What Is a Theory of Mental Representation? [1]

STEPHEN STICH

1. Introduction

Theories of mental content or mental representation are very fashionable these days. And as with many fashionable products, the market offers a dizzying range of options.There are causal co-variation theories, teleological theories, functional role theories, and theories inspired by the causal theory of reference. There are single factor theories, multiple factor theories, narrow theories, wide theories, and a profusion of variations on all of these themes.[2] Indeed, it often seems that it is hard to find a current volume of a major journal in the area that does not have at least one article offering an argument for, or (more typically) against, someone's theory of mental representation. Moreover much of this literature has an unmistakable tone of urgency to it. The quest for an adequate theory of mental representation is not just a popular pursuit, many writers insist, it is a vitally important one. Jerry Fodor, who is rarely accused of understating the case, tells us that producing a naturalistic theory of mental content is an essential step in vindicating commonsense intentional psychology. And "if commonsense intentional psychology really were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species" (Fodor 1987, p. xii). Fred Dretske uses similarly apocalyptic terms. Without a suitably naturalistic theory of mental content, he suggests darkly, we might ultimately have to "relinquish a conception of ourselves as human agents" (1988, p. x).

While there is no shortage of debate about the merits and demerits of various accounts of mental content, there has been remarkably little discussion about what a theory of mental representation is supposed to do: What question (or questions) is a theory of mental representation supposed to answer? And what would

[1] Earlier versions of this paper have been presented at the Royal Irish Academy, at MIT, at Northwestern University, and at the Universities of Bielefeld, Colorado, Gothenberg, Konstanz, Montreal, South Carolina and Syracuse. Comments and criticism from these audiences have been helpful in more ways than I could possibly record. I am grateful to Eric Margolis for help in tracking down some of the references.

[2] For causal co-variation theories, see Dretske (1988), Fodor (1987), and Fodor (1990a). For teleological theories, see Fodor (1990b), Millikan (1984), and Papineau (1987). For functional role theories, see Block (1986), Field (1977), and Loar (1981). For a theory inspired by the causal theory of reference, see Devitt and Sterelny (1987). For a single factor theory, see Harman (1986). For a multiple factor theory, see McGinn (1982). For narrow theories, see Fodor (1987) and Devitt (1990). For a wide theory, see Burge (1979).

count as getting the answer right? These are the questions that will be center stage in the current paper. In trying to answer them, it will prove useful to start by asking another question: Why do so many people *want* a theory of mental representation; what makes the project of producing such a theory seem so urgent? This is the question I'll try to answer in §2.

Though it is unfortunate that questions about what a theory of mental representation is supposed to do have been so often neglected, it is hardly surprising. The sort of methodological self-consciousness that these questions engender has rarely been fashionable in philosophy. As a result, it is all too often the case that philosophers provide elaborate solutions for which there is no clear problem—or, as Fodor has put it, they offer cures for which there is no adequate disease. Thus I would urge, as a basic principle of philosophical method, that we spend a fair amount of time getting clear about the question, before we start worrying about the answer. When we apply this strategy to theorizing about mental representation, some very surprising conclusions begin to emerge.

Here is a preview of the conclusions that I will be defending in the pages to follow.

1. Once we start thinking about what a theory of mental representation is supposed to do, it becomes clear that there are actually several very different answers that might be offered. There is not one project here but several. These projects divide into two different families, though even within a single family, there are important differences to be noted.
2. With a single (and controversial) exception, the projects that I will sketch cannot readily be pursued by philosophers using the familiar techniques of philosophical analysis that predominate in the literature. Rather, they are intrinsically interdisciplinary projects in which the construction and testing of empirical theories plays a central role. However (again with a few exceptions) the sort of interdisciplinary work that would be necessary to make serious progress on these projects is notably absent in the literature.
3. This last fact might be taken as an indication that the projects people are actually pursuing are different from the ones that I will describe—that I have simply failed to figure out what those who are searching for a theory of mental representation are up to. But without some details on what those alternative projects might be, I am inclined to draw a darker conclusion: It is my contention that most of the players in this very crowded field have *no* coherent project that could possibly be pursued successfully with the methods they are using.
4. Even if we put these worries to one side, it is unlikely that any of the projects I will sketch will be of much help in responding to the concerns that have led many to feel it is a matter of some urgency to produce a theory of mental representation.
5. But I will also argue that those concerns themselves are deeply misguided.

So much for threats and promises. It is time to get to work.

2. *Why Would We Want a Theory of Mental Representation?*

No doubt there are lots of reasons why people might want a theory of mental representation. But among these many motives, one stands out. Concern about *eliminativism* has been a central theme in the philosophy of mind during the last decade, and producing a theory of mental representation is seen to be a central step in the debate. Though eliminativists have rarely been clear or careful in setting out their thesis, I think the doctrine is best viewed as making a pair of ontological claims, one of which is much stronger, and more unsettling, than the other. The weaker claim is that the representational states of commonsense psychology—states like beliefs and desires—will play no role in a mature theory about the causes of human behaviour. If we use the label "cognitive science" as a catchall for the various scientific disciplines that will play a role in the explanation of human behavior, then what the eliminativist is claiming is that the intentional states posited by commonsense psychology are not part of the ontology of cognitive science. The stronger claim is that these commonsense mental states do not even exist. *There are no such things*, just as there are no such things as phlogiston, or caloric fluid, or witches. Those who endorse both of these claims typically suppose that the first can be marshalled in support of the second, though it is far from clear how the argument is supposed to run.[3] There can be little doubt that many people think a theory of mental representation has a major role to play in the debate over eliminativism. But exactly what this role is supposed to be is less clear. In §3 and §4 we will be looking at a pair of views on what a theory of mental representation is. When we've made some progress on that topic, we will return, in §5, to the question of how a theory of mental representation might be exploited in arguing for or against eliminativism.

