THE EMPLOYMENT OF INTRINSICALLY DEFINED REPRESENTATIONS AND FUNCTIONS

by

Joel Kenton Press

_____

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Joel Kenton Press entitled The Employment of Intrinsically Defined Representations and Functions
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: March 8, 2006
Dr. Richard Healey

_____ Date: March 8, 2006
Dr. Joseph Tolliver

_____ Date: March 8, 2006
Dr. Terrance Horgan

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: March 8, 2006
Dissertation Director: Dr. Richard Healey

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Joel Kenton Press

ACKNOWLEDGEMENTS

I wish to thank my committee chair and advisor, Dr. Richard Healey, for many hours of insightful discussion and several years of patience. I am also grateful to the other committee members, Dr. Joseph Tolliver and Dr. Terry Horgan, for their equally valuable input. Furthermore, I am thankful for frequent contributions and comments offered by the faculty and graduate students of the Philosophy Department at the University of Arizona. Finally, I wish to acknowledge the critical input from my wife, Ms. Caryn Rogers, who has heard the story told in this dissertation more times than any other human being is ever likely to.

DEDICATION

For Caryn and Oscar

TABLE OF CONTENTS

TABLE OF CONTENTS - *Continued*

ABSTRACT

Nearly all of the ways philosophers currently attempt to define the terms “representation” and “function” undermine the scientific application of those terms by rendering the scientific explanations in which they occur vacuous. Since this is unacceptable, we must develop analyses of these terms that avoid this vacuity.

Robert Cummins argues in this fashion in *Representations, Targets, and Attitudes*. He accuses “use theories” of representational content of generating vacuous explanations, claims that nearly all current theories of representational content are use theories, and offers a non-use theory of representational content which avoids explanatory vacuity. According to this theory, representations are physically instantiated structures, and represent whatever other structures are isomorphic to them, regardless of how or whether these structures are used by some cognitive system. Unfortunately, since isomorphism is a rather weak constraint, Cummins’ theory underdetermines representational content so severely that it too undermines explanatory appeals to representation. One task I undertake is to develop an alternative non-use theory which avoids this difficulty.

My second task is to adapt Cummins' argument to criticize most current analyses of "function," which undermine scientific explanation in an analogous way. Though Cummins does not explicitly argue in this manner, his own analysis of "function," by avoiding any appeal to use, avoids the explanatory vacuity to which they succumb. Consequently, I endorse Cummins' notion of function, both as it appears in cognitive science, and elsewhere.

However, although use theories fail as analyses of the terms "representation" and "function," I argue that they can still make significant contributions to the sciences employing these terms. For, while philosophers seeking to define "representation" and "function" must avoid incorporating representational and functional uses into their definitions, scientists must still find a way to determine which representations and functions are being used. Suitably re-construed use theories of representation and function may in many cases assist them in this task by providing principles for theory choice in the face of empirical underdetermination of facts about representational and functional use.

INTRODUCTION

If the theories of cognitive science are to be a part of scientific psychology, they must be able to explain the behavior of the cognitive systems to which they apply. These theories make heavy use of the concept of a representation. Thus, it is important for cognitive science that we clearly define the notion of a representation, and do so in a way that leaves the project of cognitive science intact. This dissertation aims to provide a suitable philosophical theory of representation, and to explore the ways that this term can be employed in explanations of the phenomena of cognitive science.

In the first chapter, I examine and attempt to clarify a conceptual problem, described by Robert Cummins in his book *Representations, Targets, and Attitudes*, inherent in what he calls "use theories" of representational content. Such theories define representational content in terms of representational use, but Cummins argues that this undermines attempts to use appeals to representational use in explanations of behavior. I also examine the theory of representational content with which Cummins intends to solve this problem, but find that his theory fails to

assign unique contents to representations, and hence also fails to allow for the explanation of behavior.

In the second chapter, I offer a new theory of representational content which avoids both problems, and defend it against several anticipated objections. I also show that the same explanatory problem which Cummins raises for use theories of representational content arises in theories of functions as well. As it turns out, this problem can be solved by adopting Cummins' theory of functions, and so I do. Though it is hoped that this discussion of functions will be of interest on its own merits, it will be of further importance insofar as the notion of a function plays a crucial role in the application of my theory of representational content.

The third chapter examines how the adoption of my theory of representational content would bear on the methodology of cognitive science. The main point to be learned here is that scientists attempting to employ the conceptual framework I have identified would attempt to explain behavior by postulating that cognitive systems use various representations. Since this task will be significantly underdetermined by empirical considerations, they, like all scientists, will need to appeal to some sort

of principle governing inference to the best explanation. I will argue that the criteria to which use theorists inappropriately appeal in their theories of representational content are in fact precisely the sorts of criteria to which cognitive scientists will want to appeal in their inferences to the best explanation. Consequently, use theories of representation may still be of great value to cognitive science if they are reinterpreted in this way. The third chapter will conclude with a discussion of the role of introspection in cognitive science and a discussion of scientific realism regarding assertions of representational and functional use.

CHAPTER 1

USE THEORIES, INTRINSIC THEORIES, AND THEIR RESPECITVE PROBLEMS

Use Theories of Representation and the Problem of Explanatory Vacuity

Representations, Targets, and Attitudes (henceforth *RTA*) is devoted to arguing against use theories of representational content and presenting an alternative intrinsic theory. As Cummins uses the term, a "use theory" is one which assigns meanings to representations according to their use. This idea can be traced to Wittgenstein.

> 3.326 In order to recognize the symbol in the sign we must consider the significant use...
>
> 3.328 If a sign is *not used* then it is meaningless...(If everything in the symbolism works as though a sign had meaning, then it has meaning) (Wittgenstein, 1994a, p. 10)

Though this use thesis is most commonly associated with conceptual role theories of mental content, Cummins' argues that all the major theories of mental representational content currently on offer are use theories, the only difference between them being a matter of which aspects of

a representation's use fix its content. I will discuss this claim shortly. However, since Cummins' primary argument against use theories applies equally to any theory which proposes that any aspect of a representation's use fixes its content, it will be best to keep our discussion as general as possible in the first presentation of his argument.

Cummins' argument against use theories begins with the problem of representational error, or misrepresentation. As Cummins points out, the intuitive description of cases of misrepresentation is quite straightforward. Call any cognitive system which has as its function the representing of states of affairs of type *t* a *t*-intender, and call these states of affairs of type *t* the targets of the intender. Furthermore, call whatever is specified by a representation's satisfaction conditions its content. "Then tokening a representation is error when the target of tokening it on that occasion fails to satisfy its content" (Cummins, 1996, p. 6, 13). For instance, we might say that the human visual system is a CURRENT-VISUAL-ENVIRONMENT-intender (i.e. it has as its target the current visual environment of the human of which it is part). A representation produced by this intender in some human is a

correct representation just in case it is a representation of that human's current visual environment, and it is an incorrect representation if it is a representation of anything else. Or to put it more generally, if *r* is a representational type, *c* is its content, and *t* is the state of affairs upon which it is targeted, then *r* correctly represents *t* when $c = t$, and misrepresents *t* when $c \neq t$.

Misrepresentation has always been the Achilles' heel of use theories. The problem is that the intuitive analysis above is not available to proponents of use theories. In one way or another, all use theories attempt to assign contents to representations according to the representational use to which those representations are put. While different use theories may disagree about precisely how representations are used, they obviously must all agree that representations are used to represent their targets. But if the content of a representation is simply anything it is ever used to mean (i.e. its targets), then, by definition, $c = t$. If *c* is simply defined to be *t*, there can never be any case in which $c \neq t$, and thus, there can be no misrepresentation at all.

Of course, only the simplest use theories fall prey to the problem of representational error in such spectacular fashion. All currently contending use theories have incorporated mechanisms for avoiding it. However, as Cummins points out, all use theories can be construed as responding to the problem of misrepresentation in essentially the same way, because they inevitably respond to it by singling out, in some characteristic and hopefully principled way, some subset of all uses of a representational type as "ideal" uses in which misrepresentation is said to be impossible. Since misrepresentation is purportedly impossible in such cases, assuming in those cases that $c = t$ is allegedly unproblematic. It is then claimed that these ideal cases fix the content of the representation even in "non-ideal" circumstances, allowing for misrepresentation only in those latter cases. In other words, according to these more sophisticated use theories, it will still be the case that, by definition, $c = t$ in ideal circumstances, but it will be possible that $c \neq t$ in non-ideal circumstances, since c will in such circumstances be defined to be, not the actual t,

but rather what *t* would have been had the current non-ideal circumstances, in fact, been ideal.

Unfortunately, this sort of solution leads to Cummins' problem of explanatory vacuity. The problem is that, given the way use theories assign contents to representations, explanations of the employment of those representations by various cognitive systems, and the resultant behavior of those systems, become vacuous. If we have claimed that *r* represents *c* because, under ideal conditions, *r* is used to represent *c*, we cannot then explain *r*'s usefulness to the system in the accomplishment of this task as the result of its being a representation of *c*, on pain of vacuity. Yet, it is precisely the fact that *r* is a representation of *c* that is supposed to make it useful to the system, and it is the usefulness of the representation that is supposed to explain why the system used it.

In other words, Cummins claims that it is obviously and intuitively appropriate to explain the system's use of *r* by pointing out that 1) the system needs a representation of *c*, and 2) *r* is a representation of *c*. But if we have adopted a use theory, we end up explaining the system's use of *r* by pointing out that 1) the system needs a representation of *c*, and 2*) the system uses *r*, in ideal

circumstances, as a representation of *c*. This isn't very helpful. What we need is some account of what makes *r* usable as a representation of *c*, not simply the claim that *r* is in fact used as a representation of *c*. That the system is able to use *r* to represent *c* is what we are trying to explain, not the explanation.

Furthermore, it is the usefulness of representations to a cognitive system which is supposed to explain, in part, the system's behavior. For example, suppose we want to explain how a robot has successfully avoided an obstacle. Part of the cognitive scientist's explanation will be that the robot avoided the obstacle because it employed a representation of the obstacle, *r*, in order to plot a course around it. But if the cognitive scientist then goes on to claim that *r*'s being a representation of the obstacle is to be analyzed, in part, in terms of its being the sort of thing that would allow the robot to avoid the obstacle in certain ideal circumstances, her explanation will read, in part, roughly as follows: the robot avoided the obstacle because circumstances were ideal and the robot employed a thing which allows robots to avoid obstacles in ideal circumstances. This obviously doesn't explain anything. We need to know what it is about

employing *r* which gives the robot the ability to avoid the obstacle (Cummins, 1996, p. 38-40).

Another way of posing this problem is to point out that while any ideal use theory will be unable to accommodate misrepresentation in ideal circumstances, the fact that misrepresentation is always at least conceptually possible seems to be essential to the explanatory power of appeals to representations. In one way or another, the *explanandum* of cognitive explanations of behavior is (in a sense to be clarified in a later chapter) the appropriateness of that behavior given the situation in which it occurs. In the robot example above, the reason the robot's avoidance of the obstacle seems to require explanation is that, given what the robot has been designed to do, avoiding the obstacle is an appropriate way of interacting with its environment. Claiming that the robot employed an accurate representation is explanatory because part of what it needed in order to accomplish this task was an accurate representation. Had it employed an inaccurate representation, say a representation of the obstacle's being two feet to the right of its actual location, the robot would have been less likely to behave appropriately, perhaps crashing into the obstacle. But according to ideal

use theories, if the robot is operating in ideal conditions, any representation the robot had employed would have been just as accurate (i.e. just as good a match for the robot's target), since any representation employed under ideal circumstances is defined to have a content that matches its target.

If this is right, then either the link between accuracy of representation and appropriateness of behavior will be broken, or we will have to say that the crashing robot's behavior is just as appropriate as the avoiding robot's. If the link between accuracy and appropriateness is broken, it no longer makes sense to put a premium on accurate representation, since an inaccurate representation is just as likely to result in appropriate behavior. On the other hand, if any sort of behavior can be reinterpreted as appropriate, then the fact that the robot's behavior is appropriate no longer requires an explanation at all, since the system couldn't possibly fail to behave appropriately. In either case, it is hard to see why we should even bother with a semantic interpretation of the robot's various states, since nothing will be explained thereby.

The Breadth of the Problem

Having outlined Cummins' argument, we can return to Cummins' claim that all modern theories of mental representational content are use theories, and the consequent charge that all such theories fall prey to the problem of explanatory vacuity. The first claim is well defended by Mark Perlman in "Pagan Teleology." In order to specify the use of a representation, he says, one specifies its conceptual role. So all conceptual role theories are use theories. But as we have already seen, it will not do for such theories to identify the content of a representation with its total conceptual role (i.e. with all its actual uses), since misrepresentation then becomes impossible. If every use of the representation, whether intuitively correct or incorrect, contributes to the representation's content, none of its uses can be incorrect by falling outside that content. Instead, a conceptual role theory must identify the content of a representation with only some particular portion of its total role, which is said to constitute ideal use of the representation. In cases of ideal use, misrepresentation will still be

impossible, but it will become possible when the representation is used under non-ideal circumstances.

Perlman then argues that most modern theories can be construed as conceptual role theories which differ in their accounts of which portions of the conceptual role are ideal, and therefore fix content.

> Fodor (1987, 1990a) and Dretske (1981, 1986) identify special uses by causal features (uses by detectors under optimal conditions, or that exhibit asymmetric dependence). Harman (1982) identifies special uses as those in 'normal contexts'. Papineau (1987) identifies special uses as the ones adapted to satisfy desires and Millikan (1984, 1986) identifies special uses as the ones that maximize biological fitness (Perlman, 2002, p. 268).

Once we have shown how some particular use theory can be thus construed as a conceptual role theory that identifies a representational type's content with the ideal portion of its conceptual role, and once we have determined what that theory identifies as the ideal portion of the conceptual role, we will be in a position to show how that particular theory falls prey to the problem of explanatory vacuity. All we need to do is show that explanations of behavior become vacuous in whatever ideal conditions the theory proposes.

For example, consider Ruth Garrett Millikan's adaptational role theory of representational content. Reasoning that representations are biological phenomena, Millikan suggests that we should treat representations like other "biological categories." In general, she claims, the entities of biology are defined, not in terms of their current properties, but rather in terms of those properties which historically led to their being replicated. In particular, biological entities have proper functions. The proper function of a biological system is whatever disposition that system possesses which accounts for the past reproductive success of organisms possessing systems of the same type in accordance with a Normal explanation. A Normal explanation invoking the proper function of a particular system is one which appeals to those historically typical cases of successful reproduction which depended upon the system.

Since it is entirely possible, even probable, that only a small proportion of all historical, or even all historically typical, cases of successful reproduction actually depended upon the relevant system performing its proper function, the conditions referred to in a Normal explanation need not be statistically average (i.e. Normal

conditions are not necessarily normal.) For example, it can be the proper function of the sickle cell gene to protect against malaria even if historically only a small percentage of those with the mutation were ever actually exposed to malaria. Furthermore, since selection pressures can be generated by even small variations in fitness, a disposition of some biological entity need only contribute to successful reproduction in some small proportion of Normal circumstances in order to be a proper function of the entity. Using the same example, even if those with the sickle cell gene are only marginally more likely to survive an encounter with malaria than those without the gene, protection from malaria can be its proper function (Millikan, 1996, p. 307-309).

Application of this account to mental representation yields the theory that a representation's content is to be determined by an appeal to the representation's proper function. Basically, Millikan argues that the proper function of a representation is to be accurate. The proper function of a desire (broadly construed as any propositional attitude with a world-to-mind direction of fit) is to bring about the fulfillment of the desire, and in Normal conditions this fulfillment is more likely to

come about if the desire represents the desired state of affairs accurately (Millikan, 1996, p. 314-316). Beliefs (construed similarly broadly) also have desire fulfillment as their proper functions. In this case, however, accuracy of representation is a condition which is part of the Normal explanation of that fulfillment. In Normal conditions, cognitive systems reasoning with accurate beliefs will tend to satisfy more of their desires than systems reasoning with false beliefs (Millikan, 1996, p. 316-319). As with other proper functions, the representation's accuracy function need not always, or even often, be performed, which is to say that representational accuracy could be quite rare, and yet still be the proper function of representation.

It should now be fairly clear that Millikan's theory is what Cummins calls a use theory. If the proper function of a representation is to be accurate and if the content of a representation is to be determined by what it does when it performs its proper function, then its content is to be determined by what the representation does when it is accurate. And what it does when it is accurate is accurately represent whatever it is that Normally needed to be represented in order to yield reproductive success.

Cummins would call this state of affairs that the system needs to represent the system's target. So, Millikan's theory tells us that the content of a representation is identical to its target when the intender that employs it is functioning properly in Normal circumstances. Proper intender functioning under Normal circumstances is the ideal condition of her ideal use theory. As Millikan herself puts it,

> The "content" of an intentional icon is described by telling what sort of structure or feature would have to be in the organism's environment...in order for its consumer to use it successfully in the [N]ormal way (Millikan, 1995, p. 100).

So, since Millikan's theory is a use theory, the problem of explanatory vacuity should arise. Here is Cummins' statement of the problem.

> The adaptational theory is motivated by the fact that it is plausible...to suppose that correct uses of a representation are adaptive. But what makes this plausible is the idea that the uses in question are adaptive because they are correct. That motivation is undermined on the assumption that the uses in question are correct because they are adaptive (Cummins, 1996, p. 46).

Suppose we want to explain the ability of frogs to pick flying bugs out of the air with their tongues. Suppose further that we are looking for the explanation of this ability in Normal circumstances where all relevant systems

in the frog are functioning properly. That is, we are not trying to explain any cases of successful bug-catching that might occur in cleverly designed artificial experiments, unusual environments, or with malformed or artificially altered frogs. Part of our explanation will involve the frog's ability to accurately represent the position of its small, swiftly moving targets. In order to perform this mental task, the frog will need a representation with an appropriate content, say, the content "bug." According to Millikan's theory, one of the frog's representations will have this content just in case it is the representation which frogs have historically employed in Normal circumstances to represent the similarly situated flying bugs they were trying to eat, and which allowed them to catch those bugs when the intender that produced them was functioning properly.

But then, when we offer to explain the ability of frogs to catch bugs by appealing to the fact that frogs have representations of bugs and are able to use those representations to accurately represent bug-positions, our explanation becomes vacuous. On Millikan's account, the representation employed by current frogs only has the content "bug" because of the way the employment of this

representation enhanced the fitness of the ancestors of current frogs. But these ancestral representations could only enhance the fitness of the ancestors if they also had "bug" as their content. Again, on Millikan's account, the representation can only have this content in virtue of its enhancement of the fitness of the ancestors of the ancestors of current frogs. We can keep following this regress as long as we like, but it is clear that it must terminate in the assignment of a content to the representation, and not with a further appeal to a new generation of ancestral use.

In fact, this feature of Millikan's theory undermines the very sort of adaptational explanation of the biological origin of content upon which it is predicated. Suppose that some cognitive sub-system has been replicated because it produces a certain sort of representation, but that the organisms of which it is a part initially employ that representation in an inefficient way. That is, they exploit, even when they are functioning properly in Normal circumstances, only part of the content of the representation. It seems quite reasonable to expect that at some point selection pressure might result in organisms that do exploit this previously unexploited content.

But according to Millikan's theory, this reasoning appears to be incoherent. Since the representation was not initially used, even in ideal circumstances, in the more efficient way, it does not initially have any unexploited content. The representation only gains that extra content after its new and improved cognitive use has been selected for. But, as Cummins points out, "natural selection can only select what is already there." What makes the natural selection story plausible in the first place is the claim that the unexploited content was already there to be selected. So, if Millikan is right, representations aren't even the sort of thing that could have a selection history, and if they can't have a selection history, their contents can't be determined by their selection history (Cummins, 2000, p. 123).

Furthermore, as Cummins points out, Millikan's identification of accurate use with adaptational use does not even appear to be true, much less conceptually necessary. In many cases, inaccurate use of representations will probably be adaptive. In particular, if we assume that accurate representation generally requires more cognitive resources than sloppy representation, the fact that cognitive resources are often

limited implies that natural selection will often prefer systems which strike an economical balance between accuracy and the consumption of those resources, rather than systems which maximize accuracy at all costs. For example, natural selection will probably often prefer predator-detection systems which produce lots of false positives to ones that can always distinguish dangerous predators from other non-dangerous creatures. Mice that scurry away from everything that moves may waste some energy in needless scurrying, but probably not as much energy as they would have to expend to grow, maintain, transport, and operate the more complex brain structures needed to reliably sort sparrows and prairie dogs into one category and hawks and cats into another (Cummins, 1996, p. 45).

In other cases, a combination of systems with mutually compensating errors might be more adaptive than a combination of systems which makes fewer errors. For example,

> You can design a trout that, in spite of refraction, correctly represents the positions of insects flying just above the water, but this will not be adaptive in a trout already equipped with a jumping routine that compensates for refraction (Cummins, 1996, p. 45).

Of course, one might object that cases of compensating errors could and should be better described as cases without any errors at all. Indeed, this would probably be a good *prima facie* principle for cognitive science. However, it is not hard to imagine cases where such a reinterpretation might be, and perhaps ought to be, shunned.

For example, early fish presumably had little need to see anything above the surface of the water since, in their day, nothing even remotely interesting to fish (i.e. fish food, fish mates, fish predators) existed up there. Consequently, early fish probably lacked the ability to compensate for refraction. However, at some point some trout ancestor must have hit upon the strategy of trying to eat the bugs that eventually appeared over the surfaces of ponds and streams. This would have been rough going at first, until natural selection found some way of compensating for the refraction. One way to compensate would, of course, be to improve the accuracy of the visual system, but another would be to decrease the accuracy of the jumping routine, and we can imagine empirical evidence that would at least bolster the case for the latter alternative. For example, we might find that trout had

visual systems that were extremely similar to the visual systems of their deep-sea cousins, but significantly different motor cortexes. Or, suppose we find that just before they jump, trout always align their bodies so that they point directly at the spot where an uncompensated visual system would think the bug was, but then jump in a direction several degrees away from center.

The point is not necessarily that these sorts of considerations would or should trump the *prima facie* prohibition of compensating errors, but rather that the prohibition is merely a *prima facie* one. According to Millikan's theory, however, the fact that the trout's visual system is well adapted to the task of representing airborne bug positions entails that it is representing them correctly. Thus, what should at most be a *prima facie* prohibition becomes a conceptual necessity.

Since Millikan's theory is more or less self-consciously an idealized conceptual role theory, showing that it succumbs to the problem of explanatory vacuity is a relatively straightforward exercise. However, the problem develops equally for use theories that do not initially present themselves in this way. For example, causal theories, like the one developed by Jerry Fodor in

Psychosemantics, attempt to assign contents to representations by noting which environmental conditions causally co-vary with which representations. The basic idea is that we could say that a representation *r* has content *c* if and only if it is a law of nature that every *c* causes an *r* and that every *r* is caused by a *c* (Fodor, p. 100, 1987). So far, there is no explicit appeal either to ideal conditions or to use. Whatever causal relations do or do not hold between a brain state and various aspects of the environment presumably hold whether or not those states are used, or correctly used, as representations. However, this basic idea runs into several immediate complications.

The first pair of complications is that causal covariance is both too restrictive and too permissive. The excessive permissivity of causal covariance arises from a need to account for misrepresentation. As Cummins would put it, sometimes a cognitive system will incorrectly use a representation *r* with content *c* to represent a target *t* such that $c \neq t$. In Fodor's hackneyed example, I am trying to represent the cow over in the next pasture, but I employ a representation of a horse. The problem is that if I am going to employ any representation at all in response to

the presence of the cow, this will presumably be because the presence of the cow causes the representation. But if the purported representation of a horse (henceforward, a |horse|) can be caused by a cow, then it isn't really a |horse| at all: it's a |horse or cow|. If nothing that triggers a representation can fall outside its extension, misrepresentation will never happen. Because of the disjunctive nature of the problematic representations, Fodor calls this the "disjunction problem" (Fodor, 1987, p. 101-102).

On the other hand, causal co-variance is too restrictive simply because not every element of a representation's extension is appropriately situated to actually cause a token of that representation. Chinese horses belong in the extension of my |horse| even though I, having never been to China, have never been directly causally affected by a Chinese horse (Fodor, 1987, p. 111).

