no prima facie foundational problems at all with the serious application of the model to inner states (some will be discussed in the next section). But it is also, I suggest, a mistake to think that semantic and representational concepts are so inextricably tied to social interpretive practices that using the model in psychology is just a massive error.

Here is another way to think of the situation. Consider the specific case of maps, which various psychologists and philosophers have thought may be akin to internal representational states. The familiar phenomena of map use in the public arena have both an "empirical skeleton" and a rich social embedding. Here I do not mean that there must be a way of picking out in bare causal terms all and only the events that would normally described as the use of maps, and assigning the maps definite semantic properties. I mean to absorb the possibility that this is impossible, because of the open-endedness and context-sensitivity of interpretive practices. I just mean that at least many cases of map use have some typical local causal features. Our habits of interpretation of these phenomena are affected by more than the empirical skeleton, but the empirical skeleton can be used as a source of hypotheses about how the mind works. I see this as applying to both the case of maps, and representations more generally.

So the empirical skeleton of public representation or map use might be made the basis for a scientific understanding of the mind -- in principle it can do this, but this may or may not be a good idea. Perhaps it is a good idea; perhaps there are special kinds of adaptive or intelligent dealing with the world that are only made possibly by representation use, where this phenomenon is found in public contexts and also in the mind. Ruth Millikan's theory, for example, treats internal and external signs as merely differently located instances of the same natural kind. Alternatively, this might all be a bad idea. One kind of anti-representationalist holds that the only empirical-skeletal features of representational phenomena that might be found in the mind's workings are trivial ones. The critic may also argue that using representationalist ideas when formulating structural hypotheses about the mind tends to lead to subtle regresses and pseudo-explanatory traps. So when we use the representationalist model about the mind, we get very little return and we face persistent dangers. It might, alternatively, be a good idea in some sub-fields and at some stages in our understanding, while being misleading elsewhere.

In some readers I imagine a feeling of impatience at this point. Do we really need yet another "back to square one" exercise? Surely it is perverse to deny, at the present time, that representationalism has been fruitful in many areas of cognitive science; the problems to work on now in this area are problems of detail. I sympathize with one form of this impatience -- a form that accepts that representationalism is something like a model, and insists that the model has done well in recent years. But I would add that it is easy to work within the representationalist model without properly resolving some acute foundational issues. (Indeed, that is one of the things models are good for.)

3. Three features and a challenge

Let us look more closely at the "basic representationalist model," and also at versions that make use of a resemblance relation. In this section I will discuss three characteristics of the model, and will also discuss in more detail the problem of regresses and pseudo-explanations.

The first feature of the model I will discuss perhaps looks harmless, but I will say quite a lot about it. When we have a situation that fits the basic representationalist model, the representation being consulted must be, in some sense, a distinct thing from whatever is consulting it. As the model has it, one thing is used to guide behavior towards another. If we are describing a particular situation as an instance of this phenomenon, and if the model is not being used in a merely instrumentalist way, then there must be a way of recognizing a separation between the representation and whatever is using it. Paradigmatically, there is also some generality or portability to the rule being used to interpret the representation, but I will not treat that as so important here. My focus is just on the issue of the separability of the representation from a reader, processor, interpreter, or consumer.

So if we are applying this representationalist model to the mind, and doing so in a "realist" way, then we must have some confidence that representations can in fact be separated from their users or readers. Much of mainstream philosophy has simply accepted this. There is a standard way of talking in philosophy of mind that treats this as no problem. We often posit representational states, or structures, while supposing that in some sense they can be identified as distinct parts of the system. The

availability of different "levels of description" is sometimes taken to allay any worries that might arise on this front. This tendency is not exceptionless, but a great deal of representationalist talk simply assumes a separation between a representation and something else that deals with it. This is common in teleosemantics and especially explicit in Millikan. Millikan's account is focused on things called "intentional icons" (which include beliefs and other mental representations) that are situated "midway" between "producer and consumer" mechanisms. One must make also make a separation assumption in order to say what Cummins said in the quote I gave in the previous section -- that intelligence requires that the mind operate on something with the same structure as the domain it is dealing with.

In large parts of cognitive science, standard ways of talking also assume separation, without worrying much about it. On the "classical" computationalist side of cognitive science, there is a good reason for this. One of the distinctive things about ordinary digital computers is the fact that there is a good separation between the data stored in memory and the processing apparatus that makes use of the data. (You can upgrade your memory and your processor separately.) One can talk about a computer in a way that violates the particular location of this distinction that is laid down by the hardware; one can talk of a virtual processor with a different structure from the one in the hardware, for example. But in the machine itself, there is a separation of the right kind from the point of view of the basic representationalist model. So if the mind is being seen as similar to an ordinary digital computer, there is no reason to worry too much about the possibility of data structures being inextricably tied to the processing. From the point of view of mainstream cognitive

science, it is presumably important and non-accidental that we have ended up building computers with this good separation.

In less orthodox parts of cognitive science, especially parts associated with connectionism, situated cognition, and dynamical systems, the question of separation is more vivid.[5] Sometimes a questioning of separation is seen as antithetical to representationalism; sometimes instead it is just described as "distributed representation." Connectionists quite often want to hang onto familiar kinds of representationalist talk. I do not deny that they can do this, but they may have to depart from the basic representationalist model, or interpret the model in a very "low-fidelity" way, to do so. Sometimes it seems that advocates of distributed representation want to talk in two ways at once, both inside and outside the structure of the basic model. The separation problem also has an interesting role in neuroscientific work. Talk of "inner maps" can be very appealing when talking about various cognitive functions in an abstract way, but it is the neuroscientist who has to deal with the possibility that no straightforward separation may appear between "map" and "reader."