There are various arguments for the weaker of the two eliminativist theses—the claim that beliefs and desires won't play a role in a mature cognitive science. One family of arguments focuses on the *structure* of the cognitive processes and mechanisms portrayed by folk psychology. These structures, it is maintained, are incompatible with the structures posited in one or another putatively promising scientific paradigm.[4] A second family of arguments focuses on the *semantic* or *intentional* properties of mental states, as these are construed in commonsense psychology. Some of the arguments in this second family are fairly fussy and technical. They exploit sophisticated notions like supervenience, individualism, and meaning holism. But, as Fodor has noted, for many people the most worrisome fact about semantic properties is their intuitive ontological oddness.

> [T]he deepest motivation for intentional irrealism derives not from... relatively technical worries about individualism and holism... but rather from a certain ontological intuition: that there is no place for intentional

[3] For some not entirely satisfactory discussion of the point, see Stich (1983), Chapter 11, §1.

[4] See, for example, Ramsey, Stich and Garon (1990).

> categories in a physicalistic view of the world; that the intentional can't be *naturalized*. (1987, p. 97)

This worry goes a long way toward explaining a widely accepted *constraint* on any acceptable theory of mental representation. The theory must be *naturalistic*. It must show how representational properties of mental states can be explained in terms that are compatible with the broader, physicalistic view of nature provided by the natural sciences. Despite its widespread acceptance, I am inclined to think that the naturalism constraint is deeply misguided. I will say a bit about the reasons for my misgivings in §6.

Even if one accepts the naturalism constraint, however, it is obvious that this constraint can only be part of the story about what it is to get a theory of mental representation right. For on any plausible unpacking of the naturalism constraint, it will be possible to tell lots of naturalistic stories about mental representation, and these stories will differ from one another in lots of ways. We surely don't want to say that all of these accounts are correct. So let us now ask what distinguishes the good ones from the bad ones. What counts as getting a theory *right*?

3. Describing a Commonsense Concept: A First Family of Projects

A prominent feature of our everyday discourse about ourselves and about other people is our practice of identifying mental states by adverting to their content. Examples are everywhere:

> Bush believes that Gorbachev is in Moscow.
>
> I think it is going to rain this afternoon.
>
> My wife hopes that I won't be late for dinner.

In these, and in a vast range of other cases, the attribution of content is effortless, unproblematic and unquestionably useful. Moreover, in the typical case, there is widespread inter-subjective agreement about these attributions. Plainly, there must be a mental mechanism of some complexity underlying this ubiquitous practice, and it seems plausible to suppose that the mechanism in question includes a store of largely tacit knowledge about the conditions under which it is (and is not) appropriate to characterize a mental state as the belief or the desire *that p*. If we adopt the relatively loose use of the term "concept" that prevails in psychology, this amounts to the assumption that the mechanism underlying our practice embodies a concept of mental representation. And one perfectly plausible goal for a theory of mental content would be to describe that concept. To get the theory right is to give an accurate description of the concept, or the body of tacit knowledge, that underlies our quotidian practice.[5]

[5] For a rather different story about the mechanism underlying our ability to attribute mental states by characterizing their content, see Gordon (1986) and Goldman (1989). For an extended critique of the Gordon/Goldman view, see Stich and Nichols (forthcoming).

The project of describing the conceptual structure underlying judgments about content is at least roughly analogous to a variety of other projects that have been pursued in philosophy and cognitive science. In generative linguistics it is common to assume that a speaker's linguistic judgments and practice are subserved by a substantial body of tacit grammatical knowledge, and that the task of the linguist is to give an explicit account of what the speaker tacitly knows. In cognitive psychology there has been a fair amount of work aimed at making explicit the concepts and knowledge structures underlying various social and practical skills. One of the most fascinating projects along these lines has been the effort to uncover the concepts and principles of "folk physics"—the system of information about the physical world that we exploit as we wander around in it. What makes this research particularly intriguing is the finding that many people exploit a folk physics that is mistaken about the physical world, and not just in detail. The tacit theory that apparently guides these people's physical judgments and their actions is closer to medieval impetus theory than it is to Newtonian physics.[6] Findings like this may make the eliminativist's thesis a bit more plausible. If people can rely on a seriously mistaken physical theory to assist them in moving around in the world, surely it is at least possible that they rely on an equally mistaken psychological theory when they describe, explain and predict people's behavior.

A third endeavor that bears a significant resemblance to the project of describing our commonsense concept of mental representation is the sort of conceptual analysis that has provided intermittent employment for philosophers since the time of Socrates. The rules of the game have changed very little over the last 2500 years. It goes something like this:

> S: (Socrates, as it might be): Tell me please, what is *X*? (where "*X*" may be replaced by "justice" or "piety" or "knowledge" or "causation" or "freedom"...)
>
> C: (Cephalus, perhaps, or Chisholm): I will tell you gladly. To be an instance of *X*, something must be *y* and *z*.
>
> S: But that can't be right. For surely you will grant that *a* is *X*, but it is neither *y* nor *z*.
>
> C: You are quite right. Let me try again. To be an instance of *X* something must be either *y* and *z* or it must be *w*.
>
> S: I'm afraid that won't work either, since *b* is *w*, but clearly it is not *X*.

The game comes to an end when S runs out of counter-examples, or C runs out of definitions. And, though no one has kept careful records in this sport, the smart money usually bets on S.