The third complication involves a different sort of excessive permissivity. The causal relations that causal covariance accounts call to mind are the causal relations involved in the operation of the senses. |Horse|s refer to horses because the presence of a horse in my visual field will reflect photons into my eyes and onto my retinae,

which will eventually result in the production of a |horse|. But, of course, outside of detection, |horse|s can be caused by all sorts of processes and events. To switch to an example of Cummins', Fido's |food|s mean food because food causes Fido's |food|s. But if Fido belongs to a certain Dr. Ivan Pavlov, Fido's |food|s can also be caused by the ringing of a bell. Or, to put it more precisely, Fido's conditioning by Pavlov has given Fido a |bell ⊃ food|, and the combination of a |bell| and a |bell ⊃ food| causes a |food|. For that matter, even without conditioning, if Fido is hungry enough to hallucinate, Fido's |food| could be caused simply by Fido's desire for food. But neither |bell & (bell ⊃ food)|s nor desires for food belong in the extension of Fido's |food|. If the causal theorist wants to eliminate these cases, he will need to limit reference-fixing causes to those features of the environment that cause tokens of the representation during sensory detection. As Cummins points out, "only in detection is a representation caused by its target" (Cummins, 1996, p. 58).

But detection is the function of a detector, which is to say, an intender targeted on some aspect of the sensory

environment. If we are talking about the functions of intenders, then we are talking about how the cognitive system of which that intender is a part uses the representations employed by the intender. So, causal theories are use theories, and hence, conceptual role theories. Furthermore, since not all uses of a representation are detection uses, we have at least the beginnings of a specification of idealized conditions for an idealized conceptual role theory. Perhaps causal theories are not so different after all.

In any event, Fodor proposes to avoid the second of these complications by an appeal to psychophysical laws. While it is certainly not the case that all horses cause |horse|s it is at least reasonably close to true that all suitably situated horses do.

> Of course there are instantiations of *horse* (horses in Peking and so forth) that don't affect the contents of one's belief box; arguably, however, that's only because one doesn't occupy a psychophysically optimal viewpoint with respect to those instantiations. For, plant a horse right there in the foreground, turn the lights up, point the observer horsewards...and surely the thought 'horse there' will indeed occur to him (Fodor, 1987, p. 115).

The fix is actually a bit more complicated than this. For one thing, it doesn't work for unobservable entities, since

there is no viewpoint, much less a psychophysically optimal viewpoint, from which to view them. Furthermore, it doesn't even work for horses, since although psychophysics may be able to guarantee that an optimally viewed horse will be seen, it can't guarantee that even an optimally viewed horse will be seen as a horse, since not every cognitive system has or is good at applying the concept of a horse. However, Fodor claims that it will work for certain very simple representations of sensory primitives, like |red| or perhaps |rough|, and that these cases form a foundation for assigning contents to more complex representations.

Whether or not this foundational project can be carried out in the way that Fodor suggests is debatable. However, that debate isn't particularly germane to the issues at hand, so let's assume that it is at least in principle possible to specify psychophysically optimal conditions under which the presence of any representable entity causally co-varies with some representation. If so, then Fodor can avoid the second complication by modifying the basic causal covariance account so that it asserts that *r* has content *c* if and only if it is a law of nature that, in psychophysically determined optimal detection

conditions, every *c* causes an *r* and every *r* is caused by a *c*. Since Chinese horses are horses, and since all horses in psychophysically optimal conditions cause |horses|, then Chinese horses fall into the extension of |horse|. Furthermore, a causal theorist could use the same appeal to psychophysics to avoid the first complication, the disjunction problem, as well. Since the cow in the next pasture wouldn't cause a |horse| if it were right next to me in good lighting conditions, etc., it does not belong in the extension of |horse|.

Unfortunately, since this is just an appeal to idealized use, it is bound to succumb to the problem of explanatory vacuity. What psychophysics can give us is a specification of circumstances in which some representation *r* causally co-varies with some environmental condition *c*. But this simply does not guarantee that *r* represents *c*, because there is no guarantee that, even in detection, the function of a representation is to be accurate.

The mouse example that we applied to Millikan's view above applies equally here. Suppose that the psychophysicists tell us that a certain representation is triggered in mice whenever birds of any kind are placed in whatever conditions turn out to be optimal for mouse

vision. Does it follow that the representation is a |bird| rather than, say, a |raptor| or |hawk|? Not necessarily. Given the tremendous importance to mice of avoiding raptors, and the relative inconsequentiality to mice of other birds, it is perfectly reasonable to suppose that a well-designed mouse might incorrectly represent any bird as a raptor even in optimal detection conditions. After all, unless our mouse has been endowed with particularly good eyesight, the mouse had better be designed to avoid any situation in which birds fall within its optimal detection range at all. If you rub a mouse's nose in hawk, what you generally end up with is a snack for the hawk. On the other hand, if you rub a mouse's nose in sparrow, nothing of consequence follows for the mouse. In neither case does successful detection under optimal conditions have much impact on the fitness of the mouse. Whether or not detection is successful, if the nearby bird is a hawk, the mouse is probably going to die, and if the bird is a sparrow, the mouse is going to go on about its business.

So, since situations in which birds do fall into the mouse's optimal detection conditions rarely have any impact on fitness, incidents of successful or failed detection in such circumstances won't exert much selection pressure, and

hence there is no reason to expect them to be particularly accurate. The situations that will exert selection pressure will be those where the bird is detected at a sufficient distance such that, if it is a raptor, the mouse will have a chance to do something about it. Consequently, it appears that causal co-variance in optimal conditions is more or less irrelevant to determining the content of the representation.

This case undermines the appeal to optimal covariance as a solution to the disjunction problem, but optimal covariance won't even work in the way that Fodor wants it to, as a way of dealing with the fact that inappropriately situated elements in a representation's extension won't cause tokens of the representation. Suppose we find that, even in optimal circumstances, a certain organism produces a certain representation only in some small percentage of cases where one of its predators is present. It doesn't follow from this that the extension of this representation must be some proper subset of these predators, rather than the entire set, because detection needn't always be successful to be selected for. Admittedly, it is a bit implausible to imagine a mouse whose nose has been rubbed in hawk failing to detect the hawk. Try instead a

paramecium whose nose has been rubbed in amoeba. Even if paramecia have amoeba-detectors that only work sporadically at best, the detectors may confer a significant fitness advantage.

Both of these examples were generated by an appeal to natural selection, but nothing essential hangs on this. By appealing to what we know about natural selection, we can make a plausible case for claiming that certain instances of detection in optimal circumstances are cases of misrepresentation. We could just as well tell stories about robots, and appeal to design considerations in order to make the same point. If I were designing a mouse-sized robot scout, and wanted it to avoid being damaged by hawks that might try to attack and eat it, but didn't want to expend the resources necessary to build a foolproof hawk-detector, I might opt for a hawk-detector that produces lots of false positives, just like natural selection.

The reason these cases provide an objection to causal theories is not necessarily that they correctly assign content in a way contrary to that theory (indeed, I'm not even claiming that these assignments are correct), but rather that it seems plausible that they assign contents correctly, and yet, according to Fodor's theory, it is

conceptually impossible that they be correct. According to Fodor, causal covariance under psychophysically determined optimal conditions defines representational content. If that were true, then these cases would be incoherent, rather than plausible.

Although a causal theorist might appeal to psychophysically optimal detection as a way of avoiding the disjunction problem, Fodor does not in fact take this route. Instead, he proposes to solve the disjunction problem by an appeal to asymmetric dependence. Representation *r*'s being caused by environmental condition c_1 is asymmetrically dependent upon *r*'s being caused by environmental condition c_2 if *r* would have been caused by c_2 even if it were not caused by c_1, but *r* would not have been caused by c_1 if it were not also caused by c_2. In such a case, Fodor claims that *r* represents c_2, rather than the disjoint content c_2 or c_1, even though both sometimes cause *r*. For example, even though cows sometimes cause |horse|s, |horse|s still mean horse, rather than horse or cow, because cows would not cause |horse|s if horses didn't, but horses would cause |horse|s even if cows didn't.

The intuition behind this is that

> Mistakes have to be *accidents*: if cows aren't in the extension of horse, then cows being called horses can't be *required* for 'horse' to mean what it does. By contrast, however, if horse didn't mean what it does, being mistaken for a horse wouldn't ever get a cow called horse. Put the two together and we have it that the possibility of saying 'that's a horse' falsely presupposes the existence of a *semantic setup* for saying it truly, but not vice versa (Fodor, 1987, p. 108).

That is, in order to misapply a representation with a particular extension to something that falls outside that extension, something has to fall within the extension, else there would be no representation with that extension to misapply. But successful applications of the representation do not seem to depend in any way on there being causes of the representation that fall outside its extension. If it weren't for horses (the things that |horse| represents), there wouldn't be any |horse|s, and hence cows couldn't cause |horse|s. But there don't have to be any cows for there to be |horse|s.

The problem with this is that it makes the causal relations which are supposed to define representational content dependent upon representational content. Granted, cows could not cause |horse|s if there were no |horse|s. But it does not additionally follow from this that cows

could not cause |horse|s if there were no horses, unless |horse|s depend ontologically on horses, but not on cows. Whether or not |horse|s do depend ontologically on horses depends on what sort of thing |horse|s are. Well, |horses| are representations of horses. So, in order to determine whether or not cow-caused |horse|s are asymmetrically dependent on horse-caused |horse|s, we need a theory of representational content. But the whole point of introducing asymmetric dependence in the first place was to patch up Fodor's causal theory of representational content in a way that allowed for misrepresentation. So, if we have Fodor's theory of representational content, we can evaluate asymmetric dependence, and if we can evaluate asymmetric dependence, we can complete Fodor's theory of representational content. Unfortunately, a theory of representational content that requires an antecedent understanding of representational content is not, in fact, a theory of representational content.

If asymmetric dependence is going to do what Fodor wants it to do, we are going to have to find some way of evaluating the relevant counterfactuals without appealing to an antecedent theory of representational content. Fodor's assumption that we already have such a theory is

implicit in his labeling of the representation in the horse/cow case. By calling the representation a |horse|, he has assumed that the representation means horse. It would be better to describe the case as one in which cognitive scientists have determined that both cows and horses sometimes cause a particular brain state they call *r*. A theory of representational content ought to be able to tell them whether *r* is a |horse|, or a |horse or cow|, or, for that matter, a |cow|, a representation of something else, or not a representation at all.

The asymmetric dependence criterion counsels the scientists to include in *r*'s extension only those of its causes that would continue to be causes if the rest of its causes failed to be causes. As with any counterfactual, the scientists' only guide to evaluation must come from their experience of the actual world. If they are to discover that *r* will behave differently in the presence of cows in horseless worlds, their discovery will have to depend on differences in the actual causal relations between cows and *r* and horses and *r*. Perhaps in the actual world cows cause *r*'s only when they are off in the distance, or under poor lighting conditions, or when it is stormy out, whereas horses cause *r*'s even when they are nearby and fully

illuminated. But this is just a retreat to the detection in optimal circumstances criterion that we have already rejected.

A further argument for the explanatory vacuity of any causally defined notion of representation is that causal theories force representations to be arbitrary symbols, which prevents any of their intrinsic features from playing an explanatory role. As Cummins argues

> If it is the causal connection to a property that gives a representation its content, then it cannot be any intrinsic properties of the representation that matter to its content. If you can build a horse detector at all, you can build one that responds to horses by tokening any symbol you like, and that symbol will be a |horse| according to [causal theories] (Cummins, 1996, p. 69).

But if mental representations are functionally completely interchangeable, it is hard to see why cognitive scientists should care which representations are being employed by the cognitive systems they are studying. Learning that the system employed representation r_1 rather than r_2 doesn't help to explain why the system behaved as it did if r_2 employed in the same circumstances would have had the very same content as r_1, or if r_1 could just as easily have had some other content.

So neither Millikan's nor Fodor's use theory can avoid the problem of explanatory vacuity. There are lots of other use theories on offer, but they all seem bound to founder in more or less the same sort of way. Consequently, Cummins argues that, if we want to have an explanatory role for representation, we will need to adopt an intrinsic theory of representational content. Such a theory will assign contents to representations according to the intrinsic properties of those representations, rather than in terms of the use of those representations by one or another cognitive system. If certain things intrinsically represent others, then, and only then, will it be possible to explain a cognitive system's behavior in terms of its possession of those representations.

The General Problem of Explanatory Vacuity

Shortly, I will present Cummins' intrinsic theory of representational content. However, before we leave the problem of explanatory vacuity, I think it may be a worthwhile diversion to briefly consider two cases outside of cognitive science where the same problem arises. Hopefully, this exercise will be interesting on its own

merits. However, I include it in the present context primarily in the hope that seeing how the problem of explanatory vacuity crops up in other disciplines may illuminate the difficulty it poses for cognitive science. Given this largely illustrative purpose, I don't intend to produce more than a superficial discussion of these issues.

In evolutionary biology, the problem of explanatory vacuity appears as the so-called tautology problem. According to the theory of natural selection, the persistence of certain biological traits is to be explained as the result of the advantages for survival and reproduction those traits confer upon the organisms which bear them. This advantageousness of an organism's traits is its fitness. Roughly, Darwin argued that if a fitness-conferring trait is heritable, and if the individuals in a population of organisms vary in their possession of or lack of this fitness-conferring trait, and if the possession of the trait is more advantageous than the lack of it, and especially if only some small portion of the organisms in any given generation survive, then the trait will tend to become more common in the population over succeeding generations. Since all of the antecedents of this conditional are frequently satisfied by most populations of

biological organisms, we can explain why those organisms tend to be well adapted for their environments. This explanation is sometimes summarized with the slogan "the survival of the fittest," or, if we'd like a proper sentence, "the fittest survive."

Certain objectors, creation scientists most tenaciously, have claimed that given the way that "fitness" is defined by biologists, this slogan becomes a tautology, and hence fails to provide any sort of explanation of biological adaptation. These objectors claim that the definition of fitness offered by evolutionary biologists is simply that an organism is fit if it actually survives. If fitness were defined in this way, then the explanatory claim made above, that organisms successfully survive and reproduce because they tend to possess fitness-conferring traits, would become the vacuous claim that organisms successfully survive and reproduce because they possess traits that allow them to survive and reproduce. In slogan form, "the fittest survive" becomes "those that survive, survive."

As I understand it, this problem is precisely analogous to the problem that worries Cummins. Representations are supposed to underwrite explanations of

cognitive behavior. When a cognitive system does A, we want to be able to say that the system was able to do A, in part, because it employed a certain representation. But use theories assign representational contents according to the uses the system has for the representation, and one of these uses is that the representation allows the system to do A. So consequently, we end up trying to explain how the system was able to do A by claiming that it employed something that would enable it to do A.

Cummins' solution to the representational problem of explanatory vacuity is to develop a way of assigning contents to representations intrinsically, without any reference to their use. This solution is analogous to Stephen Jay Gould's solution to the tautology problem.

> In nature, A's "superiority" over B will be *expressed* as differential survival, but it is not defined by it – or, at least, it had better not be so defined, lest [the creation scientists] triumph and Darwin surrender...[Rather] certain morphological, physiological, and behavioral traits should be superior a priori as designs for living in new environments. These traits confer fitness by an engineer's criterion of good design, not by the empirical fact of their survival and spread (Gould, 1983, p. 143).

In other words, fitness is to be defined in terms of the intrinsic propensity an organism has to survive and reproduce in its environment in virtue of its possession of

a certain set of traits. The fitness of New World monkeys, for example, consists, in part, in their possession of prehensile tails in an arboreal environment. *Ceteris paribus*, having a prehensile tail is likely to enhance such a monkey's chances to survive and reproduce. But it certainly doesn't entail reproductive success, and the sentence "arboreal creatures with prehensile tails survive" is no tautology. Rather, it is a deduction from empirical generalizations about the deleterious consequences to monkeys of falling out of trees and about what it takes for monkeys to avoid falling out of them. In fact, it is entirely conceivable that, despite the fitness advantage of its members, a population of monkeys with prehensile tails might be out-competed, even to extinction, by a population of monkeys with less-dexterous tails (say, because the prehensile-tailed monkeys all happen to be unlucky enough to fall out of trees that lack conveniently placed branches for their tails to grasp on the way down.)

A second example of the problem of explanatory vacuity appears in Socrates' argument against what is these days called the divine command theory of ethics. In religious ethics, of course, one is searching for a criterion of goodness, rather than an explanation *per se*. However, the

problem that is often raised for the divine command theory is that it leaves certain of God's attitudes without an explanation, and that this makes morality arbitrary. According to the divine command theory, goodness is the product of God's approval. That is, whatever actions or things God approves of are good or permitted or perhaps obligatory, and whatever actions or things God disapproves of are evil or forbidden. So the theory, in effect, defines goodness and evil in terms of God's approval or disapproval.

However, if the goodness of a thing is a product of God's approval, the thing's goodness cannot also be the source of that approval. Yet the only reasonable explanation of God's approval of a thing is its goodness. Given the divine command theory's definition of goodness, claiming that God approves of a thing because it is good reduces to the vacuous claim that God approves of that thing because it is something of which he approves. So, if the only reasonable explanation of God's approval of an action is its goodness, but this explanation reduces to a tautology, then God's approval must apparently be inexplicable, and hence arbitrary. If his approval is arbitrary, God could just as well have approved of

torturing kittens and disapproved of feeding and housing starving orphans.

Theists who find this objection telling but who still wish to reserve an ethical role for God typically avail themselves of something like Cummins' and Gould's strategy. Since in order to avoid the arbitrariness of morality we must be able to explain God's approval of a thing in terms of its goodness, we had better find some independent, intrinsic criterion of goodness, say, natural law theory or utilitarianism or the categorical imperative. God's approval can then be explained as the result of his perfect appreciation of this intrinsic goodness, and his commands can be reasonably held up as a model for our behavior on the grounds that he is much better than we at discerning it (*modulo*, of course, any concerns we might have about the authenticity and/or authorship of the commands, or the existence of the commander.)

In all of these cases, an attempt is made at constructing a satisfactory explanation. However, in each case it appears as though the authors of the explanation have offered an explanation of the form "(1) *x* has F because (2) *x* has G and (3) things with G always (or usually) have F," where the concepts F and G are defined in

such a way that it is impossible to fully understand that *x* has G without also understanding that *x* has F. Such explanations are vacuous insofar as anyone in a position to understand (2) thereby already understands (1), and hence has no need for (3). For example, in the natural selection case, the creation scientist claims that the explanations of natural selection vacuously assert that organisms (*x*) survive (F) because they are fit (G). In the divine command theory case, the critic of the divine command theory argues that if the theory were true, the fact that an action (*x*) is approved of by God (F) could not be non-vacuously explained by the fact that the action is good (G). And use theories of representation claim that a cognitive system (x) behaves in a certain way (F) because it employs a representation with a particular content (G). In each case, the problem is that F and G are defined in such a way that *x*'s having G entails[1] *x*'s having F, and

[1]However, the vacuity does not always arise from entailment. Recall Cummins' objection to Millikan, in which organisms learn to exploit the previously unexploited content of a representation their ancestors used less effectively. In this case, the organisms' (*x*) novel behavior (F) is explained as resulting from their more efficient use of the representation's content (G). However, according to Millikan's use theory, *x*'s having G is defined, not in terms of *x*'s having F, but rather in terms of *x*'s *ancestors* having F. Since *x*'s having G does

consequently, (3) becomes a tautology. The cases also share a solution – don't define F and G in such a way that *x*'s having G entails *x*'s having F. Instead, define G in its own terms, and then use *x*'s having G along with an independently justifiable claim (3) to explain x's having F.

Vacuity and Alternative Models of Scientific Explanation

So far, I have described the problem of explanatory vacuity in a way that more or less assumes a deductive-nomological or causal model of scientific explanation. I hope by now it is clear that at least according to these sorts of models of explanation, vacuous explanations are not explanatory. If "*r* represents *c*" is defined in terms of the behavioral effects of the employment of *r*, then

---

not entail that *x* has F, Millikan's theory avoids the circularity present in the other cases. However, the fact that *x*'s having G does entail that *x*'s ancestors had F is still problematic, since, by hypothesis, *x*'s ancestors lack F (the novel behavior). As a consequence, Millikan's theory implies that *x* does not have G (the detailed content), and hence, *x*'s possession of F cannot be explained by an appeal to the false claim that it does have G. In this case, the reason it is impossible to fully understand that *x* has G without also understanding that *x* has F is simply that it is impossible to understand that *x* has G, since it is impossible to know a falsehood.

according to these models, when we assert that the employment of *r* explains the production of those behavioral effects, we are either saying that the behavior is explained by the system's use of a) something that is constantly conjoined with that sort of behavior as a matter of natural law, or b) something that causes that sort of behavior. Either way, we haven't learned anything new that bears on the question of why the behavior occurred beyond that which is presupposed by the model of explanation itself. We learn that the behavior resulted from the operation of the laws of nature on antecedent conditions, or that the behavior had a cause, but we don't learn anything about the relevant antecedent conditions or about the cause.

However, these models of explanation are not the only game in town, and (especially if one is a hard-pressed use theorist looking for a way out) one might wonder whether some other plausible model of scientific explanation might tolerate this vacuity better. The proper way to deal with this worry would be to exhaustively map out all the plausible alternative models of scientific explanation and show that use theories produce unacceptably vacuous explanations on any of these models. Unfortunately, this

is not a realistic tactic within the scope of this work. The literature on scientific explanation is so voluminous that anything approaching an exhaustive survey would turn this small section of the first chapter into a complete book of its own. Consequently, I intend to briefly discuss only two alternative models that seem to me to be the best bets for the desperate use theorist, in order to demonstrate, at the very least, that developing a model of explanation satisfying her requirements will be forbiddingly challenging.

The most significant challenge to the problem of explanatory vacuity probably comes from Bas van Fraassen's insistence that scientific explanation is pragmatic (van Fraassen, 1980, p. 97-157). Explanations are, he says, answers to "why questions," and such questions generally presuppose an implicit, contextually determined range of relevant answers. So, for example, if we ask the question, "Why is the porch light on?" both the answer "Because the switch is flipped to the on position," and the answer "Because company is coming," may be considered good explanations. Which one is the correct explanation depends on whether the context of the question suggests that the interlocutor is interested in learning something to do with

electrical engineering or something to do with the social conditions of the coming evening.

If this is right in general, then it might similarly be true of explanations involving representations. I have been asserting that the *explanandum* of cognitive science is the production of intentional behavior, so the relevant why question in cognitive science will generally be of the form "Why did that cognitive system behave in such and such a way?" But if van Fraassen is right, this question does not, by itself, determine what sort of *explanans* will be appropriate. We also need to know the relevant context.

Thus far, since I have been presenting the problem in deductive/nomological and/or causal terms, I have been more or less assuming that the context is one according to which the why question gets disambiguated as something like "Why did the cognitive system behave in such and such a way, rather than in this or that other way?" That is, the cognitive scientist is trying to determine why, say, the robot managed to avoid an obstacle rather than crashing into it, or stopping dead in its tracks, or sitting down and sobbing at the futility of life's challenges, or whatever. If this is indeed the right context, then it seems clear that an explanation in terms of the cognitive

system's use of representations will remain unacceptably vacuous so long as representation is defined in terms of use. Since we learn nothing about what representations are like from such a theory, we can't explain why any given representation should have this behavioral effect rather than that one.

However, it seems possible that there may be other contexts in which the same why question might receive a non-vacuous answer. This is because the use theorist's definition of representation at least arguably falls slightly short of total vacuity. For, if the use theorist defines a representation as something that is used by a certain cognitive system to produce certain behavioral effects, he seems to have at least told us that, in addition to having those effects, a representation is something that gets used by cognitive systems.

This rather meager bit of information about the representation won't help us non-vacuously explain why the system engages in any particular behavior, since that would require knowing something about the employed representation as opposed to various other representations that might have been employed. Explaining why the robot avoided the obstacle rather than crashing into it requires that we know

something about what made the representation it used, say, an accurate rather than an inaccurate representation of the object's position, and use theories can't give us that.

But the meager bit of information might allow us to answer the why question about behavior if we loosen up the context. For example, what if the context suggests that we should disambiguate “Why did that cognitive system behave in such and such a way?” as asking "Why did that cognitive system behave in such and such a way rather than fail to engage in any behavior at all?" That is, rather than asking how our robot differs from other robots that fail in the avoidance task, the questioner is asking how our robot differs from the wastebasket in the corner of the lab, which is and always has been completely inert. In this case, learning that the robot, unlike the wastebasket, employs representations might allow us to answer the question, and hence to non-vacuously explain something. Roughly, we learn that the robot's behavior is somehow the result of an internal process of some (unelaborated) kind, as opposed to its being somehow externally motivated. So, a use theorist who adopted van Fraassen's pragmatic model of explanation might be able to claim that representations

defined in terms of use can at least be employed in these sorts of non-vacuous explanations of behavior.

One response to this challenge would be to question how much one really learns in this case. In particular, as I will discuss in a later chapter, the claim that a representation has been used by a cognitive system is quite theoretical, and thus arguably what one learns from this explanation is even more minimal than it already appears to be.