I turn to a second feature of the basic model. When we engage in the familiar interpretive practice outlined earlier, saying that X's state is being used as a guide to Y, we assume an answer to a question about specificity. Why is it Y that is the "target" here? In the everyday cases, a person can say that it is Y that they are using X as a guide to. In the case of maps, for example, they can say that they are treating X as a map of Y. Mapping talk of this kind fits into a larger assumed semantic framework, in which maps, rules of interpretation, and target domains can be picked out and distinguished. Clearly a somewhat different story must be told when using the basic representationalist model to describe

internal processing. But I take it that some way of picking out the target domain must be available.

This general type of problem has been discussed extensively by Cummins (1996). He sees giving a theory of "targets" and giving a theory of what a representation says about a target as two distinct parts of a theory of mental content. As far as I can tell, Cummins and I do not have exactly the same issue in mind when we talk about the problem of targets. Targets in my sense are bigger and vaguer than they are in his; a typical target for me will not be a particular object but a whole region of the environment. And I do not hold out hope for a unified naturalistic theory of how targets are determined in all real cases. But we are thinking of similar problems, clearly.

To make the problem vivid, consider a scientific case. Suppose that there is a structure in a rat's hippocampus that is said to be a "cognitive map." (This concept will be discussed some more in the next section.) The rat is guiding its behavior, in some specific spatial task, by using this inner structure. It seems we can say that this is a case of the rat using the state of X (the inner structure) as a guide to Y. But, of course, all the rat is doing is receiving input of various kinds, and combining this with various pre-existing inner states to control behavior. It does not single out X, single out Y, and decide to use the former as a guide to the latter.

From the point of view of the scientist, there is no problem here. The rat is situated in a particular environment -- a maze, for example. If the scientist has reason to posit inner representations, he or she can say that the representations are being used to deal with this particular maze. The scientist applies what I will call a "thin behavioral" specification of the target. Whatever the

organism is consulting the representation to deal with, in a thin behavioral sense, is the target.

This is fine in practice, at least in simple cases. It is also rather philosophically unsatisfying. It is natural from the scientist's point of view to say that the rat is using X as a guide to Y, but as far as the mechanics of the situation are concerned, the "as a guide to Y" claim seems extraneous. There will also be a lot of vagueness in thin behavioral specifications of targets. We have a different and richer specification of the target when it is picked out explicitly by a separate representational act. Against this, it might be argued that worrying about a richer and sharper specification of the target is worrying about something that is not part of the "empirical skeleton" of representation use, and hence should not bother us. I will return to this issue below.

The third issue I will discuss in this section is not an essential part of the basic representationalist model, but is a feature of many applications and developments of it. It is common when talking of mental representation in ways inspired by the kinds of considerations discussed above to posit a resemblance relation, albeit an abstract one, between representation and target. In what I regard as well-developed versions of this idea, the target itself is not specified by the presence of a resemblance relation; the specification of the target is a separate matter. Rather, the idea is that given that some internal structure X is being consulted as a guide to Y, this consultation can only be expected to be successful or adaptive to the extent that there is a suitable resemblance relation between the two. So the goal, in some sense, of consulting a representation is to exploit a resemblance relation between representation and target.

At first glance, it surely seems clear that this should be regarded as an optional feature of the representationalist model. Some and only some public representations work via resemblance; why should this not be true also of internal representations? However, it is quite common in this area to use the notion of resemblance far more broadly, and see the exploitation of resemblance relations as a general or invariable feature of mental representation. Sometimes, it seems to me, these claims are made in a way that uses an extremely weak concept of resemblance or similarity. In other cases, the concept of resemblance being used is not especially diluted, and a genuinely strong claim is being expressed. The underlying line of reasoning might perhaps be something like this. In the public case, the available relations between X and Y that might be exploited are roughly the three distinguished many years ago by C. S. Peirce: resemblance, indication, and conventionally-established relations. The last of these is off the table in the case of mental representation. The second can be assimilated to the first, once resemblance or isomorphism is construed in a suitably abstract way. So the only kind of relation that really matters here is resemblance.

For this or other reasons, many discussions of mental representation extend the language of resemblance to cover a very broad class of cases. In Randy Gallistel's entry for "Mental Representation" in the Elsevier Encyclopedia of the Social and Behavioral Sciences (2001) he insists that all representations exhibit an isomorphism with the represented domain. In correspondence, Gallistel confirmed that cases usually discussed by philosophers using concepts of information or indication (thermostats, fuel gauges, etc.) are treated by him as involving abstract isomorphisms. Millikan's teleosemantic theory uses concepts of

mapping and correspondence in similarly broad ways; occasionally she explicitly says that her theory vindicates the idea that inner representations "picture" or "mirror" the world (1984, pp. 233, 314). And earlier I quoted Cummins (1994), who claimed that the exploitation of structural similarity is the key to all sophisticated cognition.

I do not want to deny that there are some very subtle but still reasonable notions of resemblance that may be used here, especially those employed in logic and mathematics. My aim is not to restrict the talk of resemblance and mapping to cases where some very obvious notion of picturing is involved. Yet I resist the idea that some suitably abstract resemblance or isomorphism relation is always involved in mental representation. When X is consulted to guide behavior towards Y, this may involve the exploitation of an antecedently specifiable resemblance relation, but it may not. It can be tempting to add here that there must be some natural relation between representation and target that makes the representation worth consulting. And from there, it can seem that resemblance or isomorphism is the only genuine candidate. But this is not so. Once we have an intelligent brain, it can generate and adaptively manipulate representations that do not have any simple, easily-exploited relation to their targets. (Strong versions of the "language of thought" hypothesis are expressions of this possibility: Fodor 1975.) For this reason, I see no reason to accept the Cummins hypotheses that was quoted earlier. That hypothesis arises out of a desire for an overly simple explanation for when and why it is worth consulting X to deal with Y.

More precisely, there are (as we often find) strong and weak ways to read the Cummins hypothesis, with the strong way

unjustified and the weak way misleading. In strong forms, the hypothesis was criticized in the previous paragraph. In weak forms, the notion of similarity or resemblance is extended too far, and becomes post-hoc in its application. (If a representation was successfully and systematically consulted to deal with some target, there must have been a similarity or isomorphism present of some kind....)