This philosophical game of definition and counter-example makes little sense unless we make a pair of assumptions about the concepts it aims to analyze. The first of these is that the target concept can be characterized—or defined—by specifying necessary and sufficient conditions. To win a round, S can either produce an example which is an instance of the concept but is not captured by the

[6] See McCloskey, Caramazza and Green (1980), and McCloskey (1983).

definition, or he can produce an example which fits the definition but is not an instance of the concept. Moreover, it is generally assumed that the definition will be a Boolean concatenation of properties, or some relatively straightforward variation on that theme. The second assumption is that the players come equipped with enough information about the target concept to enable them to judge whether or not it applies in a wide range of cases, real and hypothetical, that they have never before imagined. To see this second point, consider a pair of well known examples:

(*i*) If someone asks you to keep his weapons, and then asks for them back after he has gone insane, does justice demand that you return them?

(*ii*) Suppose that Smith has just signed the papers to buy the Ford in the dealership showroom. Though Smith doesn't know it, the dealership does not have clear title to the car. However, moments before and far away, Granny Smith died, and title to her old Ford passes to Smith. So Smith believes he owns a Ford, and his belief is both justified and true. Does Smith *know* he owns a Ford?

It is hard to see how we could expect people to answer questions like these, or why we should take their answers seriously, unless we suppose that they already tacitly know something very much like the set of necessary and sufficient conditions that we are trying to make explicit.

There are two reasons why I have gone on at some length about the traditional philosophical approach to conceptual analysis. The first is that much of the philosophical literature on mental representation seems to fit squarely within the definition and counter-example paradigm. Philosophical theories about the nature of mental representation typically offer what purport to be necessary and sufficient conditions for claims of the form:

Mental state M has the content p.

And objections to these theories typically turn on intuitive counter-examples—cases in which the definition says that M has the content p, but intuition denies it, or vice versa.[7] The second reason is that there is now a fair amount of evidence suggesting that the assumptions underlying this traditional philosophical project may be simply mistaken. And if they are, then the project which dominates the philosophical literature on mental representation will be seriously undermined.

In the psychological literature, perhaps the most widely known challenge to the assumptions underlying traditional philosophical analysis derives from the

[7] See, for example: Block (1986), p. 660; Field (1986), p. 444; Jones, Mulaire and Stich (1991), §4.2; Loewer (1987), p. 296.

It is worth noting that on several occasions Fodor has claimed that he would be satisfied with sufficient conditions "for one bit of the world to be about (to express, represent, or be true of) another bit", even if they are not necessary (Fodor (1987), p. 98. See also Fodor (1990a), p. 52 ff.).

But, as noted in Jones, Mulaire & Stich (1991), if we read him literally, then it is hard to believe that this is what Fodor really wants. For providing conditions that are merely sufficient is just too easy.

> If x is Fodor's most recent utterance of "Maria Callas" (or: if x is the concept that underlies that utterance) then x represents Maria Callas.

work of Eleanor Rosch and her co-workers.[8] On the Roschian view, the mental structures that underlie people's judgments when they classify items into categories do not exploit tacitly known necessary and sufficient conditions for category membership, or anything roughly equivalent. Exactly what they do use is an issue that has motivated a great deal of empirical research during the last fifteen years, and continues to be actively explored. Early on Rosch proposed that categorization relies on *prototypes*, which may be thought of as idealized descriptions of the most typical or characteristic members of the category. The prototype for *bird*, for example, might include such features as flying, having feathers, singing, and a variety of others. In determining whether a particular instance falls within the category, subjects assess the *similarity* between the prototype and the instance being categorized. However, the features specified in the prototype are not even close to being necessary or sufficient conditions for membership. So, for example, an animal can lack one or many of the features of the prototypical bird, and still be classified as a bird. Emus are classified as birds though they neither fly nor sing. An alternative to the prototype theory is the hypothesis that categorization is subserved by *exemplars*, which can be thought of as detailed mental descriptions of specific members of the category that are familiar to the person doing the categorizing. On this account, too, people determine whether an item is a member of a category by making a tacit similarity judgment. However, on the exemplar theory, the item being classified is compared to exemplary members of the category.[9]

More recent research has made it clear that for many concepts neither the prototype nor the exemplar account will explain all the data comfortably. For some concepts it has been proposed that subjects' judgments rely on something very much like a tacitly known scientific theory. In other cases it has been suggested that there is no enduring concept underlying categorization judgments. Rather, it is argued, subjects construct concepts of various different sorts "on the fly", in response to the situation in which the need to categorize arises.[10]

Although there has been an enormous amount of work on concepts and categorization in recent years, there has been no systematic empirical study of *intentional* categories—categories like *believing that p*, or *desiring that q*. Thus at present we can only speculate about what such an investigation would reveal. Perhaps the safest bet is that whatever the mental mechanism underlying inten-

---

If *y* is Fodor's most recent utterance of "Meaning Holism is a crazy doctrine" (or the thought that underlies it) then *y* is about Meaning Holism, and *y* is true iff Meaning Holism is a crazy doctrine.

There are two sufficient conditions, and for a few pennies each I will be happy to provide indefinitely many more.

[8] As Rosch frequently notes, her work in this area was inspired by Wittgenstein's *Philosophical Investigations*.

[9] For an excellent review of the literature on prototype and exemplar theories, see Smith and Medin (1981).