However, the easiest way to respond, I think, is to allow, at least for the sake of argument, that this rather minimal sort of explanation really is non-vacuous, while insisting that cognitive science additionally requires other explanations that do become vacuous when one adopts a use theory. If the only thing cognitive science can non-vacuously explain is the fact that some things, rather than others, exhibit behavior, it isn't going to be a very impressive science. Furthermore, the meager bit of information that explains this fact is a basic assumption of cognitive science. If we didn't think that there were internal processes governing behavior, we wouldn't engage in cognitive science at all. Rather, the point of cognitive science is to explain, not simply that some

things exhibit behavior, but that various different sorts of things exhibit various different sorts of behavior under varying circumstances. Nothing in the pragmatic model of explanation will allow us to simultaneously accept a use theory and produce these latter sorts of essential explanations.

The second alternative model of explanation I will consider is Kitcher's unification model of scientific explanation (Kitcher, 1999). According to this view, an argument forms the basis of an explanation when it instantiates an argument pattern that unifies the explanation of many phenomena. In identifying such patterns, we are supposed to look for patterns of argument that are simultaneously relatively specific and yet widely applicable to the phenomena to be explained. So an argument in physics will count as explanatory when it fits one of several detailed and yet common patterns for explaining broad classes of physical phenomena. Appeals to the law of gravity and the distribution of masses, for example, will be explanatory because explanations of this precise form can be found for not only planetary motions, but also terrestrial ballistics and so on. Other patterns, such as Darwin's form of explanation in terms of natural

selection, will also count as explanatory since explanations of this more or less equally rigorous form also accommodate a (rather different) wide range of phenomena.

The reason that this sort of unification model of scientific explanation might be a promising route for the use theorist is that the model contains no specific requirement that explanations mention natural laws or causes. Kitcher, for example, claims that the explanations of natural selection need not assert any laws (Kitcher, 1985). So a use theorist might adopt Kitcher's unification model of explanation and try to claim that, despite the vacuity of explanations which appeal to representations defined in terms of use when they are construed in a deductive/nomological or causal way, if we can identify the right sort of pattern in these arguments, they can still be explanatory.

However, despite dropping any requirement that explanations appeal to laws or causes, Kitcher's unification model still won't allow us to tolerate the vacuity. This is because vacuity is a property of arguments, and according to Kitcher, explanations still have to involve arguments. How could it be otherwise?

Even if we allow that explanations need not explain phenomena by giving reasons citing laws or causes, we surely must accept that explanations have to give reasons of some sort. But if explanations involve giving reasons, they involve arguments, and it is a matter of logic that vacuous arguments don't provide good reasons. Consequently, even if representations defined in terms of use could figure in explanations that instantiate the sort of patterns Kitcher accepts as explanatory, if those explanations all involve examples of bad arguments, they won't count as explanatory on Kitcher's view.

Admittedly, this brief survey doesn't prove that there is no conceivable plausible theory of explanation that would count some sort of explanation of behavior that appeals to representations defined in terms of use as a good explanation. But I think it provides fairly good reasons for skepticism. The essential problem is that explanations that appeal to representations defined in terms of use tell us little or nothing about how having particular representations with particular contents figures in the production of behavior. If cognitive science can't tell us this, then it is simply not clear what cognitive science is for.

Cummins' Theories of Representational Content, Target Fixation, and Propositional Attitude Content (A Solution to the Problem of Explanatory Vacuity)

Before I introduce Cummins' theory, I should briefly mention a second problem which shapes it. In addition to the problem of explanatory vacuity, Cummins also claims that use theories conflate two distinct sorts of mental content: propositional attitude content and representational content. This is because it is propositional attitudes, rather than representations, that are directly used by cognitive systems in the production of their intentional behavior. A mouse, for example, needs beliefs that there is a bird in the sky, desires that the bird go away, and fears that the bird will harm it, if it is to avoid predation by raptors.

But the survival of the mouse does not necessarily require its having representations of birds available to its various intenders. This is because the fact that the mouse believes that there is a bird in the sky does not necessarily imply that the mouse is employing a |bird|. The mouse might believe there is a bird in the sky by having a THINGS-IN-THE-SKY-intender which employs a |bird|.

But it might also produce this belief by having a CURRENT-VISUAL-ENVIRONMENT-intender which employs a |bird-in-the-sky|, or even a PLACES-THERE-ARE-BIRDS-intender which employs a |sky|, or a BIRD-IN-THE-SKY-intender that employs a |true|. To be sure, cognitive systems do use representations, as they need them to form propositional attitudes, which they in turn use in the production of intentional behavior. But it is a mistake to take intentional behavior directed at *x* as directly indicative of the system's use of |*x*|s. Rather, intentional behavior directed at *x* is merely evidence that some representation is being used to form certain propositional attitudes about *x* (Cummins, 1996, p. 15).

Furthermore, those who fail to draw the distinction between attitude content and representational content will also fail to draw a distinction between falsity and misrepresentation. Representations play a role in determining the content of many types of propositional attitudes which cannot be appropriately labeled as true or false. If I believe that some aspect of the world can be accurately represented by *r*, and this is not the case, I have a false belief. But if I desire, or suppose, or fear that the world is representable by *r*, I cannot be said to

have done so falsely, no matter what *r* is a representation of.

However, I still could have misrepresented. Perhaps I am engaging in a bit of hypothetical reasoning about what will happen if John Ashcroft succeeds George W. Bush as president. If my "SUPPOSITIONS-ABOUT-THE-FUTURE-intender" malfunctions and employs a |Jodie Foster is president| when it should have employed a |John Ashcroft is president|, it has misrepresented my hypothesis. This may lead to certain false beliefs, such as the belief that when Ashcroft is president I will probably want to watch more presidential addresses on television, but it is not itself a falsehood. Furthermore, even if my SUPPOSITIONS-ABOUT-THE-FUTURE-intender employed a |George W. Bush is president|, this would be a misrepresentation, even though that representation accurately represents the way the world now is. The SUPPOSITIONS-ABOUT-THE-FUTURE-intender is not targeted on how the world now is, but rather on how I am supposing the world might be in the future, so a |George W. Bush is president| is the wrong representation for the job at hand (Cummins, 1996, p. 11).

Thus, Cummins claims that in order to deal with propositional attitudes, we really need three sorts of

semantic theory. First, we need a theory of intrinsic representational content. Second, we need a theory of target fixation. This theory is supposed to tell us how to identify the target of any given employment of a representation. For example, it will tell us whether the mouse needs to employ a |bird| or a |sky| in order to avoid predation, by telling us that its intenders are targeted on this or that feature of its environment. This theory will involve use, since targeting determines what the use of any given representational token is, but it will not be a use theory of representational content. Instead it is a theory of how representations are used. Considerably more will be said about functions in the second chapter.

Once we have these two theories in place, we are able to say what propositional attitudes are and what their semantic contents are. Propositional attitudes are attitudes we can take towards applications of representations to targets, where "the semantic content of an application is that the representation hits the target," (Cummins, 1996, p. 16). In other words, the content of a propositional attitude is the state of affairs which obtains when the representation toward which the attitude is held has a content that is identical to its target.

Cummins' theory of representational content is simply that a representation *r* has a content *c* just in case *r* is isomorphic to *c*. In this sense, mental representation is just like mathematical representation. As he puts it

> If a structure R represents a structure C, then
> 1. An object in R can represent an object in C.
> 2. A relation in R can represent a relation in C.
> 3. A state of affairs in R – a relation holding of an n-tuple of objects – can represent a state of affairs in C (Cummins, 1996, p. 96).

Furthermore, "since isomorphism is a relation between structures, it follows that, strictly speaking, only structures can represent or be represented" (Cummins, 1996, p. 109). The structure Cummins is speaking of is the same sort of structure common in mathematical logic. A simple relational structure is just an ordered pair consisting of a set of objects, and a set of relations defined over those objects and/or other relations in the set. Two structures A and B are isomorphic just in case there exists a one-to-one, onto function which maps each object in A to an object in B, and each relation in A to a relation in B.

This is clearly an intrinsic theory of representational content. Whether or not a representational token is isomorphic to whatever it is that

it is currently being used to represent can be determined simply by comparing the representation with the target structure. If the former is isomorphic to the latter, the representation is accurate; otherwise we have a case of representational error. In the determination of accuracy, no mention is made of any use by any cognitive system. In fact, any two isomorphic structures represent each other, regardless of whether any cognitive system employs one as a representation of the other.

Cummins' theory of target fixation is to be derived from a theory of intender functions, perhaps not unlike Millikan's. As we have seen, Cummins does not think that teleological theories of representational content will work because such theories simply amount to particular ways of fixing the ideal cases in use theories. But even though teleological considerations cannot provide an adequate theory of representational content, he claims that they are well suited to the task of determining the functions of intenders. In Cummins' theory, teleology tells us how representations are used by telling us what they are targeted on, but it does not tell us what they mean. Thus, Cummins avoids the problem of explanatory vacuity, not by

banning teleology outright, but by putting teleology in its proper place (Cummins, 1996, p. 118).

So, Cummins' total theory can be summarized as follows:

(1) A mental representation is a structure S, and its content, *c*, is whatever structure is isomorphic to it.
(2) A mental representation is targeted on whatever structure, *t*, the intender that employs it has the function of representing.
(3) A propositional attitude's meaning is determined by
    (a) the attitude type (i.e. belief, desire, etc.), and
    (b) the content that *c* = *t*.

Cummins' theory of representational content is stated in (1). It tells us what mental representations are, and which mental representations have which contents. It says nothing about how mental representations are used. Rather, it is a philosophical analysis of the term "mental representation."

Cummins does tell us how mental representations are used in (2) and (3), but (2) and (3) are not part of his theory of representational content. Rather, (2) and (3) are part of empirical cognitive science, which is, in part, the study of the ways that cognitive systems use representations, and the explanation of various facts about those systems and their behavior in terms of their use of

those representations. More specifically, (2) tells us what the cognitive system is attempting to represent with the representations it uses. (3b) tells us that the content of a propositional attitude is always that the content of the representation is identical to the target, or in other words, that the representation is not a misrepresentation. (3a) then asserts that we can take various attitudes towards the claim that our representations are not misrepresentations. For example, if I believe that my representation of a tiger in my living room is accurate, I will run away, if I doubt that the representation is accurate, I will rub my eyes and put on my glasses, and if I desire that the representation be accurate, I am apparently strangely fond of close encounters with large carnivores.

The Problem of Underdetermination of Representational Content

In *RTA*, Cummins himself raises the problem which is fatal to his theory of representational content. There are simply too many isomorphisms of any given representation for isomorphism alone to pin down the representation's

meaning. Cummins defends isomorphic mental representation because it seems to be able to explain, independently of use, how certain structures in a cognitive system's brain come to represent their contents. But these structures will also be isomorphic to lots of other things in the system's environment. If Cummins is right, and these mental structures really represent their contents merely by being isomorphic to them, they will additionally and equally represent a whole bunch of other things, and consequently will end up effectively representing nothing at all.

Consider, for example, Cummins' description of what he calls the "Autobot" (Cummins, 1996, p. 94). The Autobot is a little car which is designed to turn right or left in response to input from a grooved card placed into a slot at one end, which is then pulled through the car as it moves. A pin on the steering mechanism fits into this groove, so that notches in the groove will push the pin to the right or left as the card passes through the car, causing it to turn to the left or right, respectively. By properly spacing these notches on the card, the car can be programmed to make a specific series of left and right

turns at specified intervals. We could, for example, program the car to navigate a maze.

Cummins' claim is that the groove structure of such a card represents the path through the maze. As he himself notes, however, if representation is just isomorphism, this is not all it represents. If we insert the card upside down, the car will follow a mirror image of that path. It is tempting to appeal at this point to the proper use of the card: to say that the card, properly used, only represents the first path. But this would be to retreat to a use theory. Instead, Cummins claims that representational content can underdetermine the causal efficacy of the propositional attitude in which it appears.

> The fact that the attitudes have the same content does not imply that they are the same state...There is more to an attitude (though not to its content) than the representation's content, the target, and the [attitude type]. There are, in addition, lots of causally relevant physical facts about the representation that do not affect its content or what it is applied to, or which kind of attitude happens to be at issue, such as the orientation of the card relative to the car (Cummins, 1996, p. 100).

So, the content of the card representation is indeterminate between the first path and the second path. If we wanted to make the car a more reliable navigator, we could design it so that it rejected the card in one or the other of

these conditions, but this would not, in Cummins' view, change the content of the representation.

This solution might work if there were only a few alternate uses of some representation. Part of the explanation of the behavior of the cognitive system would appeal to the presence and application of the representation, and part of it would appeal to other design features of the system. Representation would not be a complete explanation of behavior, but it would be a significant component, and that might be enough. Certainly, some behavior, even in sophisticated cognitive systems, is not the result of representation. A human being, for example, might stop running because his "CURRENT-GOALS-intender" has produced a representation of his body at rest, but he might also stop because he has become fatigued, or because he has just hit his head on a low tree branch and knocked himself out cold. In the latter two cases, non-representational processes are involved in determining the person's behavior. So if there are other non-representational processes involved in ensuring the proper use of representations, this is just another case where representation is insufficient, by itself, to explain behavior.

Unfortunately though, this is not the case. In general, there will be many content structures isomorphic to any representation. In a footnote, Cummins cops to two additional ways the Autobot card could be inserted: it could be run through the car backward, turning the sequences {right, right, left, left, right} into {right, left, left, right, right}, and {left, left, right, right, left} into {left, right, right, left, left}. But at this point we are just getting started. Since Cummins claims that the meaning of a representation is not determined by the use to which it is put, the card's structure will also represent anything that any conceivable cognitive system could use it to represent. For example, we could devise an autopilot for an airplane which would take the card to prescribe a series of altitude changes to which its notches are isomorphic, such as {climb, climb, dive, dive, climb}. Placed in a suitably designed drawing machine, it could represent a topographical map, such as {hill, hill, valley, valley, hill}. As it happens, it could also represent the series of long and short tones produced when signaling the Morse code letters "MU" or "IG." Obviously, there are many more such examples.

Thus, the problem of underdetermination leads to a different sort of explanatory problem. If the card representation has a content that is indeterminate between all of these things, and anything else to which it might be isomorphic, the representational content of the card will play almost no role in the explanation of the behavior of cognitive systems. With so many bizarrely unrelated isomorphisms for a typical representation, we could only say what the representation means by listing as a giant disjunction all the possible isomorphic structures. (Now that's a disjunction problem!) The vast majority of this explanatory work will instead be done by non-representational "causally relevant physical facts" about the way particular systems use representations. This essentially amounts to a retreat from Cummins' own requirement that our theory of representation leave a significant explanatory role for representation. Since the failure to leave such a role for representation is what drove Cummins away from use theories in the first place, the fact that his alternative theory leads to a similar result clearly calls that theory into question.

Before we go on, it is worth noting that this objection, significant though it is, can easily be

overstated. Though Cummins is careful to point out that his representation relation holds between structures rather than the objects that instantiate them, he also notes, and then frequently succumbs to, the temptation to speak as if it were the instantiating objects themselves that were the representations (Cummins, 1996, p. 109). For example, he says that the Autobot card represents a path through a maze, rather than that the structure of the notched groove in the card represents that path. So long as we take the former claim as an abbreviation for the latter, there is no harm in this. However, it can be easy to slip into thinking of the instantiating objects as the representations themselves, and if one does this, the underdetermination becomes much worse.

Millikan appears to have interpreted Cummins this way in her review of *RTA*.

> Cummins asserts, for example, that if you take a certain map, it either is or is not perfectly isomorphic with Chicago! But there is an indefinite and perhaps infinite number of ways to cull from the things contained in that concrete map a set of objects to consider. Nor does the notion of isomorphism require that the members of the set be designated in any principled way. They could be any objects one felt like listing. And there is an indefinite number of relations one might consider among members of these sets of objects (Millikan, 2000, p. 108).

To return to the Autobot example, the card (in this misleading manner of speaking) will represent the path through the maze, the mirror image of the path, a series of hills and valleys, Morse code “MU”, etc., all in virtue of isomorphism between the structure of its notched groove and structures instantiated in all of these various objects or states of affairs. But there are all sorts of other structures instantiated in the card. It has a chemical structure, a color structure, a mass-distribution structure, and so on. All of these structures will be isomorphic, according to one mapping rule or another, to structures instantiated in a variety of objects. If Cummins had claimed that representations are the objects which instantiate various representational structures, and that these objects represent whatever other objects instantiate structures isomorphic to any of those representational structures, his theory would indeed fall prey to this more severe sort of underdetermination.

However, Cummins can avoid this aspect of the problem simply by being more scrupulous in his presentation. His actual view is that representations are not objects instantiating structures, but rather the structures themselves. If a particular concrete map instantiates

multiple structures, it thereby instantiates multiple representations, each of which represents only those content structures to which it is isomorphic. Nevertheless, as I have noted above, the less severe problem of underdetermination is still fatal to Cummins' theory.

Mapping Rules

(A Solution to the Problem of Underdetermination)

Despite this problem of underdetermination, there is something very appealing about the idea that mental representation, like mathematical representation, has something to do with isomorphism. Indeed, many other philosophers and scientists seem to have assumed that it does. Millikan, for example, makes frequent reference to isomorphic relations between representation and content.

> Intentional icons are akin to those structures mentioned above that map onto or are isomorphic to environmental features to which they are adapted...They exhibit a dimension or dimensions of possible variance running parallel to possible variances in the environment. A rule of projection, in the mathematician's sense, maps the one onto the other. We can call this rule the icon's "mapping rule" (Millikan, 1995, p. 99).

Another example of isomorphism at work in the context of mental representation may be found in Kathleen Akins' humorously titled reply to Thomas Nagel, "What is it Like to be Boring and Myopic?" In this paper she describes in great detail the efforts of researchers to determine what environmental features bats can perceive with their sonar-like sense. After describing the mechanics of bat sonar, she surveys the results of cortical mappings produced by single-neuron recordings. Here are the summarized results for one region of the bat's motor cortex.

> Using roughly polar coordinates, the map plots signal frequency against signal amplitude...Radiating out along the spokes, the preferred frequencies of the neurons start at about 61 kHz and extend out to about 63 kHz at the periphery; going around the wheel...the preferred amplitudes increase (Akins, 1995, p. 143).

So patterns of firing in this section of motor cortex are isomorphic to frequency/amplitude pairs in the bat's auditory environment. Akins goes on to describe various hypotheses about what this graphically represented information might represent to the bat. Various aspects of this plot may allow the bat to determine the rate of a moth's wing beats, the size of the moth's wings, the moth's orientation relative to the bat, or the subtended angle of

the moth. In each case, this is because some aspect or other of the frequency and amplitude data (and hence the motor cortex) is isomorphic to the relevant aspect of the moth or its behavior.

In both cases, there are two elements involved in the implied account of mental representation: (1) a mental structure that is isomorphic to the representation's content, and (2) a particular "rule of projection" or "mapping rule" which relates that mental structure to its content. The second component is absolutely crucial. Just as with the Autobot card, the bat's polar coordinate graph of frequency vs. amplitude can only represent aspects of the bat's acoustic environment if locations of activation in the graph are mapped to sound-frequency/sound-amplitude pairs, rather than, say, time/Dow-Jones-Average pairs, or GNP/baseball-score pairs. It is because his theory lacks this component that Cummins' is afflicted by the problem of underdetermination. However, if we could determine which mapping rule to apply to the bat's isomorphic mental state, or to the structure of the Autobot card, it would be possible to cut down the number of possible contents which those states represent. This suggests that there are two essential components to any mental representation: a

mental structure which is isomorphic to its content, and a mapping rule that maps the mental structure to its content structure.

However, Cummins' discussion of language suggests a possible objection to this claim. It would appear that language is, or at least can be, entirely conventional. The word "cat," for example, is not obviously isomorphic to any cat by any mapping rule. So it seems that linguistic representation cannot be a simple matter of isomorphic structure plus mapping rule. Furthermore, there seems to be no good reason why at least some mental representations might not be similarly conventional. As we saw above in Cummins' criticism of Fodor, causal theorists are committed to the claim that they all are.

This fact suits Cummins just fine, as he explicitly denies both that the sort of conventional relations found in language are representational, and that mental representation is ever similarly conventional. Language, he claims, does not represent at all, but merely communicates.

Cummins' insistence on this view appears to be motivated by his intrinsic theory of representational content. In order to make his theory intrinsic, Cummins

relies on similarity of structure between representation and content to tie representations to their contents. But if representations can be linked to their contents conventionally, they needn't share structure. For example, the word "dog" isn't obviously "dog-structured," and even if it were, such similarity of structure wouldn't be required by the rules of language. "Cat" or "nine-iron" could be conventionally tied to dogs just exactly as well as "dog" is. So similarity of structure isn't going to ground intrinsic content assignments for linguistic representations, or similarly conventional mental representations. The problem of underdetermination arose because isomorphism alone includes too much in the extension of a representation, but with conventional representation nothing at all will be excluded. Consequently, Cummins fears that barring some other intrinsic way of assigning contents to representations, accepting linguistic representations would require accepting a use theory of, at least those, representations.

Thus, Cummins claims that communication schemes based on convention are non-representational.

> The distinction I have in mind here is like the old distinction between eminent and formal reality...A physical triangle or dog is physical

> matter informed by *triangularity* or *doghood*. An idea of triangularity or doghood is mental matter informed by *triangularity* or *doghood*...To mentally represent triangularity is to have your mind stuff informed by the very same form that makes something physical triangular. When you see a triangular thing, there is a transfer of form from the thing to the mind, in something like the way that pressing a triangular object into a block of wax is a transfer of form. When I say "triangle" to you, the form in my idea has to be carried to your mind by my word. The word, however, is not informed by triangularity, for while one can have a triangle in physical matter, and a triangle in idea, one cannot have triangular sound. So, *triangularity* does not inform the word, but it is still in it eminently, *conveyed* by it in such a way that your mind, on the receiving end, can reconstitute the form in your mind stuff (Cummins, 1996, p. 131-132).

In Cummins' hands, form becomes structure, being informed by the same form becomes isomorphism, and mind stuff becomes brain stuff. The eminent transfer of form in linguistic communication becomes the activation (or attempted activation) by a speaker of certain concepts in the receiver by the use of signaling conventions.

> Words, then, are signals, not representations. Signals have communicative functions; representations have semantic contents. When Paul Revere displayed his famous lantern, he signaled that the British were coming by land, but he did not represent that fact...The lantern-waving was governed by a temporary convention shared by only a few parties. The sentence ["the British are coming by land"] is governed by a convention shared by all speakers of English. The sentence no more represents the British route

> than does displaying the lantern (Cummins, 1996, pp. 140-141).

Since words and sentences merely convey the structures of the associated mental representations of speaker and hearer, they are not themselves representations, and hence need not be covered by an intrinsic theory of representational content. Words don't have to be given intrinsic contents, because they aren't the sorts of thing that have contents at all.

However, I think that drawing this distinction between structural representation and conventional signaling is unnecessary, and ultimately a mistake. Instead, I intend to claim that all representation, including linguistic representation, is a function of both isomorphism and a mapping rule. If we incorporate mapping rules into our theory of representational content, we can solve the problem of underdetermination, even for conventional representation. Furthermore, as I will show in the next chapter, the addition of mapping rules need not commit us, as Cummins fears it will, to a use theory.

First, however, I want to discuss further my claim that both isomorphism and mapping rules are essential components of any sort of representation. The role of one

or the other of these components in determining the content of any particular representation may be vanishingly small, but both components are always present.

In the case of single words or of similarly conventional mental representations, the isomorphism component can be trivial. Suppose I have an extremely simple language which consists entirely of a set N of nouns which correspond to a set O of sets of physical objects. Then my simple noun-language models the structure ⟨O, ∅⟩ with the structure ⟨N, ∅⟩ according to a mapping rule which maps nouns from N to their corresponding sets of physical objects in O. Of course, these "structures" haven't got much structure at all since their set of relations is empty. We might call such structures "degenerate structures" and the mapping that holds between the structures a "trivial mapping." The mapping rule that assigns "cat" to cats and "dog" to dogs is no more or less appropriate than the mapping that assigns "cat" to dogs and "dog" to cats, or any other mapping. Only convention tells us which mapping rule to use in determining the meaning of the nouns in N. The isomorphism is there, but it does none of the heavy lifting in the content determination. Still,

if these structures were not at least trivially isomorphic in this way, no mapping from N to O would exist at all.