Before leaving this topic, I should note in fairness that the Cummins hypothesis I have focused on here was expressed in a note attached as commentary (1994) to a reprinting of an earlier chapter. The same ideas were followed up in his 1996 book, but I have chosen to focus on a formulation that Cummins presented in a rather "unofficial" way. Secondly, I am aware that the representational role of abstract but not-trivial resemblance relations, especially those with mathematical description, needs a far more detailed treatment than I have given it here.

The final topic I will discuss in this section is a general challenge to the usefulness of the representationalist model. I call it a challenge to the "usefulness" of the model, but the challenge is derived from stronger arguments, often directed against the model's very coherence. My aim here is to modify and moderate an older form of challenge.

I argued that the empirical skeleton of public representation use might be used as a model for some kinds of mental processing. But might it be possible to see, in advance, reasons why this will be a bad or misleading model? Famous arguments due to Wittgenstein (1953) and the tradition of work following him are relevant here. One form of argument that is especially relevant holds that if we import the basic structure of representation use into the head, we find that the reader or interpreter part of the

mechanism has to be so smart that we have an apparent regress, or pseudo-explanation.

A version of this challenge to representationalist explanation in cognitive science is expressed by Warren Goldfarb (1992). He is discussing a hypothesis that people with perfect pitch make use of "mental tuning forks." This concept was introduced in a newspaper discussion of a piece of neuroscientific work on the different neural activity of people with and without perfect pitch. Goldfarb regards the hypothesis of mental tuning forks as pseudo-explanatory in the extreme.

> Tuning forks! Are they sounding all the time? If so, what a cacophany! How does the subject know which fork's pitch to pick out of the cacophony when confronted with a tone to identify? If they are not always sounding, how does she know which one to sound when confronted with a tone?
>
> Real tuning forks give us the means to identify pitches, but they do so because we have the practices and abilities to use them. The internal standard is supposed to give us the means to identify items, but without practices and abilities, for the internal standard is also meant to operate by itself, in a self-sufficient manner. (If it were not, it would be otiose: why not settle for practices and abilities themselves?...) (Goldfarb 1992, pp. 114-115)

This line of thought might also be used to express a challenge to the Cummins hypothesis that I have discussed several times in this paper. Cummins wants to explain intelligence by giving the mind access to something with the same structure as its target. Call this structure S. If the mind's problem is dealing with things that exhibit S, how does it help to put something with S inside the head? The mind still has to detect and respond to S, just as it did when S was outside.

When the challenge is expressed in these strong sorts of terms, the right reply to it is to connect the representationalist model to some ideas used in the "homuncular functionalist" tradition (Dennett 1978, Lycan 1981). The internal representation is not supposed to be "self-sufficient," to use Goldfarb's term. It would need a reader or interpreter; there must be something akin to "practices and abilities." But the mind's interpreter mechanism need not have the whole set of practices and abilities of a human agent. The interpreter can be much less sophisticated than this (more "stupid," as the homuncular functionalist literature used to say), and might operate in a way that is only somewhat analogous to a human agent using an external representation. The representationalist holds that positing this kind of separation between a representation-like structure with an exploitable relation to a target and a subsystem to make use of that structure is a good hypothesis about the mind. If we put these two components together, some special cognitive capacities become possible.

So if the challenge is expressed by saying that we can see in advance that no explanatory progress can be made with the basic representationalist model, then the challenge can be defused. But the fact that we have this in-principle answer does not mean that we will necessarily make progress in the actual world, by using the representationalist model. It may well be that, for reasons akin to those expressed in the traditional challenge, there is little in fact to be gained by employing the model. This will depend on what the mind's structure is actually like. In order to have some explanatory usefulness, there needs to be the right kind of interaction between a representation and reader in the mind. The reader needs to be smart enough for its interaction with the

representation to be reader-like, but not so smart that the model collapses into homuncularism of the bad kind.

4. Inner maps in the cognitive sciences

This section will look at one family of applications of the basic representationalist model in psychology and other cognitive sciences. The work discussed in this section makes use of the concept of a mental or cognitive map -- a representational structure with some analogy to familiar external maps, like street maps. This is obviously not the only way to develop and apply the basic representationalist model in trying to understand mental processes, but it is a very natural way to do so. As I noted in sections 2 and 3 of this paper, there is a way of thinking about the representationalist model that leads people to think of resemblance or isomorphism as a crucial relation between internal and external states. Looking for inner map-like structures is a way to develop this idea.

The literature on inner maps is also, as I see it, a rather pure and direct way to use the basic representationalist model to think about the mind. The literature on inner maps in the cognitive sciences is partially separate from the tradition that emphasizes computation, logic, and language-like representation. The empirical work on cognitive maps in question is often (unsurprisingly) concerned with spatial skills, usually in non-linguistic animals. So this is a somewhat simpler arena in which the role of the representationalist model can be investigated. In particular, we do not have to worry about the possible effects of public language capacities on the representational powers of thought.[6]

The notion of inner maps is also interesting because it seems to be a kind of "attractor" concept, one that people come back to over and over again and from different parts of science and philosophy. There is something very appealing about this idea, but of course it also raises in a vivid way the pitfalls discussed at the end of the previous section. I should also emphasize that the discussion in this section is an initial foray into this literature; I hope to discuss it in more detail on another occasion. Here I will also discuss scientific work rather than philosophical work (see Braddon-Mitchell and Jackson 1996 for a relevant philosophical discussion).