[10] Murphy and Medin (1985), Barsalou (1987), Rips (1989).

tional categorization may be, it will not utilize "classical" concepts—the sort that can be defined with a set of necessary and sufficient conditions. The argument here is straightforwardly inductive: *No* commonsense concept that has been studied has turned out to be analyzable into a set of necessary and sufficient conditions. Indeed, given currently available evidence, it looks like there are no classical concepts. A second plausible speculation is that the concepts or "knowledge structures" underlying intentional categorization are much more complex than those traditionally offered in philosophical analyses. It's my guess that our "concept" of mental content is going to look more like a theory than like a Platonic definition.

Suppose these speculations are right, what follows? The most obvious consequence is that in seeking to build a theory of mental representation, the traditional philosophical method of proposing definitions and hunting for intuitive counterexamples will have to be abandoned. That method tries to specify a set of conditions that all and only the cases which intuitively fall under the target concept will satisfy. But if our intuitions about whether a state has the content *that p* are guided by prototypes, or exemplars, or tacit theories, or if the mental structures that determine our intuitive judgments are constructed partly in response to the circumstances in which the judgment is called for, then there will be no such conditions. So if using the method of definition and counter-example is the hallmark of a philosophical theory in this area, and if the commonsense concept of mental representation is like every other concept that has been studied empirically, there is a sense in which *there can be no philosophical theory of content.*

It is important not to read too much into this conclusion, however. For, although the traditional method of philosophical analysis may have to be abandoned, there is no reason why we cannot use other methods in constructing a descriptive theory about the ordinary concept of mental representation. Linguists, cognitive psychologists and cognitive anthropologists have developed a variety of methods for exploring the structure of commonsense concepts, none of which presuppose that these concepts have a classical structure that can be captured by a set of necessary and sufficient conditions. With a bit of ingenuity, one or more of these methods might well be used to probe the mechanisms underlying our intuitive judgments about mental representation.

We began this section by asking what a theory of mental representation was supposed to do. And we now have at least the outlines of one plausible answer. A theory of mental representation is supposed to describe the concept or knowledge structure underlying people's ordinary judgments about the content of beliefs, desires and other intentional states. However, if *this* is the sort of theory that philosophers want when they set out to build a theory of mental representation, then it is a good bet that they will have to give up "doing philosophy" (as traditionally conceived) and start doing cognitive science instead.

5. *Mental Representation as a Natural Phenomenon: A Second Family of Projects*

The description of commonsense concepts, when not encumbered by a priori philosophical requirements on what such a description must look like, is a perfectly reasonable activity. But it is not the only project that those who seek a theory of mental representation might have in mind. To see what the alternative might be, consider the concept of disease. There is a substantial anthropological literature aimed at describing the concept of disease as it is used in various cultures.[11] And if you are interested in how people conceive of disease, this is the place to look. But if you are interested in what disease is then it is biology or medicine you should be studying, not cognitive anthropology. An entirely analogous point could be made about *gold*, or *space*, or *mass*, or *heredity*. If you want to know how people conceive of them, then the description of commonsense concepts or knowledge structures is the project to pursue. But if you want to know what gold, or space, or mass, or heredity is really like, then you should be studying chemistry or physics or genetics.

Sometimes the relevant science will be pretty explicit about how it conceives of the item of interest. *The Handbook of Physics and Chemistry* will tell you all you want to know about gold, and then some. But in lots of other cases a science will use a concept quite successfully without providing a fully explicit or philosophically satisfying account of that concept. In those cases, philosophers of science often step in and try to make the notion in question more explicit. In recent years, there have been illuminating studies of *fitness*, *grammaticality*, *space-time* and a wide variety of other notions.[12] Part of this work can be viewed as straightforward conceptual description—trying to do for scientific concepts and theories what linguists, cognitive psychologists and cognitive anthropologists have tried to do for commonsense concepts and theories. Indeed, in recent years a number of philosophers of science have begun using the techniques of cognitive science in the analysis of science, often with intriguing results.[13] Sometimes, however, the concepts philosophers find, and the theories in which they play a role, are uncomfortably vague or poorly developed. And in these cases it is not at all uncommon for philosophers of science to propose improvements in the concepts and theories they are describing. It is often no easy matter to say where description stops and construction begins, and for most purposes it hardly matters.

It looks like we now have the beginnings of a second, rather different, answer to the question of what a theory of mental representation is supposed to do. On this second account, a theory of mental representation doesn't much care about

[11] See, for example, Murdock (1980).

[12] For fitness, see Sober (1984); for grammaticality, see Fodor (1981); for space-time see Sklar (1974).

[13] See, for example, Giere (1988), Glymour, Kelly, Scheines, and Sprites (1986), Langley, Simon, Bradshaw and Zytkow (1987), Nersessian (1991), Thagard (1988).

the commonsense conception of mental representation. The intuitions and tacit knowledge of the man or woman in the street are quite irrelevant. The theory seeks to say what mental representation really is, not what folk psychology takes it to be. And to do this it must describe, and perhaps patch up, the notion of mental representation as it is used by the best cognitive science we have available. So on this account, a theory of mental representation begins as part of the cognitive psychology of cognitive science, though it may end up contributing to the conceptual foundations of the science it sets out to describe.

In the large literature on mental representation, I know of only one author who explicitly undertakes the project I have been sketching. The author is Robert Cummins, and in his recent book, *Meaning and Mental Representation*,[14] he offers a detailed account of a notion of mental representation. But he goes out of his way to stress that the notion he is concerned with is not the folk psychological concept that underlies our ordinary language of intentional characterization (p. 26). Rather, his goal is to give an account of the notion of mental representation that is used in one venerable and still vigorous research tradition in cognitive science—the tradition that seeks to build what Cummins calls "orthodox" computational theories of cognition. This tradition "assumes that cognitive systems are automatic interpreted formal systems" (p. 13), and much of the work on problem solving, planning, language processing and higher level visual processing that has been done during the last two decades falls squarely within the orthodox computational paradigm.