In the case of more realistically complicated languages, isomorphism arguably plays an expanded role, if only a slightly expanded one. As Millikan notes, sentences have a grammatical structure which can be used to perform various operations on them (Millikan, 1995, p. 105-106). Verbs, for example, have a structure that requires some combination of subject, direct object, and indirect object. Different nouns (or noun phrases) can be substituted into any of these slots. That this is so entails that the sentences produced by these substitutions are isomorphic to the ones from which they are produced. "John loves Mary," is isomorphic to "Mary loves John," but not to "John gives Mary the ring." The first two have the form R^2xy, while the second has the form R^3xyz. And, to the extent that these structures accurately represent the real-world relations the sentences describe, the sentences will be isomorphic to the content structures they represent as well.

Even so, though wrong to deny that language is representational, Cummins is right to point out the largely conventional nature of linguistic representation. While there is a fair amount of structure in the relations

between words, the relations of words to the objects and relations they represent appear to be largely conventional. If I hear the sentence, "The dog ate the food," the sentence may represent to me, in virtue of its linguistic structure, that two things are related in some way. But unless I know what the words are conventionally tied to, the sentence will not represent to me that the dog ate the food. Isomorphism between the structure of the sentence and the structure of the state of affairs consisting of the dog's eating the food does some of the representational work, but most of the work is done by specifying the mapping.

What about the second component? Are there no cases of representation without the specification of a particular mapping rule? We have already addressed this in criticizing Cummins' theory and the problem of underdetermination. One might think, for example, that a graph of a parabola represents the trajectory of a thrown baseball solely in virtue of its isomorphism to that trajectory. However, without specifying through some mapping rule what the points of the graph are supposed to represent, the collection of the points on the parabola could just as easily represent the trajectory of a comet

around the sun, or the distribution of grades in Philosophy 101.

Unless I have been insufficiently imaginative in my search for cases of representation with either a vanishing isomorphism component or a vanishing mapping rule component, I take the preceding observations to demonstrate my claim that all representation, including but not limited to mental representation, includes both components. Once one recognizes the existence of degenerate structures and the trivial mappings between them, this claim should really be quite uncontroversial. Any attempt to claim that the objects in one set are representations of the objects in another will obviously involve putting the objects in the first set into a one-to-one correspondence with the objects in the second set. In some cases, it may be that the corresponding object pairs all instantiate identical non-degenerate structures, and if so, there will exist a non-trivial mapping rule from the structures instantiated in objects in the first set to the structures instantiated in the objects in the second set. In such cases, this non-trivial isomorphism could be exploited for representational purposes. However, even if the corresponding objects do not share any non-degenerate structures, they must at least

share degenerate structures, which are trivially isomorphic.

So even if one thinks, like Cummins, that all representations are non-trivially isomorphic to their contents or, like causal theorists, that no representations are non-trivially isomorphic to their contents, one should still be able to accept my claim that all representations involve both isomorphism and the specification of a mapping rule. Furthermore, though Cummins explicitly rejects the representational role of mapping rules, this rejection can be reinterpreted as an insistence that the only mapping constraints are those imposed by the isomorphism itself. That is, the mapping rule of one of Cummins' representations takes it to the set of all objects isomorphic to the representation.

That said, it does at least intuitively seem as though what we end up with is a continuum of representations. At one end, the content of a representation is determined almost entirely by an arbitrary mapping rule, while at the other, the content is determined largely by isomorphism. Sensory representation provides a clear illustration of this continuum by apparently inhabiting a wide swath of it. For example, consider visual representations. Suppose that

I am at an air show examining an F-16 jet fighter. From my vantage point right next to the plane, I can see a great deal of detail: individual rivets, oil streaks, my reflection in the polished metal, and so on. I see these things because the patterns of light reflected from the surface of the plane which impact my retinae are isomorphic, in various ways having to do with the properties of light, to certain aspects of the plane. The patterns of stimulation on my retinae are then converted by my visual system into visual representations. Since both the physical properties of the light and the operations performed by the visual system preserve a good deal of this structure, my visual representation of the plane is richly isomorphic to the plane itself.[2]

However, if the plane takes off and begins a spiral climb over the airfield, my current visual representation of the plane will gradually become less and less richly

[2]Technically, isomorphism does not come in degrees. By "richly isomorphic" I mean that the mental representation has a structure which is isomorphic to a relatively large proportion of the total structure of the plane.

isomorphic to the fine details of the plane. At a few hundred feet, I can no longer see rivets and oil streaks, but I can still see the external fuel tank, pylons, and landing gear. At a few thousand feet I lose these details as well, but can still tell from the shape of the wings, tail, and nose that it is an F-16 rather than some other sort of jet fighter. At several thousand feet, I can still tell that it is a fighter because it is too pointy to be anything else, but cannot tell what sort. Eventually, I see only a dark speck at the head of a condensation trail. At this point, the isomorphism component has been reduced to the trivial isomorphism present in the noun-language example. The speck is isomorphic to the plane only in such a way that it represents that there is “a thing” at such and such a location in the sky. As the isomorphism component contributes less and less to the determination of content, that determination comes to rely more and more on the mapping rule.

Use Theories and Mapping Rules

In the previous section, Millikan appeared as an example of a philosopher who explicitly embraces the two

essential components of representation. But she is also a use theorist. As Millikan herself notes, her theory employs the same basic elements that Cummins' theory does. She agrees that representations are isomorphic to their contents and, like Cummins, she appeals to the function of the intender producing the representation in order to allow for misrepresentation. The difference, however, is in the way these ingredients are combined. For Cummins, all there is to being a representation of some content is being a structure that is isomorphic to that content. As we have seen, this leads to the problem of underdetermination, since any structure that is isomorphic to one content structure is generally isomorphic to many others. But this problem does not arise for Millikan, because she thinks that, while isomorphism is an important ingredient of representation, there is more to representation than isomorphism. What Cummins calls a representation, Millikan calls an "isomorph," and then goes on to say

> *LTOBC* claims that the proper function of the producer of an indicative or fact-stating representation is not merely to produce an isomorph of a certain sort of thing, but to produce exactly that sort of isomorph of that thing that the consumer knows how to use, hence to produce something that is isomorphic in accordance with a definite projection rule. And *LTOBC* doesn't call what is produced by the

> producer an “intentional icon,” nor claim that it has any “semantic content,” unless it is the sort of thing that is in the domain of the projection rule for icons the consumer has been designed (by natural selection or by learning) to read (Millikan, 2000, p. 106).

In other words, according to Millikan, an isomorph is not a representation unless it is produced by an intender which has the function of producing such an isomorph for the purpose of representing a particular structure to which it is isomorphic. To put the immediately relevant difference between the two as succinctly as possible, Cummins thinks that representations are isomorphs, while Millikan thinks that representations (she would call them intentional icons) are isomorphs that are used in a certain way. This is, of course, what makes Millikan’s theory a use theory. As was discussed earlier, other use theories do essentially the same thing, but have different principles for determining which uses are ideal.

One way to characterize these ideal uses is by reference to the mapping or projection rule that holds between the isomorphs produced by an intender and the targets of that intender when it is functioning in the ideal way. The mapping rules to which Millikan refers are functions, not in the teleological, but rather the

mathematical sense. Each mapping rule takes as its domain the objects and relations of the isomorphs produced by the intender, and has in its range the objects and relations that make up its targets on occasions of appropriate use. The semantic content of a representation is then determined by applying the mapping rule of the intender to the isomorph it produces. The mapping rule takes us from the isomorph to its content (which will be its target on occasions of proper functioning.)

As we saw in the previous section, this solves the problem of underdetermination. While any given isomorph is isomorphic to lots of possible content structures, the application of the intender's mapping rule to that isomorph will take us to a unique content structure. We might say that the isomorph has the potential to represent all of the structures to which it is isomorphic, but that it will not actually be a representation, according to Millikan, unless it is used by some intender. This use can be characterized by applying the intender's mapping rule to the isomorph. The same isomorph, employed differently, by an intender with a different mapping rule, would be a different representation with a different unique content.

Though many use theorists do not speak explicitly of mapping rules, any use theory can be construed as specifying one. For example, Fodor's causal theory can be construed as the claim that the mapping rule which takes mental representations employed in detection to their causes under optimal detection circumstances is the rule which assigns content to those representations whenever they are used in any circumstance.

However, we have already seen Cummins' objection to this sort of strategy. By making the content of a representation depend, in part, on a mapping rule which describes the ideal functioning of the intender that produces it, the claim that the cognitive system has employed that representation will lose its explanatory force. Any capacity of the system that we might have hoped to explain in this way will instead be taken to be constitutive of the system's having the representation in the first place.

A dilemma is threatening. Cummins' view, according to which representations just are isomorphs, solves the problem of explanatory vacuity by making representation independent of use, but falls prey to the problem of underdetermination. According to the use theorist,

however, the problem of underdetermination is solved by forcing each isomorph that is to be a representation to be associated with the particular mapping rule that characterizes proper functioning of the intender employing it. But the use theorist's appeal to such a mapping rule prevents her from avoiding the problem of explanatory vacuity. If it were really true that any theory of mental representations must take one of these two forms, it looks as though we would somehow have to learn to live with one or the other of these two problems. If we could not learn to live with either, this would perhaps be a reason to reject scientific use of the notion of representation altogether.

CHAPTER 2

THE CONCEPTUAL FOUNDATIONS OF COGNITIVE SCIENCE

Representations as Structure/Mapping Rule Pairs

(A Solution to Both Problems)

What fuels the dilemma at the end of the previous chapter is the fact that the very mapping rules use theorists use to solve the problem of underdetermination appear to cause the problem of explanatory vacuity which motivates Cummins. As he says,

> I rejected relativization [to a mapping rule] in *RTA*...because I think the only way to make sense of the idea that there is a "correct mapping rule" is the way Millikan makes sense of it: The map producer is functioning Properly when it constructs maps that are isomorphic to its target structures according to that rule...This...makes representational content a function of use, and that in turn compromises the notion of representational error (Cummins, 2000, p. 115).

In other words, Cummins doesn't dispute that it would be nice to be able to appeal to a mapping rule. The problem is simply that he thinks any such appeal generates even greater difficulties, and so he proposes to do without.

However, despite Cummins' protestations to the contrary, I see no reason why we cannot specify a mapping

rule in some other way that does not make content a function of use. After all, every function which takes us from some structure to some other structure to which it is isomorphic is such a mapping rule. Furthermore, for any pair of isomorphic structures, the very fact that they are isomorphic entails that there exists a mapping rule that takes the first to the second (and another that takes the second to the first). Though some such mapping rules may characterize the proper functioning of intenders, we need not specify the mapping rules of representations only in this way. We can specify them in whatever way we like. So long as our method of specification for the mapping rules does not make any essential reference to use, the problem of explanatory vacuity will be avoided.

So, what I propose is that we decouple the mapping rules from intenders and instead make them components of the representations themselves. According to this proposal, rather than simply being a structure, a representation is an ordered pair composed of a structure (which I will follow Millikan in calling an "isomorph") and any mapping rule. The content of a representation is the structure one arrives at if one applies its mapping rule to its isomorph. Unlike Millikan's mapping rules, which are

always determined by the proper function of the intender that employs the representation, these mapping rules can be any mapping rule that takes one structure to another structure to which it is isomorphic. Every isomorph/mapping rule pair constitutes a representation, regardless of whether that pair is or is not used by any cognitive system.

This proposal avoids the problem of underdetermination in the same way that Millikan's appeal to mapping rules does. Cummins claims that a representation is a structure S which is isomorphic to its content *c*. This generates the problem that too many structures will count as S's *c*. But according to my theory, a representation is an ordered pair of an isomorph and a mapping rule ⟨S, M⟩ such that S is isomorphic to its content *c* according to M. Since the application of M to S yields a unique structure, the representation ⟨S, M⟩ has *c* as its unique content.[3]

---

[3]Actually, strictly speaking we only want representations of individuals to pick out a unique structure as their content. Representations of properties need instead to pick out a set of content structures. This modification is easy enough to introduce by pairing the representational structure with a set of mapping rules, such that each pairing of the representational structure with one of the mapping rules picks out one of the content structures in the property set. For the sake of

However, although my theory avoids the problem of underdetermination by appealing to mapping rules, it is not a use theory. The question of whether or not something is a representation simply becomes a question of whether or not that thing is a set-theoretic object of the right sort, namely an ordered pair of a structure and mapping rule. Any such pair, used or not, used this way or that way, picks out another structure, which is its content. So the proposed theory is a modification of Cummins' intrinsic theory of representational content, which is designed to fit into the same framework he uses to account for misrepresentation and avoid the problem of explanatory vacuity. My overall theory differs from Cummins' overall theory only with respect to the altered theory of representational content.

(1*) A mental representation is an ordered pair ⟨S, M⟩, where S is a structure and M is an isomorphic mapping rule from S to another structure, *c*, which is the content of the representation.

(2) A mental representation is targeted on whatever structure, *t*, the intender that employs it has the function of representing.

---

simplicity, I will often discuss only representations of individuals in what follows, but it should always be possible to apply this modification in order to extend the discussion to representations of properties.

(3) A propositional attitude's meaning is determined by
    (a) the attitude type (i.e. belief, desire, etc.), and
    (b) the content that *c* = *t*.

For example let's suppose that, in forming a belief about the Andromeda galaxy, I correctly employ a representation of the spiral structure of that galaxy. According to the proposed theory, this means that there is some structure, presumably in my brain, S_B, which is isomorphic according to some mapping rule, M_1, to the spiral structure of Andromeda, S_A. It further means that this representation $\langle S_B, M_1 \rangle$ has been employed by an intender which has S_A as its target.

However, it may be equally true that the very same structure, S_B, when paired with a different mapping rule, M_2, picks out as its content the spiral structure of some different but isomorphic structure, say the spiral structure of the Milky Way, S_{MW}. This pair, $\langle S_B, M_2 \rangle$, is just as much a representation as the one that I actually use, and if I were to use this representation, rather than the previous one, while still attempting to represent Andromeda, I would misrepresent my target.

For that matter, since the two content structures of these representations, S_A and S_{MW}, are both isomorphic to S_B, it follows by transitivity that they are isomorphic to each other, and hence that there is a mapping rule, M_3, that takes the first to the second. It also follows that the inverse of this function, M_3^{-1}, takes the second to the first. Thus, we have two more representations, $\langle S_A, M_3 \rangle$, in which Andromeda represents the Milky Way, and $\langle S_{MW}, M_3^{-1} \rangle$, in which the Milky Way represents Andromeda.

So the proposed theory solves the problem of explanatory vacuity in the same way that Cummins' theory does. A theory of representation is supposed to allow us to explain the usefulness of representations by appealing to the use-independent fact that there are some things that represent others. In the example above, $\langle S_B, M_1 \rangle$, $\langle S_B, M_2 \rangle$, $\langle S_A, M_3 \rangle$, and $\langle S_{MW}, M_3^{-1} \rangle$ are just four such representations. According to the proposal, the universe comes fully stocked with representations, waiting to be employed. It is then the business of cognitive science to try to figure out, by determining the functions of intenders, which of these representations cognitive systems employ (or misemploy). Once this is determined, the fact that some cognitive

system employs one of these intrinsically defined representations can play a non-vacuous role in the explanation of the system's behavior.

The scientific task of determining which representations have been applied by which cognitive systems to which targets will be a complex one. For example, I described the case above as a case of my correctly using $\langle S_B, M_1 \rangle$ to represent Andromeda, but in order to determine that this is in fact what I have done, the cognitive scientists studying me will have to appeal to some theory or other that tells them a) that I have in fact employed that representation, rather than some other representation, and b) that my target was in fact Andromeda, rather than some other target. The theory of representational content that I am proposing doesn't answer these questions, and the point of the problem of explanatory vacuity is that it isn't supposed to. A theory of representational content is supposed to define the notion of representation so that it can be meaningfully employed by cognitive scientists. It is then up to the scientists to decide, on empirical and theoretical grounds, how it should be employed in the explanation of behavior.

For example, according to the theory I have provided above, if I am trying to represent Andromeda (i.e. Andromeda is my target), it would be just as correct of me to apply the representation $\langle S_{MW}, M_3^{-1} \rangle$ to the target as it would be to apply $\langle S_B, M_1 \rangle$, as both have the spiral structure of Andromeda as their content. It is, of course, preposterous to think that I use the former rather than the latter representation in correctly representing Andromeda, but the point is that it is only preposterous against a certain background of scientific knowledge. The fact remains that if I could somehow apply representations with gigantic isomorphs, like the Milky Way galaxy, to the targets of my intenders, I could represent or misrepresent with them. The reason we think this is impossible is, in the terms of the proposed theory, roughly that we accept as a basic constraint on any acceptable theory of representational use that cognitive systems cannot use representations with isomorphs instantiated well outside those systems. Structures the size of the entire Milky Way galaxy just don't plausibly have any sort of causal effect on my behavior.

Determining that I am targeted on Andromeda, rather than the Milky Way or my toaster, is similarly a matter for scientific determination. According to the theory, this determination is to be made by determining the function of the intender system which employs the representation. As Cummins points out in *RTA*, use theories of representational content may be able to be usefully reinterpreted as theories of target fixation, since the same sorts of considerations those theories use (inappropriately) to fix representational contents seem like just the sort of considerations (i.e. selection history, causal covariance, etc.) that cognitive scientists will want to use in determining the functions, and hence the targets, of intenders (Cummins, 1996, p. 120). I suspect that this may be right. However, I intend to discuss these matters in more detail in the next chapter.

Pictures

Before I move on to consider an assortment of possible objections to the theory of representational content I have proposed, I would like to clear up a potential misconception about the role of isomorphism in the theory.

Cummins calls his own theory of isomorphic representation a version of the "picture theory" of representation. Both his theory and my own, in claiming that representations (or in my case, isomorphs, the structural components of representations) are isomorphic to their contents, assert a certain analogy between mental representations and conventional pictures, like photographs, paintings, and so on. The analogy is that in both cases, the representation is, at least in certain respects, similar to the content. However, this analogy can be improperly read to imply that mental representations must be similar to their contents in ways that are themselves similar to the ways in which pictures are similar to their subjects.

So, for example, while the spatial relations between elements in a photograph correspond in a way defined by the laws of perspective to the spatial elements of the photograph's subject, there needn't be any significance at all to the spatial arrangement of the isomorph of a mental (or any other non-pictorial) representation. If the content structure includes spatial relations, these could in principle be mirrored in the isomorph by temporal relations, relations among neural activation thresholds, or whatever. For that matter, since representations need not

be used in order to be representations, they could be mirrored in the heights of waves on a pond, or in temperature gradients in the corona of the sun.

The overstated analogy also puts unnecessary constraints on the sorts of structures that can be content structures, since they must be of the sort that can be represented in pictures. The point is that insisting that the isomorphs of representations be isomorphic to their contents is a much weaker constraint on the ways in which isomorphs can be instantiated than would be insisting that isomorphs be pictures, in the conventional sense, of their contents.

This is as it should be, since we do not want our concept of representation to unnecessarily constrain the work of cognitive scientists. For example, if we were to hold cognitive science to the really "pictury" picture theory that comes from overstating the picture analogy, the content of my mental representation of my cat would, I suppose, have to be something like a finite set of points coincident with my cat's surface, along with a set of distance relations holding between each pair of those points. (You can picture this structure as similar to a wire frame, like one over which one might drape *papier-*

mâché strips to create a statue of the cat). And the isomorph of my representation would have to be a similar sort of structure in my brain. But this is implausible. Though I may use similar isomorphs to represent certain sorts of contents, my cat probably isn't one of them. For one thing, whereas a structure consisting of a set of points and a set of distance relations is completely static, my cat is not. The described structure might occur usefully in a representation of a cat statue, or a dead and stuffed cat, but not in one which can assume multiple shapes. As the living cat moves, the distance relations between whatever surface points we have chosen will not remain constant.

We might represent a living cat as a four-dimensional structure of space-time points with spatio-temporal interval relations (the four-dimensional analogue of distance relations) defined over them, but the isomorph I use to represent my cat probably isn't like this either. I do not, thankfully, know or even have detailed beliefs about all the places my cat has been or all the motions and movements she has made, so my collection of space-time points will be patchy. Furthermore, I do mentally represent my cat as having a variety of properties beyond

the surface shape of her four-dimensional “time worm.” For example, I represent her as having kidneys.

In fact, I really don’t have any idea exactly what mental structures I might use when I mentally represent my cat. However, if we consider a couple of extreme cases, I think we can get an idea of the potential flexibility of isomorphic representation. First consider a narrow-minded geneticist. He is narrow-minded in the sense that, to him, a cat is simply an organism that has a certain type of genome. Naturally, he mentally represents cats with a structure that is isomorphic to the genetic structure of cats. This is not to say that he uses the same mental representation to represent cats that he uses to represent cat genomes. That would be true only if he thought that cats just were genomes. Rather, it is just that, though the geneticist distinguishes between cats and their genomes, his detailed beliefs about cats are limited to their genomes. So when he mentally represents a cat, he will employ an isomorph which is richly isomorphic to the tiny structures instantiated in cat DNA, but the isomorph he employs will not be richly isomorphic to the phenotypic structure of cats – claws, pointy ears, purrs, and so on.

For this geneticist, a cat is just a vaguely defined something that instantiates the cat genome.

On the other hand, consider a narrow-minded evolutionary biologist. She thinks of cats solely in terms of their relation to ancestor and descendant species – as nodes in the "tree of life." Consequently, we might expect that her mental representation of cats contains an isomorph that is richly isomorphic to the ancestor/descendant structure into which cats fit, but not richly isomorphic to any structure instantiated in individual cats themselves. Just as was the case with the geneticist, her mental representation of cats is not identical to her mental representation of the ancestor/descendant structure, but it is true that the isomorph employed in her mental representation of the ancestor/descendent structure is the only detailed substructure of the isomorph she employs to mentally represent cats. Thus, in this case, the content structure picked out by the mental representation of cats is not instantiated in cats at all. It is a gigantic structure, reaching millions of years into the past and spanning the entire planet (i.e. all the times and places where there are or have been cats and/or cat ancestors or descendants.) It might seem a bit strange that the content

structure of her representation of cathood is not instantiated in individual cats, but then, in a sense, anyone who thinks of cats solely in terms of evolutionary theory doesn't really think that being a cat supervenes on individual biological organisms. A purported cat without ancestors is not, according to this conception, really a cat.

So far, I have used as examples mental representations of cats employing isomorphs which are isomorphic to cat surfaces, cat anatomy, cat genomes, and the evolutionary history of cats. Rather than employing any of these mental representations, I suspect that most reasonably educated people employ a mental representation that picks out a content structure composed of all these sorts of structures. Unless our geneticist has somehow never seen a cat, he will employ an isomorph that is isomorphic to at least some phenotypic structures of cats, and unless a cat sculptor is completely unaware that her models are living, non-hollow creatures, she will employ an isomorph which is at least grossly isomorphic to their anatomy. Nevertheless, different people will probably all employ slightly different mental representations of cats. Unlike

the narrowly conceived picture theory, isomorphic representations can accommodate all of these cases.

A Survey of Anticipated Objections

Pansemanticism

The fact that representations are so ubiquitous can easily lead to the misconception that my view implies a radical pansemanticism. The same charge might be leveled at Cummins. While this is a misconception, it is worth investigating since it will illuminate an important feature of both our views. Let's consider how this misconception might arise for Cummins first. In the previous chapter's discussion of the problem of underdetermination, we noted Cummins' tendency to refer to representations and contents as if they were the objects instantiating representational and content structures, rather than as the structures themselves (for example, when he says that the Autobot card represents a path through a maze, rather than that the structure of the notched groove in the card represents the path). We also noted that if this had been Cummins' actual view, his underdetermination problem would have been exacerbated, since it will nearly always be possible to

find some pair of isomorphic structures instantiated in any pair of objects. In fact, if we allow trivial mappings between degenerate structures, then, since there exists a mapping rule which takes any degenerate structure to any other, and since each object, when paired with an empty set of relations, counts as a degenerate structure, laxness about the above distinction will yield a theory according to which every object represents every other.

As we said, though, this is just a misconception. Neither on my own view nor on Cummins' view are non-set-theoretic objects ever representations or contents. For Cummins, representations are structures, and hence different structures instantiated in the same physical object are different representations. This deals with the problem of each object's instantiation of multiple structures, but leaves the problem of multiple isomorphic mappings untouched. On my view, representations are ordered structure/mapping rule pairs, so different such pairs are different representations. This view deals with both problems, leaving each representation with a unique content even in the case of trivial mappings, since the mapping rule component of the representation will take even

a degenerate structure to a unique degenerate content structure.