In psychology, the father of the idea of inner maps is E. C. Tolman (1948). For Tolman, the hypothesis of "cognitive maps" was put forward in response to some particular forms of intelligent behavior, studied primarily in rats and seen especially (though not exclusively) in dealing with space. The crucial contrast that Tolman had in mind when he developed this idea was with strict "stimulus-response" models; the hypothesis of cognitive maps was motivated by the inability of stimulus-response models to account for what his rats could do. In his "sunburst maze" experiment, for example, a rat first learned a highly indirect route to a food source, and was then presented with a large range of new paths, some of which led more or less directly to the food source. Rats chose a nearly direct path much more often than chance would predict. Tolman's idea was often ignored in its mid-20th century context, but has since become much more influential. There has been a revival of the idea both in comparative psychology and also in neuroscience.[7]

Earlier I distinguished a basic sense of representation in which the state of one thing is used to guide behavior towards

another, and a richer notion in which this guidance involves a resemblance relation. Both in philosophy and in the sciences, we find the term "map" used in a range of weaker and stronger senses. In its weakest senses, any internal representation can be described as an internal or cognitive map. In its strongest senses, it involves a notion of resemblance between the map and the target domain. As far as I can tell, Tolman and many of the workers in psychology and neuroscience who have followed him use the term "cognitive map" in a way that is intermediate between the weakest and strongest senses I am distinguishing. That is, a cognitive map is not just any mental representation -- it has something extra -- but perhaps it need not work via a resemblance relation with what it represents. The term "map" is primarily invoked in connection with spatial cognition, but is sometimes used more generally.

For Tolman and others, the hypothesis of a cognitive map is often used to express the hypothesis of some kind of cognitive sophistication, over and above simple associative mechanisms and (especially) stimulus-response processes. But there seems to be a family of relevant kinds of sophistication. In a good deal of the literature that I have looked at so far, the dialectic is something like this. The researcher will have in mind a contrast between two (or more) classes of inner mechanisms that might be used in dealing with a behavioral problem (usually a problem involving space). One might be a family of comparatively simple, associative mechanisms, and the other will be a family of more sophisticated mechanisms that apparently involve a more flexible and intelligent use of information. The term "cognitive map" is associated with the latter.[8] But the boundary between what counts as a simpler or deflationary explanation and what counts as an explanation in

terms of cognitive mapping tends, perhaps, to shift around as background knowledge changes.

One of the main contrasts that is salient here is between what is described as an "integrated" representation of spatial structure, and the possession of special-purpose behavioral rules that can only applied in specific circumstances (Macintosh 2002). So a good deal of the empirical work tries to find whether animals are able to use experience to come up with solutions to problems that they have not been explicitly trained on. Many of Tolman's experiments had this character, and so have various later ones. The ability to use novel short-cuts in navigation tasks is often taken to suggest the presence of a "integrated" knowledge of spatial structure, for example. However, it turns out that there are associationist mechanisms that do predict the use of some kinds of short cuts (Deutsch 1960, Bennett 1996); the explanatory resources of the "simpler" mechanisms are richer than they were in Tolman's time.

In some of Kim Sterelny's philosophical work (2003) he explores the distinction between "decoupled" representations, that can be put to use in a variety of behavioral tasks, and those that are "coupled" to some specific task or behavior. A lot of the talk of "integrated" knowledge in discussions of cognitive mapping is gesturing towards this same distinction; integration is largely the ability to use learned information in a flexible range of tasks and contexts.

The most important landmark after Tolman was the work that introduced the idea of cognitive mapping into neuroscience, O'Keefe and Nadel's The Hippocampus as a Cognitive Map (1978). O'Keefe and Nadel argued for the existence of a special kind of learning system, the "locale" system, which constructs and uses map-like representations. O'Keefe and Nadel also offered a

specific neurological hypothesis: the hippocampus is the part of the brain where this happens, at least in some animals.[9] The neurological hypothesis was supported with various studies of deficits (especially those associated with failure at harder spatial problems) and also with the discovery of "place cells" in the rat hippocampus. These cells fire when the animal is in a certain place, but are not dependent on the simplest place-related stimuli; they keep firing when the rat's angle of view is changed or it is plunged into darkness, for example.

In much of this literature, I think it would be quite accurate to say that the idea of cognitive mapping is used as a model, in the specific sense developed in the earlier sections of this paper. A lot of discussion of the distinctive features of maps is discussion at the level of semantic properties; maps are taken to be associated with specific kinds of representational power. The psychologist's focus is often on something like the question of whether we have reason to think that the organism's brain contains something with those representational properties. Some of the questions that are most pressing from a philosophical point of view are not much discussed. In particular, there is not a great deal of discussion of how a physical structure in the brain might come to have the distinctive representational properties associated with maps.

To illustrate this, let us look at how the target and separation problems appear within this empirical work. The target problem seems to be all but invisible. Suppose the researcher is watching the rat deal with a water maze. (These are pools of colored water in which a platform is hidden just below the surface in a particular location, which the rat must find and later re-locate.) The rat solves the problem. The researcher posits a cognitive map, as the water maze is a hard problem that should defeat simple associative

mechanisms. What is the target of the hypothesized map? Obviously, it is the water maze -- that is what the rat is dealing with. Once one gets inside the mindset of the empirical worker, the idea that there is a foundational problem with the specification of the target seems strange. The focus of the empirical work is a set of contrasting hypotheses about what is going on inside the rat's head. Whatever is in there is obviously directed, in some sense, on what the researcher can see is the rat's current behavioral problem -- the water maze. The idea that mapping talk might be jeopardized by the need to give some substantive story about why that is the target is quite odd. So the cognitive scientific literature operates with what I called earlier a "thin behavioral" specification of the target. The target of a map is just whatever the map is in fact used to deal with.

With respect to the separation problem, we see a similar lack of concern to that found in much philosophical literature. The practice in the empirical literature is often to describe inner maps within the context of an imagined or assumed division between map and reader mechanisms, without treating this as a very substantive hypothesis. This might be seen as a harmless way of talking that may involve an idealization. If so, then from the philosophical point of view this idealization may be bearing quite a lot of weight.