An essential part of Cummins's project is an explication of the explanatory strategy of computational theories of cognition. He offers an account of what these theories are trying to explain, and of what successful explanations in this paradigm must do. This explanatory structure imposes strong constraints on an account of mental representation since the notion of representation used in computational theories must make sense of the explanations being offered. Here's how Cummins characterizes his approach.

> First determine what explanatory role representation plays in some particular representation-invoking scientific theory or theoretical framework; then ask what representation has to be—how it is to be explicated—if it is to play that role. (p. 145)

Though Cummins's target is the notion of representation exploited in computational theories of cognition, he recognizes that this is not the only promising research tradition in cognitive science. "There are a number of different frameworks in the running in cognitive science today" (p. 26), including "orthodox computationalism, connectionism, neuroscience" (p. 12) and a variety of others. Much the same approach could be used on the notions of representation exploited in these other traditions, though the results might well turn out quite different. "[T]o suppose that... [these other research traditions] all make use of the same notion of representation seems naive" (p. 12). If we ask what each of these frame-

[14] Cummins (1989). Page references to Cummins' book will be given in parentheses in the text.

works takes mental representation to be, "we are not likely to get a univocal answer" (p. 26).

This pluralistic picture is one I vigorously endorse. It adds an important dimension to the account of theories of mental representation that I have been sketching in this section. For if different paradigms within cognitive science use different notions of representation, then there isn't going to be *a* theory of mental representation of the sort we have been discussing. There will be *lots* of theories. Moreover, it makes no sense to ask which of these theories is the right one, since they are not in competition with one another. Each theory aims to characterize a notion of representation exploited in some branch of cognitive science. If different branches of cognitive science use different notions of representation, then there will be a variety of correct accounts of mental representation. Of course it might be thought that the various branches of cognitive science are themselves in competition, and that the correct theory of mental representation is the one that describes the notion of mental representation exploited by the correct cognitive science. But I see no reason to suppose that there is a unique correct framework for theories in cognitive science. There are lots of phenomena to explain, and lots of levels at which illuminating and scientifically respectable explanations can be given. Thus I am inclined to be a pluralist in this domain as well.

This is not the place for a detailed discussion of Cummins's account of the notion of mental representation, as it is used in computational theories of cognition. But there are a few themes in Cummins's work that I want to pursue a bit further, since they will lead us back to the question of how theories of mental representation are supposed to function in the debate over eliminativism.

5. *Theories of Mental Representation and the Eliminativism Debate*[15]

As Cummins sees it, the notion of mental representation that he is trying to describe abstracts from both the history of the system, and "the actual items in a system's current environment" (p. 81). "According to computationalism, cognitive systems are individuated by their computational properties" (p. 82). And a pair of systems can have the same computational properties, even though they differ in history, in environment and even in physical make-up. The taxonomy generated by this notion of mental representation is *individualistic*—if a pair of organisms or systems have the same physical make-up, then their representational states represent the same thing. However, following Putnam, Burge and others, Cummins also maintains that the taxonomy of intentional states exploited by folk psychology is *anti-individualistic*—"beliefs and desires cannot be specified in a way that is independent of environment" or history (p. 140). What Cummins concludes from all of this is that beliefs, desires and the rest of the

[15] Parts of this section were borrowed in Stich (1991).

intentional states of commonsense psychology are not among the items recognized by the computational theory of cognition. "What the anti-individualist arguments of Putnam and Burge prove from the point of view of the [computational theory of cognition] is that beliefs and desires aren't psychological states in the sense of 'psychological state' of interest to the CTC" (p. 140).

It looks like what we have here is the beginnings of an argument for eliminativism in which both sorts of theories of mental representation that we have been sketching play a role. In outline, the argument works like this: First describe the notions of mental representation exploited by commonsense psychology and by computational theories of cognition. Next, compare the two. If they are significantly different, then the representational states of commonsense psychology are not part of the ontology of computational theories. Of course, this sort of argument won't make the eliminativist's case if it is restricted to the computational theory of cognition, since as we've lately noted, contemporary cognitive science is a variegated discipline, and there are lots of other research traditions around. So to develop a plausible defence of eliminativism, this argument would have to be repeated for each of the viable research traditions in the cognitive science marketplace. As an alternative to this case-by-case approach, the eliminativist could try to compress the process by showing that there are some features that any scientifically respectable notion of mental representation will have to have, and then arguing that these features are not endorsed by the account of mental representation implicit in folk psychology.[16]

However, even if all this goes well for the eliminativist, it is not clear that he will have made his case. To see why, let's go back to Cummins's contrast between beliefs, as they are construed in commonsense psychology, and the representational states of the computational theory of cognition. According to Cummins, commonsense psychology views beliefs anti-individualistically—they can't be specified in a way that is independent of environment. The psychological states posited by computational theories of cognition, by contrast, are individualistically individuated—they can be specified independent of environment. From this Cummins seems to conclude that the ontology of commonsense psychology is different from the ontology of the computational theory. The two theories are talking about different things. And Cummins is not alone in reasoning in this way. I have myself offered a similar argument in a variety of previous publications, as have some other authors.[17] But despite having such distinguished advocates, it is not at all clear that the premises of the argument support its conclusion. What the premises do entail is that folk theory and computational theories make different and incompatible claims about the states they talk about. But that surely is not sufficient to show that they are talking about different things. If it were it would be all but impossible for theorists to disagree. Could it not be the case that folk

[16] This is, in effect, the strategy I tried in Stich (1978). Still another strategy would be for the eliminativist to argue that one or another of the competing research traditions in cognitive science is not a serious contender, and thus need not be considered.