Nevertheless, there is still a sense in which representations are ubiquitous on my view. Since every object instantiates a degenerate structure isomorphic to every other degenerate structure instantiated in any other object, and since trivial mappings hold between any two of these degenerate structures, it will be true that for any two objects A and B, A instantiates at least one structure, S, isomorphic to at least one structure, *c*, instantiated in B by a mapping rule, M. It follows that ⟨S, M⟩ represents *c*, and thus that some structure in A, when paired with M represents a structure in B. So while it is false that every object represents every other, it is true that every object instantiates a structure which, when paired to an appropriate mapping rule, represents some structure instantiated in the other. This is true despite the falsity of the claim that "everything represents everything else."

The Spookiness of Intrinsic Theories

A second misconception worth dispelling is that any rejection of the use thesis and acceptance of an intrinsic theory of representational content commits one to a spooky or mysterious theory of intentionality. Defenders of the use thesis often write and speak as though it does, and they do have a point. There seems to be a significant difference between the case where a shipwrecked sailor is rescued after he writes "HELP" in seashells on the beach, and the case where he is rescued when the seashells just happen to wash up on the beach in a similar configuration. And indeed, the difference seems to be that the sailor is using the first arrangement of seashells in a way that he is not using the second arrangement. As Wittgenstein puts it

> The mistake we are liable to make could be expressed thus: We are looking for the use of the sign, but we look for it as though it were an object *co-existing* with the sign. (One of the reasons for this mistake is again that we are looking for a 'thing corresponding to a substantive.')...One is tempted to imagine that which gives the sentence life as something in an occult sphere, accompanying the sentence (Wittgenstein, 1994b, p. 61).

This notion is further supported by one of the staunchest defenders of intrinsic intentionality, John

Searle. In discussing artificial intelligence and his famous Chinese Room thought experiment, he says,

> It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality...but as far as we know it is because I am a certain sort of organism with a certain biological (i.e. chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing...intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality (Searle, 1991, p. 516).

This does make it sound as though Searle thinks that if we just uncover the right rock, peer behind the right subatomic particle, or just look at a brain state really, really closely, we might find the special extra ingredient that makes some physical objects intrinsically representational. Furthermore, if we were to doubt that any such special ingredient were scientifically discoverable, it looks as though we would have to accept that it existed “in an occult sphere.” If intrinsic intentionality really required either a bet at long odds on future scientific discoveries or acceptance of non-physical properties, one might well be tempted to a use theory, where the special extra ingredient is not a property of the

representation, but rather the way the representation is used.

However, my theory (and Cummins' theory too, for that matter) is a non-spooky intrinsic theory. It is true enough that one can stare at brain states all day, in as much detail as one likes, without ever observing the special ingredient necessary to make them intrinsically representational. But this is because nothing could ever make brain states representations, since brain states are not set-theoretic objects of the right sort. When we are faced with a brain state, we must identify a structure instantiated in the state, and identify a mapping rule with which to pair it. These are the missing special ingredients that prevent us from immediately identifying objects which play a representational role, and the representations constituted from them which play that role. And yet, there is no more mystery to the existence of these set-theoretic objects than there is to the existence of the set of the change on my dresser. If the change exists, so does the set.

Ontology

Another likely source of objections to the proposed theory is the oddity of its objects. However, odd as they may be, they do not seem to me to pose any insurmountable ontological difficulties. Many writers on this topic, including both Cummins and Millikan, already appear to be committed to the existence of structures, which we said were ordered pairs of the form ⟨O, R⟩, where O is a set of objects, and R is a set of relations, each of which is defined over the objects in O and/or other elements of R. As we have seen, Cummins just takes these structures to be representations, while Millikan takes those structures (or isomorphs) which are used in the right way to be representations. Either way, representations are structures.

In his reply to Millikan's review of *RTA*, Cummins states explicitly that he is untroubled by an ontological commitment to structures.

> I am a realist about structures, meaning I don't think they are in the eye of the beholder. Space-time, for example, has an affine structure, and its having that structure is an objective fact about space-time...I don't propose to argue for this sort of realism beyond noting that, if you are not a realist about structure in this sense, then you ought to think most science

> worthy of the name is trivial, a major strategy since Galileo having been to get a grip on the universe by finding mathematical structures that are isomorphic to the structures of nature (Cummins, 2000, p. 113-114).

He further goes on to note, correctly I think, that Millikan endorses something similar, in that

> Although she sometimes [i.e. in her discussion of the problem of underdetermination] writes as if she thinks anything goes, her worry is not that you can find any old structure you want anywhere you want to find it, but that there are lots of different structures in the same thing (Cummins, 2000, p. 114).

Now what I have proposed is that representations are slightly more complex set-theoretic objects of the form ⟨⟨O, R⟩, M⟩, where M is a one-to-one, onto function from the isomorph ⟨O, R⟩ to some other structure, *c*, which is the representation's content. If anything makes my representations look odder than Cummins' or Millikan's, it is the incorporation of the mapping rule as a constituent of the representation. However, the existence of the mapping rule is entailed by the fact that the isomorph (which Cummins takes to be a representation and which Millikan takes to be a representation only if it is used in a certain way) is isomorphic to its content. In fact, the claim that the isomorph is isomorphic to the content

structure is identical to the claim that there exists such a one-to-one, onto mapping function between the two structures. Since both Cummins and Millikan agree that this isomorphic relationship holds between certain pairs of structures, both are committed to the existence of mapping rules. Since both are also committed to the existence of the structures/isomorphs as well, I would expect little objection from them, at least, to the existence of the ordered pair of the two.

Of course, some might object to any and all such realist claims on behalf of set-theoretic objects, or perhaps to the identification of representations with such abstract objects. However, it is worth noting that even if realism about the objects I take representations to be turns out to be untenable, the difficulty this poses for my theory is less significant than the problems of underdetermination and explanatory vacuity which afflict Cummins and the use theorists respectively.

First, suppose that the theory I have proposed is as good a theory of representation as we are going to get. If we decide that we cannot accept the existence of representations as proposed, we could still accept these objects in an anti-realist sense. This might be seen as a

retreat of sorts, but it is hardly unprecedented. For many of the most successful theories in science, both historical and current, whether or not the entities described by those theories can be accepted realistically is at least debatable. If the proposed theory ends up in this category, it will at least be in good company.

On the other hand, suppose that either Cummins' theory or some use theory is a better theory. In either of these cases, unless the respective problems faced by these theories can be somehow avoided, the scientific employment of the concept of representation would have to be abandoned altogether. If we accept a use theory, we have to accept that it is impossible to non-vacuously explain behavior by reference to representations, which seems to be the primary scientific purpose that appeals to such representations have. If we accept Cummins' theory, we must accept that representational content is seriously underdetermined, and that it consequently plays only a tiny role in the explanation of behavior. Either way, scientists seeking to understand behavior would be well advised to seek some different sort of theory. The point is that, in science, ontological failures, while perhaps not inconsequential, are preferable to explanatory failures.

Furthermore, since the ontological status of my representations depends on nothing more than the ontological status of the objects of set theory, and since set theory seems to be indispensable, anyone who wishes to reject the existence of this ontology is going to have to find some sort of substitute. So long as that substitute is an adequate substitute for talk about sets, structures, functions, and ordered pairs, my theory of representation should be able to survive the translation.

The Redundancy of Isomorphism

Even if one does not object to the existence of these objects, one might deny that I am right to call them representations. One such objection might question the relevance of isomorphism once mapping rules have been introduced. I have claimed above that even conventional representation can be accommodated within my definition of representation because of the possibility of trivial mappings between degenerate structures. Since the isomorphs of such representations are degenerate structures, all of the work of picking out a content has to be shouldered by their mapping rules. But, one might wonder, if mapping rules are capable of shouldering the

entire burden, why bother introducing isomorphism as an essential component of the representation? Why not instead simply claim that a representation is an ordered pair of some object and a mapping function that takes that object to its content object? The objector might concede that these representational objects might sometimes happen to be structured similarly to their targets, but he sees this as inessential to the notion of representation.

Admittedly, my retention of isomorphism is to a certain extent the result of the fact that my own view grew out of a reflection on the views of Cummins and Millikan, both of whom incorporate isomorphism. This reflection convinced me that the mapping rule component was necessary, but could not be defined in terms of use. So the mapping rules became components of the set-theoretic objects I call representations by being paired with the isomorphs that Cummins and Millikan call representations. Since the addition of the mapping rules allowed me to accommodate conventional representation by shifting the entire representational burden onto a mapping rule paired to a degenerate structure, there was no good reason to take out the isomorphism requirement, and thus it remained.

This makes the role of isomorphism in my theory seem a bit like the human appendix, but I don't think it is entirely vestigial. For, so long as there is the possibility of non-trivial isomorphism between mental structures and the environmental features they represent, an analysis of representation that does not essentially include isomorphism will need to deal with this structural similarity in some other way. Suppose we have a case where, as I would describe it, a non-degenerate content structure $\langle O_c, R_c \rangle$ is represented by a representation consisting of a non-degenerate isomorph $\langle O_i, R_i \rangle$ and a mapping rule M. If we were to adopt an analysis of representation that forbade this description by replacing isomorphs with unstructured objects, $\langle O_c, R_c \rangle$ would have to be represented by individually representing all of the members of O_c.

So far as I can tell, there isn't anything wrong with this way of describing the case. Indeed, if one likes, one can describe it in just this way within my analysis by individually representing all the members of O_c with degenerate-isomorph representations. But the reverse is not true. An analysis of representation that lacks the

isomorphism component can only describe the case as one in which a non-degenerate content structure is represented by individually representing its components. So, my analysis is preferable because it is general in a way that an alternative analysis which ignores isomorphism is not. It might turn out that cognitive scientists decide to refer only to representations with trivial mappings in their cognitive explanations. But I see no good philosophical reason to define the notion of representation in such a way that reference to isomorphism is forbidden from the outset.

A Boring Solution to a Profound Problem

Another sort of objection to my analysis of representation arises from the suspicion that, in incorporating mapping rules into the representation, I have simply defined away a deep mystery about intentionality. After all, isn't it surpassing strange that there should be some things that are about others? What sort of property could this "aboutness" be?

When professors first introduce the idea of intentionality to their students, they tell the students that a representation is something that "points" to something else, as if there were an arrow drawn from the

representation to its content. Only, of course, it can't really be a matter of having an arrow. First, there aren't really any arrows pointing from representations to their contents: not even invisible ones, or ones made of streams of special sub-atomic particles or whatever. Second, even if there were, say, giant black sticks with pointy ends stuck between representations and their contents, there isn't really anything intrinsically intentional about pointy black sticks. Arrows are themselves just conventionally adopted representations of the true relation of intentionality, whatever that may be. As was mentioned above, this tends to lead philosophers either to embrace a use theory or to turn to something spooky as an account of the rare and mysterious property represented by the arrow.

However, on my analysis, the fact that something is a representation is really pretty boring. Mapping rules, in virtue of being functions in the mathematical sense, have an inherent directionality. When they are paired to isomorphs (even degenerate ones), these rules "point to" a particular content structure. So the arrows are just mapping rules. And just as the lecture in Intentionality 101 would have it, a representation is just what you get when you take something (i.e. a structure) and stick an

arrow (i.e. a mapping rule) on it. As mentioned above, this means that, while nothing that lacks a mapping rule (and hence nothing that is not a set-theoretic object) is a representation, every structure can be paired with a plethora of mapping rules to create a plethora of representations. This tends to conflict with the intuition that, whatever representation amounts to, it has to be something special.

I don't deny the appropriateness of this sense of mystery, but I do think that it has been misplaced. For, while my analysis makes being a representation a boring and trivial consequence of structural similarity wherever it may occur, determining the employment of such representations by cognitive systems remains an exciting challenge for cognitive science. In place of the mysterious question "How could some things be about other things?" my view generates the question "How is it that cognitive systems are able to exploit the common but rather abstract fact that certain of their states are the isomorphs of representations, and why do they exploit those representations, rather than all the others?" These questions are addressed to scientists, who are, after all, in the mystery-resolving business.

Representations with Nonexistent Contents

Nevertheless, even if everyone can be convinced that I've picked the right set-theoretic objects to call representations, and that these objects exist, one still might worry that cognitive scientists will be unable to complete their task without appealing to the use by cognitive systems of at least some representations that fail to exist. According to the theory I have proposed, in order for a representation to exist, both its isomorph and its mapping rule must be instantiated in some existing object or state of affairs. Since cognitive scientists are unlikely to propose that a cognitive system employs any representation with an isomorph that is not part of the system, the instantiation of isomorphs should not pose a problem.

However, mapping rules may appear to be another matter. Since mapping rules are functions that take objects and relations in the isomorph to objects and relations in the content structure, they are instantiated, as it were, at both ends. One end of the mapping rule is instantiated where the isomorph is instantiated (i.e. in a brain), while the other end is instantiated in whatever object or state of affairs instantiates the content

structure. This means that I am committed to the claim that there are no representations with nonexistent contents.

So long as cognitive scientists never need to appeal to the use of representations with nonexistent contents, this won't be a problem. Unfortunately, though, it looks at first glance as though scientists will quite frequently be tempted to appeal to just such representations. If, for example, they are studying a person who believes in unicorns, it seems natural for them to appeal to that person's use of a representation of unicorns in the formation of that belief. But since there are no unicorns, it doesn't seem that there will be any representations of unicorns which the scientists can claim the person has used. If the theory of representations that I have proposed is going to work, there had better be a way for these scientists to avoid attributing the use of representations with non-existent contents.

Consider the following case in which someone forms a belief about a unicorn. Bob is looking over in the direction of a horse standing ten meters away from him. Standing eleven meters away and directly behind the horse is an antelope. Given the relative positions of the two

animals, every part of the antelope is concealed from Bob's view except for one of its horns. The antelope is holding its head in such a way that its second horn is hidden from Bob by the first, and in such a way that the visible horn is lined up with and roughly perpendicular to the horse's brow ridge. In other words, it looks to Bob as though the antelope's horn is sticking out of the horse's head, and consequently, Bob forms the belief that there is a unicorn standing ten meters away.

I have tried to set this case up in such a way that it makes sense to treat Bob's belief as resulting from a misrepresentation of the antelope's horn as belonging to the horse. According to the analysis of misrepresentation that I have inherited from Cummins, misrepresentation is the result of applying a representation with one content to a non-identical target. If the representation's content fails to match the target, the belief will then turn out to be false, since the content of the belief, *qua* propositional attitude, is that the representation's content does match the target. So, in order to find a way around the objection, we need to figure out what the target is, and then find a fully instantiated representation that

Bob (or one of Bob's intenders) has misapplied to that target.

The target is whatever it is that Bob is trying to represent, which is determined by the function of the intender that is applying the representation. Cummins has pointed out in *RTA* that many, perhaps most, intenders are indexical (Cummins, 1996, p. 118-119). In the current case, Bob is attempting to represent what he sees before him. So, the intender that is applying the representation is (perhaps some sub-system of) his visual system, and its target is determined, basically, by where he is currently focusing his eyes. That is, it is the job of the visual system to represent whatever visually observable structures occur at whatever location the eyes are currently focused upon. As I've described the case, Bob's eyes are focused on the region around the horse, which is to say, a region about ten meters away. The only visible thing in this region is the horse, so our cognitive scientists will have to identify some visible structure instantiated in the horse as the target.

As for Bob's application of representations to that target, the description of the case suggests that Bob has employed two representations, one of the horse and one of

the antelope's horn, and done so in such a way that his mistake is in somehow putting these two representations together incorrectly. In particular, he puts them together in such a way that he misrepresents the horn as being one meter closer than it really is, and therefore, as being right where the horse's head is. So we'd like to say that Bob's representation of the unicorn he believes to be ten meters away is some sort of composite of a representation of the horn and a representation of the horse.

We can construct such a composite in the following way. Let $\langle S_U, M_U \rangle$ be Bob's representation of the unicorn, $\langle S_H, M_H \rangle$ be Bob's representation of the horse, and $\langle S_A, M_A \rangle$ be his representation of the antelope horn. Then the isomorph of Bob's representation of the unicorn, S_U, can just be the set-theoretic union of the isomorphs of the horse and antelope representations, $S_H \cup S_A$. The mapping rule for the unicorn representation, M_U, can then be any mapping rule which coincides with M_H with respect to its treatment of the S_H parts of S_U, and coincides with M_A with respect to its treatment of the S_A parts of S_U. The content of this unicorn representation will then be the union of the content structure picked out by the horse representation

and the content structure picked out by the horn representation. Thus, the content structure of the unicorn representation will be entirely instantiated in existing physical objects. The horse part will be instantiated ten meters away in the horse and the horn part will be instantiated eleven meters away in the antelope's horn. Furthermore, when this representation is applied to the target (which is whatever visual structures are ten meters away) its content will fail to match, because the entire target structure is ten meters away, while part of the content structure is not.

Now one might object at this point that this is just not the right content. When Bob forms his belief that there is a unicorn standing ten meters away, it is natural to imagine that this belief of Bob's will, for example, be accompanied by a mental image. Surely, this image will not be an image of what I have claimed is the content of Bob's representation of the unicorn. If it were, the image would be of a horse ten meters away and a horn bobbing in thin air another meter behind it! Whatever that is an image of, it isn't an image of a unicorn.

This is exactly right, but not a criticism of my theory of representational content. The temptation to

think that it is comes from a failure to distinguish, as Cummins and I do, between representational content and propositional attitude content. It is completely correct to say that the content of Bob's belief (or his accompanying mental image) is not the strange scattered object in which the content structure of his representation is instantiated. Rather, as with all propositional attitude contents, the content of Bob's belief is that the content of the representation applied to the target is identical to that target. Thus, Bob believes (falsely) that the entire content of the representation coincides perfectly with the target, and his mental image is of this (non-obtaining) state of affairs.

The resolution of the problem posed by this particular example suggests a general strategy that cognitive scientists can employ to avoid appeals to representations with non-existent contents. The trick is to analyze the non-existent content into a set of existing contents, find fully instantiated, existing representations which pick out each of these contents, use these individual representations to construct a composite representation, and then claim that the cognitive system has (mis)applied the composite representation to its target.

Representations with Missing Targets

In addition to cases of representations with missing contents, it will also be possible to have cases with missing targets. For example, suppose that Irving is visually hallucinating an elephant in his empty living room. This means that his visual system (the relevant intender) has applied a representation of an elephant to its target, which, since the room is empty, appears to be nothing.

Unlike missing contents, missing targets do not threaten to pose any ontological difficulties for the proposed theory of representational content, since no part of a representation is necessarily instantiated in its target (though it will happen to be partially instantiated in its target on occasions of correct application, since then the target will be identical to the content.) However, one might worry about whether any actually existing representation will be the right representation to employ when an intender is targeted on nothing.

Irving's hallucination will clearly count as a case of misrepresentation because the content of the representation is an elephant, whereas the target is nothing, and an elephant isn't nothing. But what if we ask what

representation would count as the correct representation - a representation of nothing? According to the proposed theory of representational content, it doesn't seem that there could be such a representation, since there is no structure/mapping rule pair that picks out nothing. Ironically, Cummins and I objected to use theories on the grounds that they cannot account for the possibility of misrepresentation, but now it looks as though my proposed theory of representational content might falter because it allows a case in which correct representation is impossible.

However, I think a little reflection will show that this worry rests on a rather old confusion. Nothing is not something, and hence needn't be represented. When a properly functioning intender is "focused on nothing," it is more perspicuous to say that it lacks a target than that its target is nothing. Furthermore, that properly functioning intender will not respond by employing a representation of nothing, but rather by refraining from applying any representation at all to its non-target. An intender that is not targeted on anything commits an error if it employs any representation at all.

Representations and Concepts

The final worry I will consider in this section concerns the relation between my theory of representational content and a theory of concepts. One might grant that my representations do in fact represent the contents I claim they do, but deny that application of such representations can account for a cognitive system's possession of concepts. To appeal to a hackneyed example, if a tennis ball rolls into view, I have suggested that my representation of that ball will be some mentally instantiated, perhaps spherically structured, isomorph paired with a mapping rule that maps the isomorph to the ball. It is supposed to be the fact that this complex set-theoretic object naturally "points to" a structure instantiated in the ball that makes the ball (or more properly the ball's spherical structure) the content of the representation. But even if this is right, it doesn't seem to go any distance towards accounting for the fact that the ball is represented "as" a tennis ball, or in other words, to account for my having the concept "tennis ball." After all, if I have a friend who comes from a tennis-lacking society, she might employ precisely the same

representational type to, say, recognize the presence of the ball, without representing it "as" a tennis ball.

However, I believe this objection has been adequately addressed by Cummins (Cummins, 1996, p. 86-90). Though I disagree with Cummins regarding the correct account of representational content (by endorsing (1*) rather than (1) above), we both accept the same account of propositional attitude formation ((2) and (3) above). Consequently, I more or less share Cummins' reply to this objection, since that reply is essentially that "representation as," or conceptualization, is determined by a system's propositional attitudes rather than its representations.

Once again, the appropriate reply turns on the fact that the objection fails to account for the distinction between representational content and propositional attitude content. It is true enough that in order to have a concept, in addition to having a representation, a cognitive system also needs a set of background beliefs about that which has been represented. What I have that my friend lacks is a bunch of beliefs about tennis: particularly beliefs related to the proper size, shape, consistency, elasticity, fuzziness, etc., that balls must have if they are to serve as tennis balls. So having a

concept is a matter of having certain propositional attitudes. Now if one conflates representational content with propositional attitude content, one is liable to conclude that, in order for me to have the concept of a tennis ball while my friend lacks it, I must also possess some representation that she lacks.

But this does not follow. We have already seen that one can have a propositional attitude about something without having a representation of that thing (for example, the mouse that forms its belief that there is a bird in the sky by having a PLACES-THERE-ARE-BIRDS-intender that applies a |sky| to its target). The same thing goes for the background beliefs necessary to the possession of a concept. My possession of the concept “tennis ball” is a complex structure of beliefs about the game of tennis and the balls one uses in that game. Each of these various beliefs will be formed by intenders applying various representations to various targets, but my possession of the concept does not directly entail my possession or lack of any particular representation. I simply need to have applied some representations to a bunch of appropriate targets in such a way that I form the background beliefs necessary to possession of the concept.

As Cummins points out, there is a strong link between this conflation of representation with "representation as" and use theories.

> One notes that one cannot be said to have the concept of assassination if one does not know that assassinated people are dead, and concludes correctly that whether or not someone has the concept of assassination depends on what they know. But if you think that concepts are mental representations, you will think you have concluded that whether or not someone has a mental representation of assassination depends on what he or she knows, and that is [conceptual role semantics, (i.e. a use theory)]. That conclusion doesn't follow, however...[because]...knowledge structures, like one's concept of an elevator, are not representations (Cummins, 1996, p. 89-90).

Consequently, we should not be surprised that if our notion of concepts has been nurtured in a use theory environment, it will have to change when the problem of explanatory vacuity drives us away from such theories.

Functions and the Problem of Explanatory Vacuity

At the beginning of this chapter, I replaced Cummins' theory of representational content, (1), with my own, (1*), but left the remainder of his theory of propositional attitudes unchanged.

(1*) A mental representation is an ordered pair ⟨S, M⟩, where S is a structure and M is an isomorphic mapping rule from S to another structure, *c*, which is the content of the representation.
(2) A mental representation is targeted on whatever structure, *t*, the intender that employs it has the function of representing.
(3) A propositional attitude's meaning is determined by
    (a) the attitude type (i.e. belief, desire, etc.), and
    (b) the content that *c* = *t*.

Just like (1), (1*) is an analysis of "mental representation," which tells us what mental representations are and which representations have which contents, but says nothing at all about use. However, nothing in the previous sections gives us any reason to alter (2) or (3), which describe how mental representations are used by cognitive systems to produce propositional attitudes. As I characterized above, (2) and (3) really belong to cognitive science rather than philosophical analysis.

This characterization is not really motivated by any strong inclination to draw a firm boundary between two independent and separate disciplines. Certainly, neither science nor philosophy can or should be practiced in isolation from the other. But I do want to insist that there are two very different sorts of questions being

answered here. The philosophical (1) and (1*) are answers to questions of ontology and analysis. They are necessary for cognitive science in the same way that the definition of the word "element" is necessary to the study of chemistry. If one doesn't know what is meant by the word "element," one cannot explain anything by reference to chemical elements, and if one doesn't know what is meant by the word "mental representations," one can't explain anything by reference to mental representations. Only after (1*) tells us what representations are will we be prepared to switch from philosophy to cognitive science, where we attempt to explain the behavior of cognitive systems by appeals to their use of representations. These issues will be the topic of the next chapter.

However, we are not yet ready to make the switch from philosophical questions to scientific ones. Though (2) and (3) make claims about representational use, they themselves rely on an as yet unanalyzed notion: the notion of a function. (2) tells us that we must know the function of an intender in order to determine its target. And, while (3) makes no explicit reference to functions, the propositional attitude types referred to in (3a) are presumably to be defined functionally (i.e. in terms of

which sorts of antecedent propositional attitudes and/or sensory and/or emotional states might cause them, and which sorts of consequent propositional attitudes and/or actions might result from them).