So a lot of the time, a representationalist model using the idea of a map is introduced into discussion as a means for describing and investigating some empirically interesting distinctions, without much concern for the foundational problems. But some discussions do take some of the extra steps, and ask how a neural structure could possibly have the properties posited in the model. One way to do this is to explicitly introduce the concept of

computation. I see this concept not as itself part of the basic representationalist model, but as something additional that can be connected to it. John O'Keefe (of O'Keefe and Nadel 1978), for example, has in his later work tried to put the notion of cognitive mapping "on a firmer computational basis" (1989 p. 280). His strategy is to show how a spatially organized array of neural activity could encode a matrix of numerical values that constitute a map of the environment. This map might be "consulted" by the rat in the control of behavior via mathematical manipulation of the matrix (multiplication, inversion and so on). So a large part of O'Keefe's paper is a demonstration that these mathematical manipulations of matrices are all operations that neural circuits could, in principle, perform.

It would be interesting (and difficult) to look closely at how O'Keefe's computational account relates to the issue of separation, and also to the homuncular functionalist treatment of the role of the "reader" mechanisms, as discussed in section 3 above.

5. What have we learned from teleosemantics?

Suppose the discussion in the preceding sections is on the right track. What then is the status of the large body of philosophical work on naturalistic theories of mental representation? In particular, what, if anything, have we learned from teleosemantics? I take these two questions in turn.

The basic representationalist model is a schematic, vague sort of structure, and also one that is not usually described in rigorously naturalistic terms. So the following question presents itself: supposing that we formulated a version of the model in

purely naturalistic terms, exactly what sorts of semantic descriptions could be given a principled basis in the model?

When this question is asked about a very simple and stripped-down version of the model, we know from decades of philosophical work that the available semantic descriptions will exhibit a range of indeterminacies and breakdowns. But there is the possibility that richer versions of the model may support more determinate and fine-grained semantic descriptions than stripped-down versions do. So it is possible to take the basic model and embed it in a more elaborate and detailed scenario, where all the extra components of the scenario are described in purely naturalistic terms. We can do this, and then ask which additional kinds of semantic description attach plausibly to the resulting structure. For example, we can try to embed the basic model in a surrounding context that will make a sharp and principled notion of misrepresentation available, or the discrimination of contents that involve co-extensive concepts.

What I am describing here is a kind of philosophical model-building, which operates by augmenting and supplementing a basic representationalist model that has its origins outside philosophy. Even those philosophers who are hostile to the general framework presented in this paper will have to agree that something like this is what a lot of work in the last 20 years has, in fact, involved. The naturalistic philosophical literature has spent a lot of time describing idealized hypothetical scenarios that have reasonable or compelling descriptions in both physicalist and semantic terms. The aim has been to describe as minimal and empirically feasible a structure as possible, while maximizing the number of features of paradigmatic semantic description that attach plausibly to the structure.

In particular, this is how I see teleosemantics. The teleosemantic literature takes the basic model and embeds it within a biological setting. A simple causal structure that conforms, roughly, to the basic model, is embedded within a context involving an evolutionary history and various forces of natural selection. We take the basic model, embed it in this setting, and see how this affects our responses with respect to the semantic description of the structures in the model. So the basic representationalist model itself has no particular relationship to such things as natural selection and biological function. But it is possible to make use of such biological concepts to show how a rather elaborate semantic description of the basic model can be grounded in naturalistic factors.

One of the intuitions that has driven teleosemantics is the idea that rich biological concepts of function pick out a special kind of involvement relation between parts of organisms and their environments. Edging even closer to the semantic domain, there is a kind of specificity or directedness that an evolved structure can have towards an environmental feature that figures in its selective history. I think this idea is basically right; there is an important kind of natural involvement relation that is picked out by selection-based concepts of function. But this relation is found in many cases that do not involve representation or anything close to it. It is found in the case of enzymes, ordinary physical traits, and all the other features of organisms that can have selective histories. There is nothing intrinsically semantic about this involvement relation. But this sort of involvement relation can be added to the basic representationalist model; that is what teleosemantics does.

In keeping with the comments at the end of section 2, I should add that selection-based concepts of function might also be

used within other approaches to bridging semantic and physicalist description of inner processes, besides the basic representationalist model. But let us see how selectionist ideas can contribute to the elaboration of the model that has been the main topic of this paper.

I will focus once again on the target problem. I claimed in earlier sections of this paper that the basic representationalist model can be employed, and is often employed, with a "thin behavioral" specification of the target. In practice, there is no problem saying that the target of the rat's inner map (if there is one) is the maze with which it is dealing. This idea was accepted, while noting that from a philosophical point of view this specification of the target seems somewhat vague and extraneous. In a teleosemantic version of the representationalist model, however, the target becomes far from extraneous. This is because of the role of the target in a feedback process that shapes the representation-using mechanisms.

I will use Millikan's theory to illustrate this. A central concept in her account is that of an "indicative intentional icon." A wide range of semantically evaluable phenomena turn out to involve these structures, including bee dances, indicative natural language sentences, and human beliefs. Millikan says that an indicative intentional icon is a structure that "stands midway" between producer and consumer mechanisms that can both be characterized in terms of biological function. The consumer mechanisms modify their activities in response to the state of the icon in a way that only leads systematically to the performance of the consumers' biological functions if a particular state of the world obtains. That state is (roughly) the content of the icon. More specifically though, if we have a set-up of this kind then the icon is "supposed to map" onto the world in a particular way -- via the

application of a particular rule or (mathematical sense) function. Given the way that the consumers will respond to the state of the icon, if the world is in such-and-such a state then the icon is supposed (in a biological sense) to be in a corresponding state.