[17] See Stich (1979), (1983), part II, and Stack (unpublished).

psychology and computational theories are talking about exactly the same things, and that folk psychology is just *wrong* about them?

What is really at issue here is the question of what determines the reference of the terms used in a theory. Those with eliminativist sympathies often write as though they accepted some version of the description theory of reference.[18] But this is a doubly dubious doctrine for eliminativists to adopt. One danger is that naive versions of the description theory tend to *trivialize* eliminativism. If minor disagreements between what commonsense says about mental states and what cognitive science says about them are sufficient to show that commonsense and cognitive science are positing different entities, then of course eliminativism is correct. But who cares? No one ever thought that commonsense psychology would turn out to be right about everything. Indeed, if we grant that minor theoretical differences always engender different ontologies, and if we assume that later theories are typically closer to the truth than earlier ones, we end up with a quite mad view—a sort of *pan-eliminativism*. For surely it is very likely that *every* theory we now accept will undergo some improvements during the next century. If that's enough to show that the entities posited by current theories don't exist then *nothing* we now believe in exists!

A second concern about the description theory of reference is that even much more sophisticated versions of the theory may well be wide of the mark. And if they are, then no interesting ontological conclusions can be drawn from the fact that folk psychology and the cognitive sciences disagree about mental states. One philosopher who has seen this point very clearly is William Lycan. Here's how Lycan views the matter:

> I incline away from Lewis's Carnapian and/or Rylean cluster theory of the reference of theoretical terms, and toward Putnam's causal-historical theory. As in Putnam's examples of "water", "tiger", and so on, I think the ordinary word "belief" (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that "belief" will turn out to refer to some kind of information-bearing inner state of a sentient being,... but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. Thus I think our ordinary way of picking out beliefs and desires succeeds in picking out real entities in nature, but it may not succeed in picking out the entities that common sense suggests that it does. (1988, p. 32)

As Lycan emphasizes, it is a consequence of this view that our commonsense theories may end up having been very wrong about the nature of beliefs and other representational mental states:

[18] Lycan (1988) correctly characterizes the "doxastaphobe's" argument as follows:

> Typically their arguments take the form: "Common sense characterizes beliefs [say] as having each of the following properties: F, G, H,.... But nothing that will be mentioned by any respectable future psychology will have all or even very many of those properties; therefore, beliefs will not figure in a mature psychology". (p. 4)

> I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire. (1988, pp. 31-32)

Unfortunately, when the issue at hand is eliminativism, Lycan's line has much the same defect as naive versions of the description theory—though in the opposite direction. For on Lycan's view it is hard to see how *anything* could show that the posits of folk psychology are not part of the ontology of a given branch of cognitive science. Indeed, on Lycan's view, it is far from clear why we should not say that phlogiston really does exist. It's the stuff we now call "oxygen", and earlier theorists were "just wrong about a lot of things". So it seems that if we accept either the theory of reference that Lycan favors or the naive description theory, then the eliminativist's claim will be *trivialized*. On the description theory, eliminativism is trivially true; on the causal-historical theory, eliminativism is trivially false.

Where does all this leave us? If we want to construe eliminativism as an *interesting* doctrine, rather than one which is trivially true or trivially false, then our account of reference will have to be less restrictive than the naive description theory, and more restrictive than the causal-historical theory. And plainly we *do* want to construe eliminativism as an interesting doctrine—or so I used to think. Thus, when I first set out the argument that I've just sketched, my initial reaction was to hunt around for a more promising story about reference. However, I now think that was a serious mistake.

The problem is not that alternative accounts of reference are hard to find; quite the opposite. It's relatively easy to construct accounts of reference that appear to do just what we want. They are more restrictive than the causal-historical theory and less restrictive than the description theory. But this raises questions that should by now sound very familiar: Which of these theories of reference is the right one? And what counts as getting the theory right? Moreover, in light of the close connection between the notion of reference and the notion of mental representation, it is pretty clear that much of what was said about the latter notion in §III and §IV could be repeated, with equal plausibility, about the former. If it can, then there isn't going to be any single, correct account of reference. Rather, there will be one account that describes our commonsense notion of reference (or several accounts, if there is more than one commonsense notion in circulation), and other accounts describing reference-like notions that may be of use in one or another project in psychology or linguistics or epistemology, or perhaps in some other discipline.

Now if all of this is right, some surprising conclusions follow. The first is that eliminativism cannot be viewed as a single thesis, nor even as a pair of theses as suggested in §II. To see the point, consider the weaker of the two eliminativist theses distinguished earlier—the thesis which claims that the posits of commonsense psychology are not part of the ontology of cognitive science. If our recent reflections are on the right track, then this thesis makes no sense—it

has no determinate truth conditions—unless it is tied to some specific account of reference. So if there are many perfectly correct accounts of reference, then there are many different readings of the "weak" eliminativist thesis. Moreover, it is plausible to suppose that on some of these readings the eliminativist's claim will turn out to be true, while on others it will turn out to be false. But if this is right, then it is far from clear that any reading of the eliminativist thesis is all that interesting a claim. Before we realized that any intelligible version of the doctrine had to be relativized or indexed to a theory of reference, it was perhaps plausible to claim that if eliminativism was true, then some grave intellectual catastrophe would ensue. And that, surely, is more than enough to make the doctrine interesting. But once we've seen the need to index the doctrine to some particular theory of reference, things look rather different. For surely no one is prepared to claim that eliminativism, no matter what theory of reference it is relativized to, will bring the intellectual roof down. Of course, one still might maintain that there is some particular theory of reference such that if eliminativism, indexed to that theory, is true, then worrisome consequences will ensue. And for all I know this might be right. But if it is right, it certainly isn't *obvious*; it is a claim that needs an argument. And I haven't a clue how that argument might go. So until we get some enlightenment on the matter, I think it is reasonable to suspect that the interest of eliminativism has been very much exaggerated.