Of course, the mere presence of an unanalyzed term in our theory need not, in itself, require us to analyze it. One cannot be expected to provide an analysis for every term, or even every important term, one uses. In this case, however, a discussion of the proper analysis of "function" is in order because Cummins has himself offered an influential analysis of it (Cummins, 1975). What is more, the analysis that he has provided parallels his analysis of mental representations in *RTA*. Furthermore, as I intend to demonstrate below, the problem of explanatory vacuity appears with respect to functions in much the same way that it appears with respect to representations.

In "Functional Analysis" Cummins is actually arguing simultaneously against two claims about functions, one regarding the form of functional explanation and the other regarding the proper analysis of functional attributions. The two claims are, respectively

> (A) The point of functional characterization in science is to explain the presence of the item (organ, mechanism, process, or

> whatever) that is functionally characterized.
>
> (B) For something to perform its function is for it to have certain effects on a containing system, which effects contribute to the performance of some activity of, or the maintenance of some condition in, that containing system (Cummins, 1975, p. 49).

(A) and (B) are supposed to complement each other in the formation of functional explanations by allowing a scientist to deduce the existence of some object from the existence of the system containing it, along with the fact that the system containing it could not exist if the object failed to perform its function.

The problem with this sort of explanation, though, is that it is typically unsound, since it is rarely, if ever, the case that no other object could perform the same function. Hearts, for example, are not necessary to the continued existence of live animals, since any sort of suitable pump could do what hearts do (Cummins, 1975, p. 50). This is hardly surprising, since the whole point of talking about functions is that it allows us to abstract away from all the causal dispositions possessed by an object and concentrate only on those involved in its fulfillment of its functions. Once such abstraction has taken place, it is a good bet that some other object

exists, or could exist, which shares the dispositions that allow the original object to fulfill its function, while lacking many of its other dispositional properties.

Consequently, Cummins rejects (A). This leaves a vacuum where (A) had been that Cummins is happy to fill. Cummins claims that the real point of functional explanation is to explain some complex disposition (or capacity) of a system in terms of the less complex dispositions of its components (Cummins, 1975, pp. 54-55, 62-64). The fact that the heart has the function of pumping blood, for example, should be seen, not as explaining why hearts exist, but rather as explaining, in part, why blood circulates in those animals.

However, if we adopt this notion of functional explanation, (B) has to go as well. As it currently stands, (B) would analyze “the function of the heart is to pump blood” as something like “the heart pumps blood, and the pumping of blood contributes to the continued survival of the animal which contains it.” This sounds fairly reasonable, but (B) would also analyze “the function of the heart is to make 'thump-thump' sounds” as something like “the heart makes 'thump-thump' sounds, and those sounds contribute to the continued noisiness of the animal which

contains it." The problem is that, intuitively, only some of the effects that an object has on its containing system contribute to its fulfillment of its function (Cummins, 1975, p. 57). Thus, (B) really needs to be replaced with something like

> (B*) For something to perform its function is for it to have certain effects on a containing system, which effects contribute to the proper or adequate performance of some activity of, or the proper or adequate maintenance of some condition in, that containing system.

Unfortunately, this modification just leads to a different problem, because restriction to the proper or adequate performance of the containing system is really just a disguised appeal to the function of the containing system. As Cummins says, "it seems clear that for something to be in working order is just for it to be capable of performing its functions" (Cummins, 1975, p. 58). So, we can equivalently re-write (B*) as

> (B*) For something to perform its function is for it to have certain effects on a containing system, which effects contribute to the function of that containing system.

This means that we cannot hope to analyze every function using (B*), since every application of (B*) in an analysis of one function introduces a new function (of the

containing system) which will itself be in need of analysis.

So, if we want to keep (B*), we will have to supplement it with some other criterion for attributing what we might call "ultimate functions:" i.e. functions which can be analyzed without appealing to the function of a further containing system. In some cases, we may seem to have fairly reasonable ways of doing this. For example, in cases where the system is the result of intentional design, perhaps we can say that its function is the function its designer had in mind for it (Cummins, 1975, p. 53). In cases of biological systems, we can perhaps say that an organism's function is to contribute to the survival of its species (Cummins, 1975, p. 59).[4]

---

[4]Actually, we probably don't want to claim that the function of an organism is the survival of its species, since it is (normally) organisms that get naturally selected, rather than species. I take it that the appearance of the phrase "give or take a nicety" in his discussion of this ultimate function indicates that Cummins is aware of this, and that his use of the much-abused phrase "survival of the species" is meant to stand in for some more carefully worded selectionist theory of functions. This imprecision is tolerable here because the problem Cummins (and I) are interested in will still crop up no matter which particular selectionist theory of functions is used as an ultimate function. Consequently, in what follows, I will continue to employ Cummins' imprecise phrase.

If we do say this, then we will be able to say what the functions of the members of two significant classes (artifacts and biological organisms) are without appealing to any further functions. The explanatory strategy would then be to attribute functions according to (B*) in all cases except those involving either a complete intentionally designed artifact, or a complete biological organism. Those objects have either the function intended by the artifact's designer or the function of preserving the organism's species, respectively. Thus, a complete analysis of the notion of a function will be a combination of (B*) with two extra rules ((D) for "designed", (O) for "organism") for attributing ultimate functions.

> X performs its function if and only if
> (B^{*1}) X has certain effects, and either
> (B^{*2}) those effects contribute to some function of a containing system, or
> (D) X is an artifact, and those effects were intended by the designer of X, or
> (O) X is a biological organism, and those effects contribute to the survival of X's species.

Note that (B*) has been broken up, since (B^{*1}) will occur in the functional analysis no matter what. Obviously, part of what it is for something to have certain effects while performing its function is for that thing to have those effects.

However, this strategy of supplementing (B*) with rules for attributing ultimate functions won't work. The reason Cummins gives explicitly in "Functional Analysis" is that all such general rules for the attribution of ultimate functions will at best be rules of thumb, as there will always be cases in which intuitively appealing functional ascriptions conflict with the rules. Cummins offers the following example. Suppose that, due to the sudden disappearance of all pigeon predators, the sudden appearance of lots of local pigeon comestibles, and a temperate climate that negates the need to migrate, flying ceases to contribute to the survival of the pigeon species. Worse yet, since flying now wastes energy, it even becomes detrimental. Even so, says Cummins, "flight is a capacity which cries out for explanation in terms of anatomical functions regardless of its contribution to the capacity to maintain the species" (Cummins, 1975, p. 60).

There are, at least arguably, even non-biological cases which "cry out" for functional explanation despite the fact that they are products of neither natural selection nor an intelligent designer. For example, Saturn's ring system is a complex, organized, and self-perpetuating system. Furthermore, that system has

components which contribute to its perpetuation. For example, the bands between the rings are maintained by small “shepherd moons” which sweep the ring debris away from the clear bands between rings and into one of the rings. Though perhaps a bit poetic, it doesn’t seem totally unreasonable to say that the function of the shepherd moons is to sweep clean the bands between the rings.

Perhaps one could come up with an additional rule or even several additional rules for the attribution of ultimate functions which would cover these cases. But this would probably just generate conflicts between rules. Even if we stick to (D) and (O), conflicting cases are imaginable. Suppose that an insane pianist/surgeon kidnaps me, induces a coma, attaches an acoustic amplifier to my chest, and places me on top of his piano to be used as a metronome. Is the thumping of my heart one of its functions? If we apply the survival of species criterion, the noise made by my heart is a non-functional byproduct. But if we apply the intentional design criterion in this macabre situation, the heart does have the production of regular thumping noises as a function, since this allows

the metronome of which it is a part to fulfill the function intended by the insane pianist.

More recently, Cummins has specifically attacked selectionist accounts of function, like (O). Such accounts will count a trait as having a function only in cases where that trait has been selected because it performs that function. However, the theory of natural selection requires that in order for a trait to be selected because it has a certain function, there must be variation within some population of organisms with respect to that trait. This means that selectionist accounts will count a trait as having a function only if that trait was at some time possessed only by some proper subset of the total population. In other words, only novel traits can be selected.

But the introduction of novel traits is quite rare. Most natural selection is selection of variations in the quality of the performance of some function, rather than in the possession or otherwise of that function. Consequently, on selectionist accounts, only novel traits can have functions. This again rules out intuitively appealing attributions of functions. For example, it will turn out that the wing of the sparrow does not have a

function, since there has never been a time when only some sub-population of sparrows had wings. Of course, some ancestral population of modern birds must have existed in which only some of the members of the population had wings, but these ancestors would not have been sparrows, and their wings would not have been particularly similar to sparrow wings.

What Cummins takes all this to show is that the appropriateness of functional attributions does not depend on whether or not there is some principled way of picking out the ultimate functions of containing systems. Rather than attempting to save (B*) by supplementing it with rules for attributing ultimate functions, we should reject (B^{*2}) along with those rules.

However, an even better reason for rejecting (B^{*2}) is that it succumbs to the problem of explanatory vacuity. Suppose that we want to explain some capacity C of some cognitive system S by analyzing C into various functions carried out by subsystems of S. In particular, let *s* be one of these subsystems and *f* be a function performed by *s* which is to figure in the explanation of S's capacity C. If we analyze "*s* performs *f*" using (B*) supplemented with (D) and (O) what we will get is,

> *s* performs *f* $=_{df}$ *s* has the effects produced by the performance of *f* and those effects contribute to the performance of C by S.

(We can ignore clauses (D) and (O) in this case, since it has been stipulated that *s* is a subsystem of S.) But, since C is itself a function of S, "S performs C" must be similarly analyzed as

> S performs C $=_{df}$ S has the effects produced by the performance of C, and either
> (1) those effects contribute to some function of a system containing S, or
> (2) S is an artifact, and C was intended by the designer of S, or
> (3) S is a biological organism, and C contributes to the survival of S's species.

Note that the definition of "*s* performs *f*" makes reference to S's performance of C. If we combine these two analyses by substituting the second in for the relevant portion of the first, we will get

> *s* performs *f* $=_{df}$ *s* has the effects produced by the performance of *f* and those effects contribute to the fact that *S has the effects produced by the performance of C*, and either
> (1) those effects (produced by the performance of C) contribute to some function of a system containing S, or
> (2) S is an artifact, and C was intended by the designer of S, or
> (3) S is a biological organism, and C contributes to the survival of S's species.

We have just seen Cummins' reasons for worrying about clauses (1) through (3). The analysis of *f* generates an

appeal to C, which is a function of S. Since C is also a function, it too must be analyzed in the same way. Depending on what sort of system S is, this will generate either a further appeal to yet another function, as in (1), or a claim that C is some sort of ultimate function, as in (2) or (3).

But the problem of explanatory vacuity arises from the portion I've italicized. Regardless of which of the clauses (1) through (3) applies to S, the analysis of "*s* performs *f*" will always include the assertion that "S has the effects produced by the performance of C." In other words, when we say that the subsystem performs its function, part of what we are saying is that the system of which it is a part has a certain capacity. But according to Cummins, the fact that the system containing the subsystem has that capacity is the very thing we are supposed to be explaining, in part, by an appeal to the function of the subsystem. Consequently, all such explanations are vacuous. Whenever we attempt to explain a complex disposition of a containing system by specifying the less complex functions of its parts, each claim that one of the parts has some function will be analyzable into a conjunction, one conjunct of which is simply that the

containing system has the complex disposition we hoped to explain. The attribution of the less complex functions entails the existence of the complex disposition, so the former cannot be non-vacuously used to explain the latter.

So Cummins could, and I do, reject (B*[2]), (D), and (O) on the same grounds upon which we rejected use theories of mental representation. In fact, these three rules for the analysis of functional ascriptions form what we might call a use theory of functions. (D) says that certain systems have the functions used by their designers, (O) says that other systems have the functions used by organisms to survive and reproduce, and (B*[2]) says that still other systems have the functions used by their containing systems. Other theories of functions fall into the same pattern. For example, according to the goal contribution account of functions endorsed by Christopher Boorse, a disposition counts as a function just in case it plays a role in bringing about the realization of a goal, where goals are to be identified in terms of systems theory (Boorse, 2002). On such a theory, functions are dispositions used by a system to attain its goals.

Having rejected all such theories, we are left with (B*[1]). What we need is some different way of selecting

which effects are part of an object's function and which are not. What Cummins suggests is that the use rules should be replaced with a restriction, not on the kinds of systems which have functions, but rather on the kinds of explanation in which the attribution of functions may serve. Ever since our rejection of (A) above, we have been operating under the assumption that the point of functional explanation is to explain the complex dispositions of certain systems in terms of the less-complex dispositions of their parts. Cummins therefore proposes that we should say that an object is performing its function if and only if this attribution of a function to the object allows us to construct an interesting functional explanation of this sort. Roughly, this will happen, he says, when there is a sufficiently significant difference between the complexity and type of the dispositions of the system and its parts (Cummins, 1975, p. 66).

For example, this restriction (call it (C) for "Cummins") intuitively identifies hearts as pumps but not circulatory-noise-makers. Hearts are pumps because they are just one part of the more complex circulatory system, and pumping is just one activity among many involved in circulation. But, also in accordance with intuition,

hearts are not circulatory-noise-makers, because most of the noise made by the circulatory system is made by the heart. Analyzing the body's disposition to make circulatory noise into the heart's ability to make pretty much the same noise just doesn't buy us much explanatory power, since understanding one is pretty much the same as understanding the other.

So, the theory of functions Cummins ends up with is

X performs its function if and only if
(B*[1]) X has certain effects, and
(C) those effects contribute to the effects of a system which contains X and is both significantly more complex than X and has effects which are of a significantly different type.

(C) is admittedly pretty vague, but Cummins hastens to add that "there is no black-white distinction here, but a case of more or less" (Cummins, 1975, p. 66). Some cases are amenable to functional explanation, some are not, and many fall in between. Perhaps we could find a less vague way of characterizing which are which, but what is really important is that, whatever belongs in the place occupied by (C), it isn't any criterion that makes appeal to some further function, or to the function's use by a designer, organism, or whatever. Rather, it must be some way of

characterizing those attributions of functions which produce satisfactory explanations.

One reason Cummins may feel free to be vague about (C) is that it doesn't really seem to be telling us what functions are. That job, I think, is pretty much filled by (B^{*1}). Functions are just dispositions that happen to feature in satisfactory explanations of more complex dispositions in the right sort of way. In the right circumstances, any disposition could be a function (think of the mad pianist/surgeon's metronome). Cummins' criterion (C) tells us when we will be inclined to call some disposition of an object one of its functions (i.e. whenever this turns out to be explanatorily useful), but it does not really add anything to (B^{*1})'s analysis of what a function is.

For example, it isn't so much that one could not coherently assert that hearts have a noisemaking function, as that one wouldn't bother. If one wants to explain why human chest cavities are so noisy, one can simply point out that they typically contain certain noisy parts without additionally claiming that those parts have the function of making noise. The additional assertion that these parts have the function of making noise is coherent, but it just

makes the explanation more verbose without making it better.

On the other hand, if one wants to explain why blood circulates in humans, asserting that hearts have a pumping function allows us to more easily distinguish how hearts contribute to this effect from the ways in which other portions of the circulatory system contribute to the same effect, while ignoring those features of hearts that are irrelevant in this context. Calling a disposition a function, then, is something like calling a person a celebrity. The only difference between an ordinary person and a celebrity is that celebrities are thought to be particularly worth noticing if one is interested in popular culture. Similarly, functions are those dispositions thought to be particularly worth noticing if one is interested in explaining the behavior of complex systems.

Given this analysis of functions, we can add it to our theory of propositional attitudes. Re-writing (B*1) in a manner more consonant with the rest of the theory, we get

(0) A function of an object is a causal disposition of that object.

(1*) A mental representation is an ordered pair ⟨S, M⟩, where S is a structure and M is an isomorphic mapping rule from S to another structure, *c*, which is the content of the representation.

(2) A mental representation is targeted on whatever structure, *t*, the intender that employs it has the function of representing.
(3) A propositional attitude's meaning is determined by
    (a) the attitude type (i.e. belief, desire, etc.), and
    (b) the content that *c* = *t*.

The new clause (0) plays a similar role in the theory to that played by (1*). Both define key terms in the theory, which are then used in the remaining clauses (2) and (3) to explain how propositional attitudes are formed.

Unfortunately, the addition of (0) uncovers a previously undetected problem with (2). (2) tells us that representations are applied to targets by intenders, and that the targets of intenders are determined by their functions. But according to (0), since any causal disposition can be a function, the intender will have lots of functions. This means that (2) is not going to give us a unique answer to questions about correct or incorrect representation. Even if we know which representation is being employed by the intender, if we cannot decide whether the intender is targeted on this or that structure, we won't be able to tell whether or not the employed representation is a misrepresentation.

This problem is an analogue of the problem of underdetermination for representational content. I think that I have convincingly argued above that we cannot define the function of an object in a way that makes reference to its usefulness to some entity or other because such a definition trivializes the explanations of functional analysis. But if we therefore define function in a way that makes no mention of use, we are stuck with an underdetermination problem. It is all well and good to say that some system (like the intenders currently at issue) has lots of different functions, but sometimes we need to be able to pick out one or the other of those functions as the one relevant to explaining some further phenomenon (like the application, or misapplication, of representations to targets).

Since the problems are similar, it should come as no surprise that I think the solutions are similar too. Given the framework we inherited from Cummins' analysis of misrepresentation, the job of an analysis of representations is to determine what representations actually represent. The way we threaded our way between the representational problems of explanatory vacuity and underdetermination was to find a way to uniquely specify

the content of a representation without any reference to use: the key point being that this allowed us to disentangle questions of use (which will still come up when scientists attempt to explain behavior as resulting from successful or unsuccessful representation) and questions of content (which need to be settled prior to those attempts). By defining representation as we have, we achieved the goal of uniquely determining the actual content of representations. Once this analysis has provided a unique content for each representation, it is still up to the cognitive scientists to decide which representations are actually employed by the system.

On the other hand, the point of determining the function of intenders is to determine what the target of a representation is, which is to say, what it is that the intender is "trying" to represent. As it currently stands, (2) expresses the notion that the intender is trying to represent a particular target by saying that it is the function of the intender to apply representations to that target. The problem with this is that, according to (0), any causal disposition of the intender is a function of it, regardless of whether or not that disposition is something the intender, or the cognitive system of which it is a

part, is trying to do. The way to fix this problem is to realize that, just as there will generally be lots of different representations instantiated in a cognitive system and its environment, intenders instantiate lots of functions. And, just as it is the job of cognitive scientists to decide which of these instantiated representations are actually being employed by the cognitive system, it is the job of cognitive scientists to determine which of the many functions an intender happens to have is the one relevant to its cognitive use by the system. Consequently, I think we should modify (2) like so:

(0) A function of an object is a subset of the total set of causal dispositions of that object.
(1*) A mental representation is an ordered pair ⟨S, M⟩, where S is a structure and M is an isomorphic mapping rule from S to another structure, *c*, which is the content of the representation.
(2*) A mental representation is targeted on whatever structure, *t*, the intender that employs it has the *employed* function of representing.
(3) A propositional attitude's meaning is determined by
  (a) the attitude type (i.e. belief, desire, etc.), and
  (b) the content that *c* = *t*.

Assuming that cognitive scientists can tell us which representations and intender functions have been employed, we can now nail down the content and the target of a

propositional attitude. The target is the structure which the intender that employs the representation has the (unique) employed function of representing. The content is the structure which is (uniquely) picked out by the representation which is (uniquely) applied by the intender to the target. This then definitively tells us whether or not the formation of the propositional attitude involves correct or incorrect representation.

Employment

This last modification, replacing (2) with (2*), highlights yet another thus far undiscussed term in our theory of propositional attitudes. What does it mean to say that certain representations or certain functions, and not others, are employed? The answer lies in the forgoing discussion of functional analysis. The *explanandum* of a functional analysis is always some complex disposition of a system, and this disposition is always to be explained by showing how it arises from the less-complex functions of the system's parts. However, only some of the functions of those parts will be relevant to the explanation of the complex disposition. To say that a particular function has

been employed is just to say that it is one of the functions which makes such a relevant contribution to the production of the complex disposition.

For example, commenting on a swimming race held during the recent Olympics, one of the announcers exclaimed, “Look at him use those legs!” This exclamation is essentially a partial functional analysis of the swimmer’s disposition to quickly complete the race. By citing the swimmer’s use of his legs, the announcer called attention to the causal contribution that kicking his legs made to his speedy forward motion. A more complete functional analysis would presumably attribute to the swimmer the use of the paddling motions of his arms, the activities of his pulmonary and circulatory systems in delivering oxygen to his muscles, and so on. However, it would not attribute to the swimmer use of any of the functions of his body that make little or no contribution to his forward motion. The announcer did not exclaim, “Look at him use those lower intestines!” because, while the swimmer’s lower intestines no doubt continued to perform various functions during the race, and while some of these functions were no doubt used by the swimmer in the maintenance of other, more long-term dispositions of his body, none of the functions of the

swimmer's lower intestines contributed significantly to his swimming performance during the race. So in addition to asserting that a certain function plays a significant role in the production of a more complex disposition, claiming that certain functions are employed also tacitly gives us permission to ignore unused functions.

If employed functions are in general just those functions that play a role in a functional analysis, this should be true of the employed functions of the intenders mentioned in (2*). (2*) tells us that the employed function of an intender is to apply representations to targets. The surface grammar here suggests that this involves the intender's using representations to do something to targets, but this is deceptive. Applying a representation to a target is not like applying paint to a house. Rather, what intenders do is produce or modify brain states. This process is entirely local and needn't affect the intender's target at all. However, in virtue of producing a single brain state, intenders thereby bring into existence lots of isomorphs (i.e. one for each structure instantiated in the brain state) and the production of each of these isomorphs brings into existence lots of representations (i.e. one for each environmental

structure to which the produced isomorph is isomorphic). Each production of one of these representations counts as one of many functions of the intender. In order to say which of these many acts of representation production is the employed function of the intender, we need to know what complex disposition of the intender's containing cognitive system is the *explanandum* of the functional analysis in which that function plays a role.

Ultimately this *explanandum* is the intentional behavior of the cognitive system. In the first chapter I said, rather vaguely, that what needed explaining about intentional behavior was its "appropriateness" to its producer's situation. The interesting thing about intentional behavior is that it can be, in various ways, coordinated. Changes in our direction of motion can be coordinated with the presence of obstacles in our path, rotating our wrists clockwise can be coordinated with the threads on a standard screw, and uttering "there is a black hole orbiting radio source Cygnus X-1" can be coordinated with there actually being a black hole orbiting radio source Cygnus X-1.

Since this coordination is the relevant *explanandum* of our functional analysis, if we want to know what the

employed function of an intender is, we have to ask which of the representations it produces plays a role in bringing about that coordination. We can now see why and in what sense the production of representations (via the production of brain states) is the employed function of the intender. A noteworthy feature of the coordinational facts we are trying to explain is that they are not, in general, facts about spatio-temporally contiguous states of affairs. For example, given some sort of remote sensing ability, a robot need not come into contact with an obstacle in order to change direction so as to avoid it, and given memory, there need not even be any current causal influence of the obstacle on the robot. This feature is noteworthy because representations, as I have defined them, are themselves generally spatio-temporally non-contiguous. By producing a certain representation, the intender coordinates certain of its inner states (the representation's isomorph) with the, frequently distant, states of affairs with which the cognitive system's behavior is coordinated (the representation's content).

The production of this representation is therefore the intender's employed function. Of course, as was already mentioned, the intender produces lots of other

representations, and thereby sets up all sorts of other coordinations between states of the cognitive system and various bits of its environment. However, if none of these coordinations results in a further coordination of the cognitive system's behavior with its environment, such coordinations will not count as employed functions, since they play no role in the explanation by functional analysis of intentional behavior.

So, by way of summing up, the employed function of an intender is the production of representations which figure in functional analyses of the intender's cognitive system's capacity for coordinated, intentional behavior. An employed representation is a representation the production of which figures in such an analysis. The target of an intender's production of a representation is whatever feature of the world is coordinated with the system's behavior (or fails to be coordinated in the case of misrepresentation). An intender is then said to have applied a representation to a target when the intender's production of that representation figures in a functional analysis of the intender's cognitive system's capacity to coordinate its behavior with that target (or a functional

analysis of the system's failure to coordinate behavior in the case of misrepresentation).