What this story involves, in abstract terms, is a combination of the basic representationalist model plus a feedback process, in which relations between actions produced and the state of the world can shape the representation-using mechanisms. We suppose that the success of actions controlled by the consultation of an inner representation is determined by the state of some particular part of the world, and these successes and failures have consequences for the modification of the cognitive system. The particular feedback process that Millikan uses is biological natural selection. Within-generation change is handled by an elaborate story about how selection on learning mechanisms can generate teleo-functional characterization of the products of those learning mechanisms. It is the abstract idea of a suitable feedback process (or what Larry Wright once called a "consequence etiology") that is most relevant here, however (see Wright 1976). If we have a feedback process of the right kind, then the representationalist model can be employed in such a way that the specification of the target becomes a natural part of the story -- a real part of the mechanism. The aspect of the world that the organism's inner representation is directed on is the aspect whose different states control feedback processes shaping relevant parts of the cognitive system, especially the ways in which representations of that kind are produced and consumed.

There is also a more famous way in which the biological embedding of the basic representationalist model can be used to motivate richer semantic descriptions. This is the problem of error

or misrepresentation. Much of the original appeal of teleosemantics was its ability to employ teleo-functional notions of purpose in order to deal with apparently normative aspects of semantic phenomena. In particular, the biological notion of failure to perform a proper function was used to attack the problem of misrepresentation, which had caused a lot of trouble for information-based theories. Millikan has occasionally said that the sole or primary contribution of the teleo-functional part of her theory (or any theory like it) is to handle this problem. I do not think this is right. As will be clear from earlier sections of this paper, I think the target problem is significant in its own right, and is not (as it is sometimes seen) merely an aspect of, or perspective on, the error problem. But the misrepresentation problem is indeed one place where the teleosemantic machinery has seemed helpful.

In the light of the preceding discussion, it is possible to take a different perspective on the use of teleo-functional concepts to deal with error. In general, the "normative" nature of semantic concepts has been overestimated and overemphasized in recent decades. Part of this emphasis came from the remarkable rhetorical power of Kripke's discussions of semantic phenomena in his book on Wittgenstein (Kripke 1982). Part of it came from other places, including the teleosemantic literature itself. Semantic phenomena do display something a bit like normativity in its more familiar senses. But the relation between the quasi-normative aspect of semantic phenomena and the quasi-normative forms of description made possible via selection-based or teleonomic concepts of function is really a kind of analogy or mirroring, not a potential reductive relationship. The ways in which misrepresentation can motivate descriptions of what "ought" to

have happened and what is "supposed" to be the case derive not from the empirical skeleton of representational phenomena, but from the elaborate network of interpretive social practices that surround it. (This is part of the real message of Kripke's book.)

Above I argued that by embedding the representationalist model in a certain kind of biological context, Millikan could give a good answer to the target question. The fact that this embedding is one way to settle questions about targets does not mean it is the only way. Other ways of augmenting or embedding the basic model may suffice to do the same kind of job. As has repeatedly been noted, the empirical commitments behind any ascription of content made on Millikan's basis are strong. A selection process of just the right kind is needed, to "aim" an inner icon at a definite aspect of the world. As I argued in an earlier paper (1994), it is possible to doubt the widespread applicability of this pattern of explanation without bringing in fanciful swampman-like cases, and without casting doubts on Millikan's interpretation of evolutionary theory and its relation to learning. All that is needed to raise problems is a "noisier" set of processes affecting the evolution of the cognitive mechanisms in question. And in cases where it is available, Millikan's solution can also diverge in interesting ways from pre-theoretic intuitions about the target of a representation. This is the message, as I see it, of Paul Pietroski's argument against teleosemantics (1992).

Pietroski uses an elaborate thought-experiment to make his case, but the argument can also be described more schematically. He describes a case where coordination with a particular environmental variable is responsible for the success of a state of internal wiring in a simple organism, but where the crucial environmental variable is not linked to anything the organism can

perceive in a single causal chain, but instead is linked to something the organism can perceive by a common cause pattern. By responding to observable variable Z, the organism coordinates its dealings with Y, where Z and Y are (roughly) products of a common cause. Applying Millikan-style principles to determine the content of an inner state by which the organism guides its behavior, we find that the inner state will represent the state of Y, the variable that matters in the explanation of success and failure. But Y may be a variable whose states no organism of this kind can perceptually discriminate in any circumstances. Pietroski makes a good case for the view that his case shows a real gap between intuition and the application of Millikan's principles. Millikan's account looks for the explanatory variable in a feedback process affecting the representational machinery, and this variable can be sufficiently far from the organism's perceptual capacities (even in optimal circumstances) that this seems an implausible target from the standpoint of everyday interpretation. In retrospect, it is unsurprising that there is no perfect match between the explanatory variable and the pre-theoretically appealing assignment of content. What I find surprising is that it takes some real work to come up with a problem case that shows this clearly.

I have discussed the relation between Millikan's work and the target problem in some detail here. I will not say much about the separation problem. It will be obvious that Millikan's account and (so far as I can tell) other accounts in the teleosemantic literature tend to help themselves to an assumption about separation without worrying much about the issue. As Millikan says, to be an intentional icon something must exist "midway between" producer and consumer mechanisms; at least to some extent, these components all have separable roles. To use one of

her standard lists of examples, bee dances, adrenalin flows, indicative natural language sentences, and human beliefs are all seen as indicative intentional icons. The assumption that beliefs are structures that have the kind of distinct location between producer and consumer mechanisms that we find in the case of bee dances and adrenalin flows is a substantial assumption.

The other topic discussed in earlier sections that might be raised here is the role of resemblance relations. I distinguished the basic representationalist model from the more specific version where a resemblance relation between representation and target is exploited. Interestingly, Millikan does present her view as one in which a mapping or resemblance relation is an essential part of the explanatory structure; she sees her account as a vindication of the old idea that thoughts can picture the world, albeit in an abstract way (1984, pp. 233, 314). Other teleosemantic theories (eg., Papineau, Neander) do not have this feature. I have argued in previous discussions (1994, 1996) that Millikan's discussions of resemblance and mapping are somewhat misleading. It is only a very attenuated sense in which her account, if true, would vindicate the idea that thoughts picture the world. This point can be made compactly by noting that the general notion of mapping that Millikan uses to describe indicative intentional icons will apply to natural language sentences, as well as beliefs. True descriptive sentences like "snow is white" map or picture the world, too. I have handled the notion of mapping in this paper in such a way that map-like representation contrasts with linguistic, conventionally mediated representation. It is possible to describe notions of reference, satisfaction and truth using the language of abstract picturing, but I do not see much point in it.