6. *Eliminativism and the Naturalism Constraint*[19]

Recall that on Fodor's view the "deepest motivation" for eliminativism, or "intentional irrealism", is the suspicion that "the intentional can't be *naturalized*" (1987, p. 97). Presumably the implicit argument for irrealism that Fodor has in mind has the structure sketched in the previous section: To be exploited in a respectable scientific theory a concept must be naturalizable. So if intentional notions can't be naturalized, then they can't be exploited in any respectable scientific theory. Fodor's own theory of content is largely motivated by the hope that this suspicion can be laid to rest. In this last section I want to consider a pair of questions about all of this. First: What would it take to allay the concern that "the intentional can't be naturalized", and how might a theory of mental representation of the sort we have been considering play a role in this process? Second: Just how bad would it be if the project fails and we discover that we can't naturalize intentional notions?

The answer to the first question seems clear enough, at least in outline. To put to rest the fear that the intentional can't be naturalized, we have to give a naturalistic account of a notion of mental representation that is (or might be) exploited

[19] For a more detailed treatment of the issues discussed in this section, see Stich and Laurence (in preparation).

in cognitive science, *and with which the commonsense notion* of mental representation may plausibly be identified. However, this answer raises a pair of problems. The first is simply a version of the problem that we were wrestling with in the previous section: What does it take to justify the cross-theoretic identification of a pair of theoretical concepts? On my view, there is no determinate answer to this question. But I have already said my piece on that topic, and won't reopen the issue here. The second problem is one that has been lurking in the shadows since the early pages of this paper, when the issue of "naturalizing" mental representation was first raised: What does it take for an account of mental representation to be *naturalistic*? Though I know of no one who has offered a detailed answer to this question, the literature strongly suggests that those who want a naturalistic account of mental representation want something like a definition—a set of necessary and sufficient conditions—couched in terms that are unproblematically acceptable in the physical or biological sciences.

Whether an appropriately naturalistic account of mental representation can be given is, of course, very much an open question. My own guess, for what little it's worth, is that the project is quite hopeless. However, in contrast with Fodor and many others, I am inclined to think that very little hangs on the matter. Fodor suggests that if we can't give a naturalistic account of mental representation, then there will be no place for the notion in serious science. And if that's the case, then the eliminativists will have won a major battle. Indeed, perhaps they will have won the war. But this suggestion strikes me as quite wrongheaded. To see why, we need only consider a few examples. Let's begin with the notion of a phoneme. What is it to be a /p/ or a /b/? If you want a naturalistic answer, one which gives necessary and sufficient conditions in physical or biological terms, then I'm afraid you're going to be disappointed. For despite many years of sophisticated research, there is currently no naturalistic answer available.[20] Of course that situation might change. Phoneticians may come up with a naturalistic account of what it is for a sound sequence to be a /p/. But then again the current situation might not change. If it does not, this is surely no reason to become a phoneme-eliminativist and deny the existence of phonemes. Much the same point could be made about lots of other notions of unquestionable scientific utility. There is no naturalistic account of *grooming behavior* in primate ethology. Nor is there a naturalistic account of *attack behavior* in stickleback ethology. But surely it would be simply perverse to deny the existence of grooming behavior, simply because we can't define it in the language of physics and biology. Suitably trained observers can detect grooming behavior (or phonemes) with impressively high inter-subjective reliability. And that, I would urge, is more than enough to make those notions empirically respectable. To demand more—in particular to demand that the notions in question can be "naturalized"—seems unmotivated and silly. The situation for *mental representation* looks entirely parallel. There may, perhaps, be good reasons to

[20] For useful overviews, see Fry (1979) and Pickett (1980).

be an eliminativist. But the fact that mental representation can't be naturalized is not one of them.[21]

Department of Philosophy
and Centre for Cognitive Science
Rutgers University
New Brunswick
NJ 08903
USA

STEPHEN STICH

REFERENCES

Barsalou, L. 1987: "The Instability of Graded Structure: Implications for the Nature of Concepts", in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*, Neisser, U. ed. Cambridge: Cambridge University Press.

Block, N. 1986: "Advertisement for a Semantics for Psychology", in *Midwest Studies in Philosophy X : Studies in the Philosophy of Mind*, French, P. et. al. eds. Minneapolis: University of Minnesota Press, pp. 615-678.

Burge, T. 1979: "Individualism and the Mental", *in Midwest Studies in Philosophy, IV*, French, P. et. al. eds. Minneapolis: University of Minnesota Press, pp. 73-121.

Cummins, R. 1989: *Meaning and Mental Representation*. Cambridge, Mass.: Bradford Books / MIT Press.

Devitt, M. and Sterelny, K. 1987: *Language and Reality: An Introduction to the Philosophy of Language*. Cambridge, Mass.: Bradford Books / MIT Press.

Devitt, M. 1990: "A Narrow Representational Theory of the Mind", in *Mind and Cognition*, Lycan, W. ed. Oxford: Basil Blackwell, pp. 371-398.