For example, suppose we have observed a robot avoiding an obstacle, and want an explanation of how the robot was able to coordinate its movements with the presence of the obstacle. We find that the robot has a subsystem which produces a certain sort of state, and that this state has a structure isomorphic to the object. It follows by (1*) that it has produced a representation of the obstacle. If we further find that the production of this representation plays a role in a functional analysis of the avoidance of the obstacle, this will make the production of that representation the employed function of the producing subsystem, and it will make that subsystem an intender targeted on the obstacle. In such a case, we would say that the intender has applied the representation to the obstacle.

Note, however, that the employed function of the intender need not be to produce an accurate representation of its target. One reason why this is so is that we only have to explain actual behavioral coordination in whatever degree we find it. Perhaps, if the robot’s coordination of its movements with the presence of the obstacle is perfect,

we will explain this as the result of the intender's consistent employment of completely accurate representations. But suppose that the robot occasionally collides with the obstacle. So long as its course corrections usually result in avoidance, the robot's behavior is still coordinated with the presence of the obstacle, but it is imperfectly coordinated. In such a case, it may make more sense to say that the employed function of the intender is to produce only more or less accurate representations of the obstacle, since the representation need only be as accurate as the system's behavior in order to explain its performance.

Another reason why we might hesitate to assign perfect accuracy as the employed function of an intender is that the intender's execution of its employed function will explain the cognitive system's more complex capacity for coordinated behavior only in combination with the other functions involved in the relevant functional analysis. Depending on which other functions of other sub-systems play a role in that explanation, even in cases of perfect behavioral coordination, an intender might not have perfect accuracy as an employed function. Cummins' trout case is an example. Even if trout were always successful in

catching flying insects, this might still be attributed to their visual systems' misrepresentation of the insects' positions in combination with a suitable story about the compensating effects of other employed functions of other subsystems.

CHAPTER 3

THE METHODOLOGY OF COGNITIVE SCIENCE

Representational and Functional Uses

The point of the last chapter was to define the structure of the theories of cognitive science. What we have established so far is that, in order to allow for the possibility of representational error, such theories must appeal to both representations and the functions of the intender systems which employ them. Furthermore, we have established that, if the explanations of cognitive science are to be non-trivial, what representations and functions are, and what contents representations have, must all be definable without referring to the uses to which either representations or functions are put. Drawing from the work of Cummins, I have offered analyses of "representation" and "function" that satisfy this requirement. Finally, I have claimed that with these notions of representation and function in hand, it is the job of the cognitive scientist to determine which representations and functions are employed by the cognitive systems she studies, and how appeals to their use and/or

misuse can be used to explain and predict the behavior of those systems.

In this chapter I want to examine the task of the scientist attempting to construct such cognitive explanations of behavior. In so doing, however, it is not my intention to contribute much towards the completion of that task. In particular, I do not intend to attempt to evaluate the relative merits of different theoretic principles for assigning representational and functional uses to which scientists might appeal in constructing a cognitive explanation. Furthermore, to the extent that I do discuss such principles, I do not intend to be particularly original.

I fear that this may leave some philosophers and cognitive scientists cold, though I don't really think that it should. One of the main points of the previous chapter is that whatever principles scientists appeal to when claiming that a cognitive system is using or misusing a representation or function must be independent of their analysis of what a representation or function is. But since use theorists do not respect this separation, and since most philosophers who write on this issue are use theorists of one sort or another, their work on

representation typically contains arguments for precisely this sort of principle, and that precedent has engendered an expectation of such arguments in philosophical works on mental representation. As I will argue shortly, re-construed as scientific theories of representational use, rather than philosophical use theories of representational content, the principles offered by these philosophers show a good deal of insight. So good, in fact, that I don't feel that I have much to add in this arena. The contribution I am hoping to make to cognitive science in these pages involves discovering what the role of such principles is in cognitive science, rather than which of these principles scientists should adopt.

The first thing I want to emphasize is just how complex the cognitive scientist's task will be. The ultimate goal of cognitive science, at least as I am conceiving of it, is roughly to produce explanations and predictions of the behavior of cognitive systems by attributing various propositional attitudes to those systems, and applying (perhaps a cleaned-up version of) the folk psychology that relates stimuli and behavior to those attitudes. That is, once we can say what propositional attitudes a system has, folk psychology is supposed to get

us from, for example, “S desires P,” and “S believes S’s doing A will result in S’s getting P,” to “S will probably do A.” Though this project is not without its difficulties, I have been more or less assuming that it will work, since if this assumption is wrong cognitive science is a dead end anyway.

Instead, I have been concentrating on the role that the notion of representation plays in these explanations. The point of the attribution of representational uses is to bridge the gap between propositional attitudes and physical facts about cognitive systems and their environments. Folk psychological explanations of behavior will only really be explanatory if we can say, in antecedently understood and non-vacuous terms, what it is to have propositional attitudes, and which attitudes any given cognitive system has.

According to our theory of propositional attitudes, there are several distinct sorts of facts that scientists are charged with determining in order to generate a cognitive explanation of behavior. This complexity can be made apparent with that favorite of philosophical devices, the classification system. We have already discussed at length one way of dividing up the task: Cummins’

distinction between determining what is actually represented, and what the system is trying to represent. Let's call the tasks associated with the former "representational tasks," and the tasks associated with the latter "functional tasks" (since what the system is trying to represent is spelled out in terms of the employed function of the intender.)

However, orthogonal to this distinction is a further distinction between three distinct stages of the cognitive scientist's investigation of a cognitive system. I call the first of these stages the physical stage. On the representational task side, this involves discovering the spatio-temporal locations, dimensions, and properties of various physical objects and processes. On the functional side, it involves determining the causal dispositions of physical objects. Cognitive scientists need to determine these facts because the representations to which they hope to refer in their hypotheses are instantiated in various physical objects and processes, and the functions of intenders are subsets of their total set of causal dispositions.

The second stage is the abstraction stage. On the representational side it involves identifying isomorphs,

content structures, and the isomorphic mapping rules that connect them. On the functional side it involves abstracting away from the total set of causal dispositions of an object to those sets of dispositions that define its various functions. These facts constitute a second distinct stage because one could in principle know all the facts of the physical stage without knowing any of the facts of the abstraction stage. This is not to say that the facts of the abstraction stage are not reducible to the facts of the physical stage. Whether or not a given physical object instantiates a particular structure or function depends entirely on that object's composition and/or its causal interactions with other objects. However, the difficulty of the abstraction stage task is in recognizing all the different structures instantiated in various objects, noticing when any two of them are isomorphic, and considering particular subsets of causal dispositions. The task is one of attention, focusing on certain structures and dispositions, while ignoring others. For example, the fact that American stop signs instantiate an octagonal structure is fully determined by the physical composition of the sign, but one could still fail to attend to this fact.

The third stage is the use stage. Once the cognitive scientist has determined which representations are instantiated in which objects, and which functions can be abstracted from the total set of causal dispositions of which objects, she is still faced with the task of determining which of these representations and functions are employed by the cognitive system she is studying in the production of its intentional behavior. Only once she has completed this third stage will she be able to attribute particular propositional attitudes to the system, and in that way explain its behavior.

Determination of the facts of the later stages is clearly dependent upon prior determination of at least some of the facts of the earlier stages. The scientist cannot determine what representations and functions are instantiated until she has determined where things are in the universe, what properties they have, and what sorts of causal relations hold between them. And, of course, she cannot determine which representations and functions are being used until she determines which representations and functions exist, and are hence at least potentially available for use.

However, none of this should be taken to suggest that the order of the stages corresponds to a methodological order for cognitive science. Since every structure instantiated in every object in the universe is at least a potential isomorph or content structure, and since every subset of causal dispositions of every object in the universe is a potential function of that object, any scientist who sets out to complete the first stage in its entirety would have to (impossibly) produce a description of something like the complete trajectories and causal interactions of every sub-atomic particle in the universe. And even if the first stage were somehow completed, completion of the second stage would be just as daunting, involving as it would the identification of every structure instantiated anywhere or when in the universe, the pairing up of every pair of isomorphic structures, and the identification of every subset of causal dispositions.

Instead, the cognitive scientist will, like all scientists, allow theory to guide her investigations in promising directions. For the cognitive scientist, this guidance will come from theories of representational and functional use. If we have at least some rough theory about how cognitive systems have to be related to

representations and functions in order to use them, we can limit our abstraction stage search for representations and functions. For example, the mere fact that I have a brain state that is strongly isomorphic to the cat sitting in front of my open eyes does not, by itself, necessarily entail that I am using a representation consisting of the brain-state structure and a mapping rule that takes that structure to the cat. It is equally true that the fact that the Andromeda galaxy, which is strongly isomorphic to the Milky Way, is gigantic and far, far away does not, in itself, entail that I am not using it as an isomorph to represent my home galaxy. Of course, the first representation is undeniably vastly more likely to be accepted as an employed representation than the second, but this is undeniable only against a background of physical theory. Though the structure instantiated in the Andromeda galaxy is a perfectly good isomorph component of a perfectly good representation of the Milky Way, it is simply not causally related to my behavior in any way that could possibly explain it. The “cat-shaped” brain structure, on the other hand, may be suitably situated for this task, or may not, depending on exactly where and how it is instantiated.

Similarly, once we have decided which sorts of representations and functions are plausible candidates for use, we can limit our physical stage inquiries to facts about the objects that instantiate those representations and functions. For example, with respect to representations, the cognitive scientist will only need to know about those objects that plausibly instantiate isomorphs (like brain states) and about the objects which instantiate the contents to which their mapping rules take them. (The mapping rules are of course instantiated at both of these ends.) Furthermore, she will only need to know about those structures instantiated in those objects that form plausibly used isomorphs or content structures. She will not, for example, have to determine the location of all the mitochondria in all the neurons in which some plausibly used isomorph is instantiated, since her theory of representational use presumably tells her that structures incorporating mitochondria do not figure in usable representations. Of course, as with all science, such theory-driven research can potentially lead the researcher astray (who knows, maybe those mitochondria are more representationally useful than we thought). Consequently, the good cognitive scientist will keep an eye

out for abstraction stage and object stage facts that might allow for previously undreamt of useful representations and functions.

The Underdetermination of Representational and Functional Use

Considering the complexity of these three stages, cognitive science appears to be afflicted with what is probably the nastiest case of underdetermination in all of science. Usually, when philosophers of science talk about underdetermination they have in mind the underdetermination of theory choice by empirical facts. According to the Duhem-Quine thesis, since there are always multiple theoretical descriptions compatible with the totality of our observations, and even with the totality of all possible observations we could make or might have made, we have to choose a theoretical interpretation of those observations based on some further criterion. Scientists choose the simplest theory, or the theory that best unifies a broad range of empirical phenomena, or even the theory that best satisfies some aesthetic criterion. The process

whereby we appeal to some such super-empirical criterion we call an inference to the best explanation.

As I intend to discuss shortly, the notion of an empirical, or observable, fact is notoriously problematic. However, in order to get our discussion of underdetermination off the ground, I propose that we temporarily identify empirical facts by appealing to Bas van Fraassen's semi-definition of "observable" from *The Scientific Image*.

> X is observable if there are circumstances which are such that, if X is present to us under those circumstances, then we observe it (van Fraassen, 1980, p. 16).

By this standard, intuitively and charitably applied, practically everything studied by cognitive scientists will count as unobservable. Of course, some of the objects of the physical stage are observable. However, many are not. In most cognitive systems of any complexity, the isomorphs of employed representations will have to be microscopic in order that they may all fit within a system of reasonable dimensions (i.e. the isomorphs of representations employed by humans are instantiated in microscopic neurons or groups of neurons because lots of them need to be able to fit inside human heads). Furthermore, though some used

representations are representations of observable structures instantiated in observable objects, some are not. The content structures of some used representations are too small, too large, too energetically feeble, or too inconveniently located to be observed. As for the functions of potential intenders, as Hume taught us, even when the causes and effects related by some causal relation are observable, the causal relation itself is not. So functions, which are just causal dispositions, are never observable.

As we move on to the abstraction stage, it seems clear that none of these facts are observable on van Fraassen's criterion. This is because all abstraction stage facts are set-theoretic facts. Identifying a representation as being instantiated involves recognizing the presence of both an isomorph and a content structure, each of which involve recognizing sets of objects and sets of relations defined over those objects, as well as recognizing the isomorphic mapping rule that holds between them. Identifying a function as being instantiated involves recognizing a subset of causal relations. But we don't observe these sets. What we observe are the physical objects that are their elements, or the physical events which are the inputs

and outputs of causal relations. At most we might say that we observe that certain objects or dispositions form a set, but then any collection of objects or dispositions does this. The trick of the abstraction stage is just to attend to certain objects or dispositions while ignoring others, calling the objects or dispositions to which we attend elements of the set. That these objects form a set is something we recognize, rather than observe.

Similarly, it is clear that we cannot observe representational or functional uses, since we cannot observe uses in general. We can, of course, observe certain objects as they are being used. But the observation we make in such a case would not differ at all from the observation we would make if no use occurred at all. For example, I can observe a person using a hammer to drive a nail, but everything I observe in such a case would be compatible with that persons' being dead as a doorstop, and hence not using anything at all, but being remotely controlled by a clever robotics engineer, or for that matter, just blown around by a very complicated random pattern of strong wind gusts. This isn't to say that there won't be times when we feel confident in our assertions about use. But our confidence will be the result of our

confidence in having made a reasonable inference to the best explanation, not merely the result of what we have observed.

According to our theory of propositional attitudes, the criterion to which cognitive scientists must appeal in making this inference to the best explanation will be a principle for guiding an interpretation of cognitive systems as employing various representations and functions. Because so much is underdetermined in cognitive science, this interpretational task will be enormous. For the sake of a concrete and familiar example, let's suppose that some cognitive scientist is trying to explain how I can have, and ultimately act on, some particular belief about the Andromeda galaxy.

Consider first the physical stage. Since the objects which form human isomorphs are, at least according to most plausible theories, microscopic, one interpretational task of the scientist will be to interpret the readings of various instruments as indicative of the existence of assorted physical unobservables, such as neurons, neural firings, etc. Similar instruments will be used in detecting intender systems, also instantiated in groups of neurons, and the neuronal events to which they are disposed

to respond and which they are disposed to produce. On the other hand, since the content structure is instantiated in Andromeda, which is too faint and far away to be detected in detail with the naked eye, the scientist must rely on an understanding of the optics of telescopes in order to identify that content structure. Furthermore, since intender functions are subsets of causal dispositions, the scientists will need physical theories which ascribe various causal dispositions to the intender targeted on Andromeda.

However, this is not nearly the most important reason for thinking that mental representational usages are empirically underdetermined. For, even if we assume that the cognitive scientist can confidently claim to know a good deal about the microscopic physical objects which compose the brain, the distant objects that compose Andromeda, and the causal dispositions of the brain's intender systems, she will not thereby have determined which structures instantiated in those physical objects are the isomorphs or content structures of employed mental representations, or the causal dispositions of employed functions. This is true even if we assume that she has already ruled out as too implausible the possibility of my

using representations with isomorphs instantiated outside my own brain. After all, my brain instantiates many structures which, when paired to the appropriate mapping rules, represent various structures instantiated in the Andromeda galaxy. First, there is the trivial mapping between my whole brain, or any single part of it, considered as a degenerate structure, and the whole Andromeda galaxy considered in the same way. Besides that, there will be lots of other structures which, completely by accident, happen to map onto, say, the spiral structure of that galaxy, or the distribution of temperatures in it, or whatever. Each of these structures, identified in the abstraction stage, could, in principle, be used by some sort of cognitive system to form a representation of Andromeda.

Yet none of them is the isomorph a cognitive scientist trying to explain my belief about Andromeda is looking for. The isomorph she is looking for is the one that I actually use in forming that belief. The fact that it, like all the others, is the isomorph component of a representation of the Andromeda galaxy explains, in part, how representation of Andromeda is possible for me, but this fact contributes

not at all to the project of determining whether I actually use that representation to form beliefs about Andromeda.

The cognitive scientist will also have to determine which mapping rule I am pairing with the isomorph she has interpreted me as employing. Since the mapping rule consists of a set of ordered pairs such that the first element of each pair is an object or relation in the isomorph and the second element in each pair is an object or relation in the content structure, she will be looking at both objects in my brain and objects in my environment. The objects in my brain will be whichever brain states she has interpreted as elements of the isomorph. The objects in my environment will be whichever environmental objects she interprets as the elements of the content structure.

Furthermore, whether or not the latter of these sets of objects is instantiated in the Andromeda galaxy depends on whether the cognitive scientist has decided to interpret my use of the representation as a case of correct representation or as misrepresentation. If she interprets me as accurately representing Andromeda, then she must choose a mapping rule which will take me from the representational structure instantiated in my brain to an isomorphic content structure instantiated in Andromeda.

This is because we have defined accurate representation as a match between content and target. If, on the other hand, she has decided that I am misrepresenting Andromeda, then she will interpret me as using a representation which picks out some content other than Andromeda.

This introduces yet another theoretical interpretational task for the cognitive scientist. As I introduced the case, she is seeking to explain one of my beliefs about Andromeda, so some structure or other instantiated in Andromeda must be the target. However, the mere fact that the cognitive scientist has interpreted me as using a representation with such a structure as its content is not sufficient to explain my ability to form a belief about Andromeda, since I might not be using the representation for that purpose. Unless the representation can be interpreted as being employed by an intender targeted on some Andromedan structure, my use of that representation can't explain my ability to form a belief about Andromeda. Consequently, she must interpret some system in my brain as having the employed function of employing this representation of Andromeda.

Of course, as we stressed before, these determinations need not occur in this order. Rather, the cognitive

scientist, faced with a great mass of behavioral, anatomical, and environmental observations, must juggle interpretations of her instruments, employed isomorph, employed mapping rule, and employed intender function in such a way that all the interpretations cohere with the observations and each other. When the overall interpretation fails, the scientist may reinterpret by a) altering the theory of one or more of her instruments, b) altering her assessment of which structure produced by some intender is the relevant employed isomorph, c) altering her assessment of some isomorph/mapping rule pairing or, d) altering her assessment of the employed function of some intender.

Of course, as I mentioned earlier, the distinction between observable and unobservable facts has been widely criticized, and as a consequence, the notion of underdetermination of theory choice by the observable facts is problematic. Underdetermination is supposed be interesting because it suggests that theoretical appeals to unobservable facts are on shakier epistemological ground than pronouncements about observable facts. But if it turns out to be impossible to non-arbitrarily sort scientific facts into epistemologically problematic

theoretical facts and epistemologically less-problematic observable facts, the philosophical interest of underdetermination is diminished.

Criticisms of the observable/unobservable distinction come in what one might call optimistic and pessimistic varieties. Both sorts of criticism put claims about observables and unobservables on an epistemological par. However, optimistic critics tend to see claims about unobservable facts as not significantly worse off than claims about observable facts, while pessimistic critics tend to emphasize that claims about observables are just as epistemologically shaky as claims about unobservables.

The optimistic critic's case is grounded in vagueness in the notion of observation. For example, Grover Maxwell argues that any attempt to sharply delineate the class of observable facts is bound to be arbitrary.

> There is, in principle, a continuous series beginning with looking through a vacuum and containing these as members: looking through a windowpane, looking through glasses, looking through binoculars, looking through a low-power microscope, looking through a high-power microscope, et cetera, in the order given. The important consequence is that...we are left without criteria which would enable us to draw a non-arbitrary line between "observation" and "theory" (Maxwell, 1998, p. 1055-1056).

Of course, one can still allow that this continuous gradient corresponds to a similarly continuous change in epistemic status. But a continuous epistemic gradient falls short of entailing the philosophically interesting consequences often thought to follow from the fact of underdetermination. In particular, it does not motivate selectively taking some sort of anti-realist stance only towards unobservables, since

> Although there certainly *is* a continuous transition from observability to unobservability, any talk of such a continuity from full-blown existence to non-existence is, clearly, nonsense (Maxwell, 1998, p. 1057).

The optimistic critic's argument is bolstered by the apparently steady encroachment upon unobservable territory by what scientists generally think of as new methods of observation. Facts discovered by microscopes and telescopes, for example, might once have been thought of as involving theoretical inferences, but are today generally thought of as observations. Even such esoteric experimental apparatus as abandoned underground mines filled with carbon tetrachloride (a common cleaning fluid) are said to produce observations of the interior structure of the sun by detecting the neutrinos it emits (Hacking, 1998, p. 1165).

The pessimistic critic, on the other hand, argues that there are no purely observable facts at all, since all observation is theory-laden. The claim here is that the theoretical beliefs of an observer affect what she sees in unavoidable ways. N. R. Hanson argues that because of differences in background beliefs,

> You see a bird, I see an antelope; the physicist sees an X-ray tube, the child a complicated lamp bulb; the microscopist sees coelenterate mesoglea, his new student sees only a gooey, formless stuff. Tycho and Simplicius see a mobile sun, Kepler and Galileo see a static sun (Hanson, 1998, p. 90).

This, he claims, is not just a matter of different observers having the same visual experiences but placing different theoretical interpretations on them, since, although the two observers may share some sort of state in common (perhaps a retinal image), these commonly held states cannot reasonably or naturally be described as observational experiences. Whatever state the microscopist and his student have in common, by the time that state results in what can properly be called an experience, it will have been contaminated by their differing theoretical understanding, or lack thereof, of microscopy.

This distinction between optimistic and pessimistic criticisms of the observable/unobservable distinction is

not meant to suggest that the two sorts of criticisms are in some way incompatible. In either case, the substantive conclusion supported by the objections is that observability has no significant bearing on epistemology or metaphysics. A critic of the observable/unobservable distinction can quite consistently employ both optimistic and pessimistic arguments.

Nevertheless, the distinction is worth noting, as I intend to deal differently with the two sorts of criticisms in what follows. This section is supposed to be about the underdetermination of representational and functional uses, since the fact that such uses are in some sense underdetermined is going to be important to what I have to say about the new and proper role for re-construed use theories. Consequently, it would be a bit of an embarrassment if the claim that representational and functional uses are underdetermined turned out to be false, vacuous, or otherwise insignificant. On the other hand, if it can possibly be avoided, I would rather not have to defend the observable/unobservable distinction against the multitude of arguments ranged against it merely for the sake of making the points I am trying to make concerning the methodology of cognitive science.

Consequently, in order to avoid the tendentiousness of the observable/unobservable distinction, I think it will be useful for present purposes to remember the fact, alluded to previously, that underdetermination need not be underdetermination by the observable. Though this has traditionally been the sort of underdetermination most discussed by philosophers of science (primarily because it bears on the viability of empiricism), theory choice can be said to be underdetermined relative to any well-defined set of facts that may be of interest to us. That is, if we assume that we can determine some set of facts with certainty, and still not be constrained to choose a unique theory to explain that set of facts, then theory choice is underdetermined relative to that set of facts. Now each of the three stages identified above constitute a well-defined set of facts, and so we may ask which of these sets are underdetermined by which others. If any of the higher-level facts is underdetermined by the facts of lower levels, I will have all the underdetermination I need for the purposes at hand, without having decided which facts are or are not observable.

I am not, however, going to get my underdetermination from the gap between physical stage facts and abstraction

stage facts, since the latter follow by definition from the former. This means that, if we assume that we could come to know all the physical stage facts, observable or otherwise, with certainty, we would be able to uniquely determine which representations and functions exist. Since, as I argued in the previous chapter, representations and functions are just set-theoretic constructions of physical objects and their causal dispositions, the existence of representations and functions is completely determined by the existence of those objects and dispositions.

But the story is different with the representational and functional uses of the use stage. Even if we assume perfect knowledge of the physical stage facts, and hence complete determination of the abstraction stage facts, multiple assignments of representational and functional uses will be consistent with the physical and abstraction stage facts. After all, unless not just dualism, but interactionist dualism, is true, there is at least in principle a complete physical stage explanation for any physical stage fact, including those physical stage facts that describe human activity. Furthermore, since abstraction stage facts are definitionally determined by

physical stage facts, it follows that there is a physical stage explanation for all abstraction stage facts as well. So the only facts that could possibly require use facts for their explanation are themselves use facts.

But explanations are only required of those facts which are known with at least relative certainty, and the question currently before us is whether or not use stage facts are underdetermined by physical and abstraction stage facts. In other words, when we substituted the totality of physical stage and abstraction stage facts for the traditional totality of empirical facts, we substituted for both *explanans* and *explanandum*. Whereas traditional empiricism requires explanations only of empirical facts, in the present context we should require explanations only of physical and abstraction stage facts.

But if we had a physical stage explanation of all the physical and abstraction stage facts, and if only physical and abstraction stage facts require explanation, then our physical stage explanation would have explained every fact that needed explanation. Use stage facts would be unnecessary. It does not, of course, straightforwardly follow that there are no facts about use. However, the immediately relevant point is that, since there is a purely

physical stage explanation which is adequate relative to the totality of physical stage and abstraction stage facts, the choice between this explanation, at least, and any explanation which appeals to representational and functional uses, will be underdetermined by the totality of physical stage and abstraction stage facts.