In Millikan's account, as I argued in the papers cited above, the talk of mapping boils down to a certain kind of requirement of systematicity. Any representation must be related to other possible representations in systematic ways, and the links between that set of representations and states of the world they represent will have related systematic features. Systematicity of this kind may well be a very important idea. I would distinguish it, however, from what I take to be much stronger uses of the idea of resemblance of the kind seen (for example) in the Cummins hypothesis discussed earlier in this paper.

I am acutely aware that the issues about mapping and resemblance raised here need a more detailed and careful treatment. That will have to wait for another paper.

The main thing I have done in this section is give an alternative account of the philosophical role of teleosemantics, focusing on Millikan's version. But some work in teleosemantics can also be seen as expressing empirical hypotheses. The idea that teleosemantics is a straightforwardly empirical theory has been asserted explicitly in the literature (eg., Papineau 2001). But I have in mind a slightly different set of empirical claims, and from the point of view of the present paper, the more philosophical and more empirical aspects of teleosemantics have often been mixed together in complicated ways.

One empirical hypothesis has remained live throughout the paper: does the basic representationalist model pick out an important natural kind, that figures in the causal organization of intelligent organisms? I have described representationalism as a "model," but models can be used to accurately represent the real structure of the world. Models in the present sense should not be associated, in any general way, with instrumentalism.

To that hypothesis about the basic model we can add hypotheses about the evolutionarily-embedded versions of the model that figure in teleosemantics. One hypothesis is especially vivid. Might it be the case that the only natural systems that instantiate the basic representationalist model have been shaped not just by evolution in general, but by the particular kinds of selective histories that figure in the teleosemantic literature? Might a particular kind of natural selection be the only feasible etiology for the kind of structure seen in the basic representationalist model?

Some advocates of teleosemantics might want to say, at this point, that exploring this sort of hypothesis was always the whole point of the program. To me, however, it seems that if an empirical hypothesis of this kind is supposed to be center stage, it has not been properly separated before now from various other ideas -- ideas of the kind discussed in the earlier parts of this section.

6. Conclusion.

I will give a brief summary of my main points.

We may need some new views about the kind of application that representational talk has to inner states and processes. In this paper I tried to develop one such view, by treating representationalism as the application of a certain kind of model. This view is not offered as an account of our most basic mentalistic concepts, although it might be linked with an account like Sellars'. When taken seriously in a scientific or philosophical context, the basic representationalist model does have interesting foundational problems, but these are not insoluble in principle. Much of the literature on "cognitive maps," especially in comparative

psychology, is a rather pure application of the basic representationalist model.

Teleosemantics, especially Millkan's work, can be seen as a philosophical elaboration of the basic representationalist model. Teleosemantics embeds a version of the basic model in a detailed biological scenario, and shows how quite elaborate semantic descriptions of elements of this model can be matched up with, or grounded in, carefully described biological factors. More of the baggage of folk semantic talk can be given a precise analogue in a naturalistic, biological scenario than one might expect. The biological treatment may not be the only approach to augmenting the basic model that yields these kinds of results, however. I deny that teleosemantics has uncovered a set of hidden historical assumptions that must be made when engaging in any psychological or neuroscientific talk of mental representation. In addition, teleosemantic ideas may possibly be applied to the problem of semantic properties without going via the basic representationalist model (a possibility I did not much discuss).

Teleosemantics also contains, in a rather philosophically entangled way, empirical hypotheses about the feasible etiologies by which structures that instantiate the basic representationalist model might arise.

I doubt that teleosemantics, or any theory like it, will deliver the direct, reductive, puff-of-papal-smoke solution that the 1980s literature envisaged. We probably need to look for a different kind of philosophical approach to semantic phenomena from what we are used to. But in the meantime, we have learned quite a lot from naturalistic semantics, including teleosemantics, even if we have not learned what we might have originally hoped to learn.

* * * * *

References

Bennett, A. (1996). "Do Animals Have Cognitive Maps?" Journal of Experimental Biology 199: 219-224.

Braddon-Mitchell, D. and F. Jackson. (1996). Philosophy of Mind and Cognitive Science. Oxford: Blackwell.

Clark, A. (1997). Being There: Putting Brain, Body and World Together Again. Cambridge MA: MIT Press.

Cummins, R. 1994. "Interpretational Semantics," in S. Stich and T. Warfield (eds.) Mental Representation: A Reader. Oxford: Blackwells.

Cummins, R. (1996). Representations, Targets, and Attitudes. Cambridge MA: MIT Press.

Davidson, D. (1984). Essays on Truth and Interpretation. Oxford: Oxford University Press.

Deacon, T. (1997). The Symbolic Species: The Co-Evolution of Language and the Brain. New York: Norton.

Dennett, D. C. (1978). Brainstorms. Philosophical Essays on Mind and Psychology. Cambridge MA: MIT Press.

Dennett, D. C. (1987). The Intentional Stance. Cambridge MA: MIT Press.

Deutsch, J. (1960). The Structural Basis of Behavior. Cambridge: Cambridge University Press.

Dretske, F. (1981). Knowledge and the Flow of Information. Cambridge MA: MIT Press.

Dretske, F. (1988). Explaining Behavior. Cambridge MA: MIT Press.

Fodor, J. A. (1975). The Language of Thought. New York: Crowell.

Fodor, J.A. (1987). Psychosemantics. Cambridge, MA: MIT Press.