Dretske, F. 1988: *Explaining Behavior*. Cambridge, Mass: Bradford Books / MIT Press.

Field, H. 1977: "Logic, Meaning and Conceptual Role". *Journal of Philosophy*, 74, pp. 379-409.

Field, H. 1986: "Critical Notice: Robert Stalnaker, *Inquiry*". *Philosophy of Science*, 53, pp. 425-428.

Fodor, J. 1981: "Some Notes on What Linguistics Is About", in *Readings in Philosophy of Psychology*, Vol. 2, Block, N. ed. Cambridge, Mass.: Harvard University Press, pp. 197-207.

Fodor, J. 1987: *Psychosemantics*. Cambridge, Mass.: Bradford Books / MIT Press.

Fodor, J. 1990a: *A Theory of Content and Other Essays*. Cambridge, Mass.: Bradford Books / MIT Press.

[21] After this paper was written I was delighted to discover that Michael Tye had independently reached very similar conclusions on the basis of very similar arguments.An extremely interesting paper in which Tye develops these themes will appear in this journal later this year.

Fodor, J. 1990b: "Psychosemantics, or: Where Do Truth Conditions Come From?" in *Mind and Cognition*, Lycan, W. ed. Oxford: Basil Blackwell, pp. 312-337.

Fry, D. 1979: *The Physics of Speech*. Cambridge: Cambridge University Press.

Giere, R. 1988: *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Glymour, C., Kelly, K., Scheines, R., and Sprites, P. 1986: *Discovering Causal Structure: Artificial Intelligence for Statistical Modelling*. New York: Academic Press.

Goldman, A. 1989: "Interpretation Psychologized". *Mind and Language*, 4, pp.161-185.

Gordon, R. 1986: "Folk Psychology as Simulation". *Mind and Language*, 1, pp. 158-171.

Harman, G. 1986: "Wide Functionalism", in *The Representation of Knowledge and Belief*, Harnish, R. and Brand, M. eds. Tucson: University of Arizona Press.

Jones, T., Mulaire, E. and Stich, S. 1991: "Staving Off Catastrophe: A Critical Notice of Jerry Fodor's Psychosemantics". *Mind and Language*, 6, pp. 58-82.

Langley, P., Simon, H., Bradshaw, G., and Zytkow, J. 1987: *Scientific Discovery: Computational Explorations of the Creative Process*, Cambridge, Mass., MIT Press.

Loar, B. 1981: *Mind and Meaning*. Cambridge: Cambridge University Press.

Loewer, B. 1987: "From Information to Intentionality". *Synthese*, 70, pp. 287-317.

Lycan, W. 1988: *Judgment and Justification*. Cambridge: Cambridge University Press.

McCloskey, M. 1983: "Naive Theories of Motion", in *Mental Models*, Gentner, D., and Stevens, A. L., eds. Hillsdale: N.J., Erlbaum.

McCloskey, M., Caramazza, A., and Green, B. 1980: "Curvilinear Motion in the Absence of External Forces: Naive Beliefs About the Motion of Objects", *Science*, 210.

McGinn, C. 1982: "The Structure of Content", in *Thought and Object*, Woodfield, A. ed., Oxford: Oxford University Press, pp. 207-258.

Millikan, R. 1984: *Language, Thought and Other Biological Categories*. Cambridge, Mass.: Bradford Books / MIT Press.

Murdock, G. 1980: *Theories of Illness*. Pittsburgh: University of Pittsburgh Press.

Murphy, G. and Medin, D. 1985: "The Role of Theories in Conceptual Coherence". *Psychological Review*, 92.

Nersessian, N. 1991: "How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science", in *Cognitive Models of Science: Minnesota Studies in the Philosophy of Science*, Vol. 15, Giere, R. ed. Minneapolis: University of Minnesota Press.

Papineau, D. 1987: *Reality and Representation*. Oxford: Basil Blackwell.

Pickett, J. 1980: *The Sounds of Speech Communication: A Primer of Acoustic Phonetics and Speech Perception*. Austin, TX: Pro-Ed.

Ramsey, W., Stich, S., and Garon, J. 1990: "Connectionism, Eliminativism and the Future of Folk Psychology". *Philosophical Perspectives*, 4, pp. 499-533.

Rips, L. 1989: "Similarity, Typicality, and Categorization", in *Similarity, Analogy and Thought*, Voisniadou, S., and Ortony, A. eds. New York: Cambridge University Press.

Sklar, L. 1974: *Space, Time and Space-time*. Berkeley, University of California Press.

Smith, E. and Medin, D. 1981: *Categories and Concepts*. Cambridge, Mass.: Harvard University Press.

Sober, E. 1984: *The Nature of Selection*. Cambridge, Mass.: Bradford Books / MIT Press.

Stack, M. unpublished: "Why I Don't Believe in Beliefs and You Shouldn't". Paper delivered at annual meeting of the Society for Philosophy & Psychology, 1980.

Stich, S. and Laurence, S. in preparation: "Intentionality and Naturalism".

Stich, S. and Nichols, S. forthcoming: "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language*.

Stich, S. 1978: "Autonomous Psychology and the Belief-Desire Thesis". *The Monist*, 61, pp. 573-91.

——1983: *From Folk Psychology to Cognitive Science*. Cambridge, Mass.: Bradford Books / MIT Press.

——1991: "Do True Believers Exist?" *Proceedings of the The Aristotelian Society*, Supplementary Volume, 65, pp. 229-244.

Thagard, P. 1988: *Computational Philosophy of Science*. Cambridge, Mass.: Bradford Books / MIT Press.