Fodor, J. A. (1998). Concepts: Where Cognitive Science Went Wrong. Oxford: Oxford University Press.

Gallistel, R. (2001). "Mental Representation, Psychology of" in P. Baltes and N. Smelser (eds.) The International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier.

Giere, R. (1988). Explaining Science: A Cognitive Approach. Chicago: Chicago University Press.

Godfrey-Smith, P (1994). "A Continuum of Semantic Optimism," in S. Stich and T. Warfield (eds.) Mental Representation: A Reader. Oxford: Blackwell.

Godfrey-Smith, P. (1996). Complexity and the Function of Mind in Nature. Cambridge: Cambridge University Press.

Godfrey-Smith, P. (forthcoming a). "On Folk Psychology and Mental Representation" in H. Clapin, P. Staines, and P. Slezak (eds.) Representation in Mind: New Approaches to Mental Representation. Amsterdam: Elsevier Publishers, pp. 147-162.

Godfrey-Smith, P. (forthcoming b). "Untangling the Evolution of Mental Representation." To appear in A. Zilhão(ed.), Cognition, Evolution, and Rationality: A Cognitive Science for the XXIst Century. Routledge.

Goldfarb, W. (1992). "Wittgenstein on Understanding." In P. French, T. Uehling and W. Wettstein (eds.), Midwest Studies in Philosophy XVII: The Wittgenstein Legacy. Notre Dame: University of Notre Dame Press.

Jeffrey, K., M. Anderson, R. Hayman, S. Chakrabrorty (forthcoming). "Studies of the Hippocampal Cognitive Map in Rats and Humans."

Kripke, S. (1982). Wittgenstein on Rules and Private Language. Cambridge MA: Harvard University Press.

Lycan, W. (1981). "Form, Function, and Feel" Journal of Philosophy 78: 24-50

Mackintosh, N. (2002), "Do Not Ask Whether They have a Cognitive Map, but How They Find Their Way Around." Psicologica 23: 165-185.

Millikan, R. (1984). Language, Thought, and Other Biological Categories. Cambridge MA: MIT Press.

O'Keefe, J. (1997). "Computations the Hippocampus Might Perform," in L. Nadel, L. A. Cooper, P. Culicover, and R. M. Harnish (eds.) Neural Connections, Mental Computation. Cambridge MA: MIT Press, pp. 225-284.

O'Keefe, J. and L. Nadel (1978). The Hippocampus as a Cognitive Map. Oxford: Clarendon Press.

Papineau, D. (1984). "Representation and Explanation," Philosophy of Science 51: 550-72.

Papineau, D. (1987). Reality and Representation. Oxford: Blackwell.

Papineau, D. (2001). "The Status of Teleosemantics, or How to Stop Worrying About Swampman," Australasian Journal of Philosophy 79: 279-189.

Pietroski, P. (1992). "Intentionality and Teleological Error," Pacific Philosophical Quarterly 73: 267-82.

Ramsey, W., S. Stich and J. Garron (1991). "Connectionism, Eliminativism and the Future of Folk Psychology" in J. Greenwood (ed.) The Future of Folk Psychology: Intentionality and Cognitive Science. Cambridge: Cambridge University Press, 1991.

Roberts, W. A. (1998). Principles of Animal Cognition. Boston: McGraw Hill.

Sellars, W. (1956/1997). Empiricism and the Philosophy of Mind. Cambridge MA: Harvard University Press.

Shannon, C. (1948). "A Mathematical Theory of Communication," Bell System Technical Journal 27: 379-423. 623-656.

Sterelny, K. (2003). Thought in a Hostile World. Oxford: Blackwell.

Stich, S. P. (1982). "On the Ascription of Content," in A. Woodfield (ed.) Thought and Object: Essays on Intentionality. Oxford: Oxford University Press.

Tolman, E. C. (1948). "Cognitive Maps in Rats and Men." Psychological Review 55: 189-208.

Van Gelder, T. (1995). "What Might Cognition be, if not Computation?" The Journal of Philosophy 92: 345-381.

Wittgenstein, L. (1953). Philosophical Investigations. Translated by G. Anscombe. New York: Macmillan.

Wright, L. (1976). Telelogical Explanations. Berkeley: University of California Press

---

Acknowledgment: I am indebted to Kim Sterelny for extensive comments on an earlier draft of this paper. This work has been influenced by discussions with Richard Francis and Alison Gopnik.

[1] The basic ideas of teleosemantics can also be used to try to explain the semantic properties of public representations (Millikan 1984). But the main focus, both in the initial work and in more recent discussions, has been mental representation. In this paper I will discuss how mental and public representation are related, but I will not discuss teleosemantic treatments of public representation itself.

[2] I have undertaken similar forays in a couple of other papers in edited collections (Godfrey-Smith forthcoming a and b). There may be some tensions across these different papers, all of which are expressed in a cautious and exploratory way.

The forthcoming "a" paper discusses problems with the mainstream 1980s program, and possible concessions to the naysayers, in more detail.

[3] Some of Stich's papers (eg. 1982) show this rather well.

[4] My account here is closer to Sellars' account of sense-impressions, given at the very end of his paper, than it is to his basic account of thoughts.

[5] See, for example, Ramsey, Stich and Garron (1991), Van Gelder (1995), Clark (1997).

[6] For some speculations on these issues that complement the present discussion, see Godfrey-Smith (forthcoming b).

[7] For a survey in comparative psychology that uses the concept, see Roberts (1998). For a good review of the neuroscientific work, see Jeffrey et al. (forthcoming).

[8] Tolman himself, after contrasting his mapping idea with the stimulus-response model, then compared more task-specific "strip maps" with more flexible "comprehensive maps." He saw this as a gradient distinction. Once we have this distinction, the contrast between strip maps and associative mechanisms is perhaps a little problematic in Tolman's paper.

[9] In humans the hippocampus seems to be associated more with general laying down of memories than with spatial cognition in particular.