So rather than worry about the underdetermination of use stage facts by empirical facts, I will just discuss the underdetermination of use stage facts by the totality of physical and abstraction stage facts. Now one might attempt to launch an objection to this "physical and abstraction stage facts/use stage facts" distinction analogous to either the optimistic or the pessimistic criticisms of the observable/unobservable distinction. However, neither sort of analogous objection will undermine my assertion that use stage facts are underdetermined.

Objections analogous to the optimistic criticisms will not work because there is no continuous gradient from physical stage facts to use stage facts analogous to the continuous gradient in the observability of physical objects and processes. The observability or otherwise of physical objects depends upon various of their physical properties, such as size, location, chemical composition,

etc., which allow them or prevent them from entering into causal interactions with our sense organs. Since these properties generally form continua, so does observability. But there are no corresponding facts that lie halfway upon some continuum between physical stage and use stage facts. Whatever exactly it means to say that a representation or function has been employed, it clearly doesn't mean that some thing, a use, has attached itself to the representation or function, and that if we could just detect this thing, we could tell the used representations or functions from the unused ones.

Similarly, there is also no cause to worry that some future instrument, the "use-o-meter," will be developed using our theory of representational or functional uses, and eventually come to be regarded as just as good as observation. The only qualm I have about this is the possibility that conscious introspection might count as a method of observing (at least some) use stage facts. I intend to devote a future section of this chapter to this qualm, but I will be ignoring it for the moment.

As for an analogue of the traditional pessimistic criticism, which argues that the theoryladenness of observation makes empirical claims just as

epistemologically shaky as other theoretical claims, I don't see how such an analogous objection could be constructed in a way that did justice to the independence of physics. The imagined analogous pessimistic criticism would have to claim that our determination of physical and abstraction stage facts is somehow theoretically dependant upon use stage facts. That is, it would have to claim that theorists who disagreed about representational and functional uses would be unable to agree about what the physical stage facts are because their different interpretations of use color the procedures by which they determine physical and abstraction stage facts.

But physicists, at least in their role as physicists, need not have any theories about use. To a physicist, a human being is just an annoyingly complex physical system which does what it does because its parts are governed by the same laws of physics that govern everything else. And unless interactionist dualism is true, he is right to do so. This doesn't mean that there is no place for cognitive science. The cognitive scientist is much more likely to correctly predict human behavior with her representational and functional uses than the physicist is with his particles and forces. But the fact that one can do physics

without any use theories at all shows that our beliefs about physical stage facts need not be contaminated by our use theories.

While we're on the subject of objections, some might wonder whether the fact that representational and functional uses are underdetermined in this way isn't a reappearance of the problem of underdetermination that I myself said was fatal to Cummins' theory awhile back. If underdetermination was bad for Cummins, isn't it bad for me? The answer is no, because my underdetermination is in the right place. Underdetermination appears for Cummins in his theory of representation, which is supposed to be an analysis of what representations are, and what it is for one thing to be a representation of something else. Underdetermination in such an analysis is like underdetermination in a definition. An underdetermined analysis fails to specify the meaning of the term analyzed, which is a disaster if one is hoping to use the analyzed concept in scientific explanation. But underdetermination appears for me in my (admittedly rough) outline of a scientific theory of representational use. As the Duhem-Quine thesis showed, underdetermination in science, though perhaps annoying and perplexing, is just an unavoidable

cost of doing business. We make an inference to the best explanation and move on.

This is why I have been trying to establish that use facts are in some sense underdetermined. Since the totality of physical and abstraction stage facts cannot even determine whether or not there are any use stage facts, it cannot determine whether I use this or that representation. So on what grounds can we say that my use of one representation is more believable than my use of the other? The answer here must be the same as the answer given in the case of empirical underdetermination. Appeal to my use of one representation must be part of a better explanation of the physical and abstraction stage facts than is appeal to my use of the other representation.

In the empirical case, the quality of an explanation is usually judged according to criteria of simplicity, unification, etc. The same criteria can probably be used to justify our preference of some attributions of use over others. For example, when cognitive scientists reject as wildly implausible any claim that I can use representations with isomorphs instantiated outside my brain (like the representation of Andromeda by the Milky Way encountered in the previous chapter), they are appealing to the fact that

such a use fails to cohere well with current theories of physics. Such gigantic structures won't have appreciable effects on the details of my behavior unless physics is badly incomplete.

However, in more interesting cases, deciding which of two competing use explanations is better using these traditional criteria alone may become problematic. Normally, the cases that cognitive scientists find truly perplexing will be ones in which they are trying to decide which of two (or more) representations I have employed, where both representations not only have isomorphs instantiated in my brain, but perhaps even have isomorphs more or less similarly instantiated in my brain. For example, they might be trying to decide whether my behavioral response to a certain stimulus was mediated by a representation used in early or late stages of visual processing. Even if they agree that simpler, more unifying explanations are best, it is easy to imagine that they might still disagree about which explanation is simpler, or more unifying.

This becomes even more plausible when one recalls that attributions of representational use must be accompanied by attributions of employed intender functions, and decisions

about whether or not to count the application of various representations to their targets as correct representation or misrepresentation. For example, in the first chapter we criticized Millikan's use theory on the grounds that it ruled out as conceptually impossible the trout which succeeds in catching flying insects, not by employing an accurate representation of the position of the insects, but rather by employing an inaccurate representation along with a compensating jumping mechanism. One might argue that such purported cases of compensating errors can always be more simply described as involving correct representation without the compensating error. However, as we saw above, given an appropriate, in this case evolutionary, context, the compensating errors account might turn out to be the simplest, or most coherent, overall. In such cases, appeals to simplicity and unification, which are hardly free of controversy even in the non-biological and non-social sciences, become particularly problematic.

A Proper Role for Use Theories

I think it is really this problem to which use theories of mental representation, such as Millikan's

teleosemantic theory or Fodor's causal theory, are potential solutions. We have already seen that, as use theories of representation, such theories succumb to the problem of explanatory vacuity. But they may yet succeed as scientific theories of representational use. While a philosophical theory of representation must not make reference to representational use, cognitive scientists must still determine which representations and intender functions are used in order to predict and explain behavior. Since these are use stage facts, and since such facts are underdetermined even relative to the totality of physical and abstraction stage facts, cognitive scientists need a theory of representational use to allow them to choose the best cognitive explanations of behavior. It seems to me that the factors to which use theorists have typically, but inappropriately, appealed in their attempts to define mental representation are just what the cognitive scientists need in order to determine which representations are used.

Consider, for example, the hackneyed case of the bug-catching frog. When a bug flies into tongue-range in his visual field, he usually catches and eats it. However, when pesky scientists fling bee-bees past, he usually

catches and eats them too. Does the frog incorrectly represent the bee-bees with a |bug|, or correctly represent them with a |little black moving thing|?

This is, of course, an instance of our good friend the disjunction problem, and use theorists have tested each other's theories of representation against it for years. We have already seen why, regardless of what they say about how this case is to be resolved, their theories of representation must fail. Basically, they have asked and attempted to answer the wrong question. They take the question raised by this case to be, "What is the content of the representation employed by the frog?" and since they try to answer this question by appealing to the frog's use of the representation, they run into the problem of explanatory vacuity. My theory of representation answers this question by assigning content to the representation employed by the frog intrinsically. However, it also intrinsically assigns contents to a whole bunch of other representations that play no significant role in determining the frog's behavior. It is quite possible that we could find multiple |bug|s and |little black moving thing|s instantiated in such a way that it is at least plausible that they might be employed by the frog to

produce the observed behavior. So cognitive scientists are left with a different question, "What representation (the content of which is intrinsically determined) has been employed by the frog?" How can the scientists answer this question?

Well, they might reason as follows. The frog is a biologically evolved system, shaped by natural selection. So frogs will, generally, employ those representations the employment of which historically lead to the successful replication of frogs. Catching bugs makes an obvious contribution to frog fitness, so there will be selection pressure towards the employment of |bug|s. But catching little black moving things doesn't necessarily make such a contribution, so there will be no selection pressure towards the employment of |little black moving thing|s. So, when the frog is placed in an abNormal environment, it will misrepresent bee-bees as bugs.

Alternatively, they might decide to try to make the frog's detection and discrimination task as easy as possible. In order to determine this, they will need to determine the frog's optimal detection conditions, and then put the frog in those conditions *vis a vis* bugs and bee-bees. Depending on how the scientists think the frog's

visual system works, they might, for example, turn up the lights, decrease the distance between the frog and the bugs and bee-bees, expose the frog to just one bug or bee-bee at a time, slow the bugs and bee-bees down, or coat the bugs and bee-bees with phosphorescent paint. If the frog is still trying to catch the bee-bees even under optimal conditions, it is a good bet that the frog's error is not one of misrepresentation. It just doesn't have the ability to employ representations that differentiate between bugs and bee-bees. On the other hand, if the frog now snaps only at bugs, it must have the ability to use different representations of the two types of stimulus, and it must sometimes use the wrong ones when the bee-bees are non-optimally situated for its detection apparatus.

Obviously, these two lines of reasoning have been contrived to mirror the considerations that appear in the use theories of Millikan and Fodor. But contrived as they are, the point is that they don't seem at all forced. Reasonable scientists, faced with the task of determining which intrinsically defined representations are and are not employed by a given cognitive system, will appeal to just the sorts of criteria that use theorists use (inappropriately) to define the content of representations.

And the disagreements these scientists are likely to run into about which criteria best determine representational use are just the sorts of disagreements that divide use theorists on the question of representational content. Consequently, most of what has passed for philosophical argument about representational content can probably survive if it is reinterpreted as argument about representational use.

Notice that treating teleosemantic and causal theories as theories of representational use, rather than use theories of representation, allows us to treat them as *ceteris paribus* principles for inference to the best explanation, rather than as exceptionless definitions. One of Cummins' chief arguments against idealized use theories of any sort was that they make misrepresentation conceptually impossible in whatever conditions they take to be ideal. But principles for inference to the best explanation are never taken to be conceptually necessary. Even the staunchest defender of simplicity as a truth-tracking virtue in physics would agree that it isn't simply a matter of definition that the simplest theory is true. At best, simple theories are more likely to be true, but nature always has the right to belie our expectations and

be perversely complicated. On the one hand, this weakens the simplicity principle, but it also lets it partially off the hook, since it needn't be exceptionless.

The same lesson applies to our re-construed use theories. Fodor's theory gets its initial plausibility because, *ceteris paribus*, it really is plausible that detectors won't misrepresent under ideal detection conditions. How could they malfunction in the really easy cases? Nevertheless, as we saw in the first chapter, for certain sorts of stimuli, like the hawk that gets too close to the mouse, this plausibility is trumped by other considerations. Similarly, Millikan's theory gets its initial plausibility from the idea that correct representation is adaptational. As a *ceteris paribus* principle, this is fine. Knowledge is, after all, power. But as Cummins' complaint about resource constraints makes clear, knowledge, like all power, is expensive, so sometimes it is best to do without. We have already seen how devastating these criticisms are for Fodor and Millikan's use theories. But re-construed as theories of representational use, the purpose of which is merely to choose the best available theoretical alternative, those criticisms provide no more serious grounds for objection

than the mere possibility of the truth of a needlessly complex particle physics provides an objection to the simplicity principle.

It is at this point that I fear that advocates of this or that use theory are about to be disappointed. Having finally gotten around to the issues they care so deeply about, I must announce that it is not my intention to debate here whether any particular use theory, suitably reinterpreted as a theory of representational use, would provide better guidance in the choice of cognitive interpretations. Many of the arguments offered by warring proponents of these theories over the years will not be significantly altered by the considerations I have advanced above, and I have nothing new to add to those arguments. All that has changed is the purpose of the theories. Rather than being conceived of as theories of mental representational content, which tell us what mental representations are and what their contents are, use theories should be re-conceived as theories of representational use, which guide us in deciding which representations a cognitive system employs.

Conscious Introspection and the Observation Base

As I noted in the previous section, the conclusion that facts about representational and functional employment are underdetermined, and hence the further conclusion that use theories of representation should be re-conceived as theories of representational use, would be undermined if there were some way of directly observing use facts. Many philosophers (and common sense) tell us that we have a perception-like faculty of introspection which allows us to examine our own mental states. Precisely which states can or cannot be introspected is debatable, but if it turns out that introspection allows us to observe what I call representational uses, it would seem that we should all uncontroversially accept these facts about use for the same reason that we accept facts about baseballs and thumbtacks: i.e. they are part of the empirical base. If so, then I have been making cognitive science out to be much harder than it actually is, and the role that I have assigned to re-conceived use theories is unnecessary.

However, I think a little reflection will show that, even if we accept that introspection is a reliable source of a certain amount of self-knowledge, introspection cannot

entirely deliver us from underdetermination. Owen Flanagan's discussion of introspection (which he refers to as "Cartesianism") in *The Science of the Mind* makes this particularly clear. Flanagan points out that there are several distinct sorts of self-knowledge that introspection might be thought to provide. For present purposes, we need discuss only three of his five categories of introspection.

> State Cartesianism...Each person has infallible and privileged access to the intentional and phenomenal states she is in...[knowing] whether she is in a state of belief, desire, hope, happiness or fear. We know our own moods, itches, pains, and afterimages for what they are and as they are.
>
> Content Cartesianism...Each individual has privileged and perfect access to the contents of the intentional states she knows herself to be in. So in addition to knowing that I am in a state of desire...I also know what I am desirous of.
>
> Process Cartesianism. One has privileged access (at the functional level, not at the neural level) to the internal mental processes. For example, if I ask you how many windows there were in the house you grew up in, and then after you answer, ask you how you figured out the answer, you will tell me that you pictured the house in your mind's eye, and then went from room to room counting windows. If Process Cartesianism is true, you know what you are talking about (Flanagan, 1991, p. 194-195).

In general, the plausibility of the reliability of introspective knowledge seems to decrease as we work our

way through this list of categories. Flanagan does offer some reasons for doubting State Cartesianism and Content Cartesianism. We sometimes seem to mistake an itch for a tickle, or a hope for a desire. Sometimes we feel that we have a desire or a belief, but aren't sure what it is that we desire or believe. We also sometimes accept third-person judgments about intentional states and/or contents by qualified experts against first-person reports based on introspection. For example, if we observe someone who appears to sincerely report a desire for one thing, but whose behavior seems entirely bent on avoiding that thing, we may conclude that he does not know his own desires.

However, even if we set these worries aside and assume that we do have pretty reliable access to our intentional states and their contents, representational use may still be underdetermined. Flanagan's distinction between Content Cartesianism and Process Cartesianism more or less parallels the distinction Cummins and I draw between propositional attitude content and representational content. I may, for example, be able to introspect that I believe there is a bird in the sky, but it will not necessarily follow that I can introspectively determine whether I form this belief by having a THINGS-IN-THE-SKY-

intender which employs a |bird| or by having a PLACES-THERE-ARE-BIRDS-intender which employs a |sky|. Introspecting that I have a belief that there is a bird in the sky involves State and Content Cartesianism, but further introspecting that this belief involves employing a |bird| rather than a |sky| requires Process Cartesianism.

But as Flanagan notes, "virtually no cognitive scientists espouse Process Cartesianism [because] the evidence for it is so bad" (Flanagan, 1991, p. 195). Cognitive scientists have frequently gathered strong evidence that contradicts many of our introspections about our own cognitive processes. As an example, Flanagan cites a famous study of memory retrieval (Sternberg, 1966). In the study, subjects were presented with lists of numbers between one and ten, and then asked to recall whether a particular test digit occurred as one of the elements in the list. Data regarding the time subjects took to answer this question was then collected. When asked how one performs such a task, most people report mentally surveying the list from left to right, stopping this search as soon as the test digit is found, if it occurs in the list at all. If this were an accurate report of the representational processes at work in memory retrieval, we

ought to find that reaction time is shorter when the test digit occurs earlier in the list. However, reaction times turned out to be unaffected by the position of the test digit within the list, suggesting that the subjects were actually exhaustively searching the entire list every time, regardless of the placement of the test digit.

In other cases, we simply don't seem to have any strong introspectively delivered convictions about our use of representations. Flanagan asks

> What mental processes account for your understanding of the words in this sentence? How is the alphabet stored in your long-term memory? Complete this word: **ppy. How did you do that?...How do you recognize msspilled wurds? Ha! You had a little trouble with that one, didn't you? How come? The point is that most people haven't a clue as to how they do any of these things (Flanagan, 1991, p. 195).

And I would add that what we don't have a clue about can't alleviate underdetermination about representational use.

Of course, there do seem to be well-documented cases in which our introspective intuitions get it right. In another famous study (Shepard and Metzler, 1971), subjects were asked to determine whether or not a pair of two-dimensional representations of three-dimensional objects were representations of different objects, or representations of the same object with different

orientations (i.e. such that one representation was merely a rotated version of the other). Once again, reaction times were recorded. Subjects reported the introspective experience of rotating one of the representations until it matched the other, if the two matched at all. This generates a prediction that reaction times should be greatest when the two matching representations are rotated through larger angles, and in this case the actual reaction times satisfy the prediction derived from introspection.

Nevertheless, considered in light of the many cases in which introspection either delivers faulty information about representational processes, or no information at all, such isolated successes are hardly a vindication of Process Cartesianism. For one thing, we might just put the successes down to luck. Furthermore, even if we become convinced that at least some of the successes are the result of a reliable capacity for introspection, if we find that we cannot discover some essential difference between reliable and unreliable introspection, we won't know which ones to trust. And finally, even if we could identify some reliable class of use facts that are reliably discoverable through introspection, so long as at least some use facts remain hidden to introspection, underdetermination will

remain. Given the spotty record of intuitions about use facts, I think it is safe to say that the underdetermination of use facts is quite significant, and that therefore the need for theories of representational use is great.

Realism and Anti-realism about Representational and Functional Use

Wherever empirical underdetermination appears, issues of scientific realism and anti-realism are right around the corner. If observation cannot settle a matter of theory choice, and our choice ends up relying on an appeal to super-empirical principles, we must ask ourselves whether we think those principles are truth-tracking. If we do, we will claim to be realists about the claims made by the theory, believing that its claims are true. But if we do not, we will adopt one or another anti-realist attitude towards it, accepting it as good science because it possesses some sort of virtuous characteristic, say, empirical adequacy, that falls short of truth.

I don't intend to argue for or against taking a realist attitude toward any particular theory of

representational or functional use. Indeed, I haven't even argued for or against scientific acceptance of any such theory, metaphysics aside, and I don't intend to start now. However, by way of further motivating careful attention to the problem of explanatory vacuity and my solution to it, it is worth pointing out that the failure of use theorists to keep theories of content and theories of use separate has caused significant confusion, and that this confusion evaporates once the distinction is restored.

Consider the following disagreement between Daniel Dennett and Jerry Fodor. Though he sometimes balks at the label, Dennett is an instrumentalist about representations and he seems to see his instrumentalism as relieving him of the obligation to (in every case) give a determinate answer to questions about representational content. Take, for example, the aforementioned case in which frogs are tricked into catching and eating bee-bees instead of bugs by flinging the bee-bees past in a way that mimics the flight of a bug. The case challenges us to say whether the frogs have misrepresented the bee-bee as a bug, and then set out to eat a bug, or correctly represented the bee-bee as a "little black moving thing," and then set out to eat one of those. Dennett argues that the content should be

determined by the frog's ancestors' environment of selection, and that therefore

> To the extent that there is just no telling what that environment of selection has been, there is also just no fact of the matter about what the frog-eye report *really* means" (Dennett, 1995, p. 408).

Consequently we don't have to worry about giving the wrong answer in this case, since there is nothing to be wrong about.

But Fodor, as a realist about representations, objects that unless there is a fact of the matter about how the frog manages to catch flies, which presumably involves there being a fact of the matter about how it uses representations to do this, its use of representations can't be selected for at all.

> Interpretivism [i.e. instrumentalism] is, *inter alia*, the view that, strictly speaking, we don't really have beliefs and desires. But, one supposes, what a creature *doesn't really have* can't help it much in its struggle for survival (Fodor and Lepore, 1993, p. 74).

However it now seems that both Dennett's position and Fodor's criticism of it are confused. Fodor is right that if the representation the frog uses has no determinate content, then its use of that representation will imply no determinate fitness. Dennett's refusal to assign a

determinate content to the representation is a refusal to completely define the notion of representation to which cognitive explanations refer. But no predictions are entailed by incompletely specified theories. It is as if a physicist were to appeal to a theory about electrons, expecting to get determinate predictions out of the theory, without specifying the strength of the charge on the electron. Dennett is being more than an instrumentalist here; he hasn't even said which theory it is that he is an instrumentalist about.

However, from the fact that the representation the frog uses has to have a determinate content in order to generate a prediction of the frog's fitness, it does not necessarily follow that there must be a determinate fact of the matter about which representation the frog uses. Regardless of whether we interpret the frog as using the representation with (determinate) content "bug" to catch bugs, or the representation (also) with (determinate) content "little black moving thing" to catch bugs, if the two interpretations predict that the frog will catch the same number of bugs in its Normal environment, both will predict the same (determinate) fitness. For that matter, we could easily construct empirically adequate theories of

use according to which the frog's representation has a more specific content like "fly," or a more proximal one like "little moving black retinal image." I suspect that even those generally predisposed towards realism will agree that the severity of the underdetermination problem for representational use at least aids the plausibility of anti-realism about such facts.

So, due to his conflation of theories of content and theories of use, Fodor has given us a good argument for being a "realist" about representational content (i.e. insisting that representations have a determinate content), and incorrectly taken it to also be an argument for realism about representational use. Due to the same conflation, Dennett inappropriately takes the plausibility of a merely instrumentalist stance towards representational use as an excuse to be an "instrumentalist" about representational content (i.e. to resist fully defining the notion of representational content). But once one distinguishes questions of content from questions of use, it is easy to see the independence of these two issues.

Similar problems arise in contemporary discussions of functions. A common complaint against Cummins' analysis of

function is that it is too indiscriminate. As Christopher Boorse puts it,

> [Cummins'] analysis is now generally seen to have a serious problem of overbreadth. For one thing, it generates functions in non-biological sciences where teleological language is absent. It implies that the function of mists is to make rainbows...Moreover, it creates false functions within biology too. Relative to our capacity to die of fluke infestation, our liver's capacity to house liver flukes is its function (Boorse, 2002, p. 65).

The force of the liver fluke case is that we do not, and presumably should not, assert that the function of our livers is to house liver flukes: that in so asserting we would be asserting a falsehood. However, when Boorse alludes to the absurdity of our endorsing this function, I think he is really thinking of our endorsing it as, in the terms of this discussion, an employed function. He cannot possibly object to its being a function in Cummins' sense (i.e. a disposition that can figure in an interesting functional analysis), since it is the fact that the liver fluke case satisfies these constraints (when it shouldn't) that is supposed to make the case a counterexample. Rather he is objecting that, even though one might successfully analyze our capacity to die from fluke infestation in part

in terms of our liver's disposition to house flukes, we should not say that this is what our liver is used for.

However, taken as a non-use analysis of functions, nothing Cummins says implies that we should or should not make any assertions at all about functional use. The fact that the liver has the function of housing flukes, and that this function could be employed in a functional analysis of our capacity to die of fluke infestation, allows us to construct an explanation of the latter in terms of our bodies' use of the former. And, to the extent that this functional analysis is a good one, it at least has the prospect of being heuristically valuable. We can imagine a pathologist saying to himself, "How did these people manage to die of fluke infestation?" and another answering, "Well, for one thing, their livers are perfect accommodations for the darn things." But even so, depending on our theory of functional use, we might still reject the claim that this function is really used by the body to produce this capacity, perhaps while insisting that other functions of the liver (say, the production of bile) are really used. If so, we would be taking an anti-realist stance towards the former and a realist stance toward the latter. When Cummins is accused of having an overbroad analysis, I think

that, in addition to his analysis of functions, he is being saddled with a theory of functional use that, rather implausibly, counts every function (i.e. every disposition) as being used. But endorsement of the former need not require endorsement of the latter.

The mist case is perhaps slightly more troubling for Cummins, since in this case even "teleological language is absent." That is, even though a functional analysis of a cloud's capacity to produce a rainbow can be analyzed into the dispositions of mist droplets to refract and reflect light, and hence might be heuristically valuable, physicists as a matter of actual practice do not say that mist has this function. But so what? Cummins has told us that if a physicist should ever claim that mist has the function of reflecting and refracting light in order to create rainbows, he means that mist does have the disposition to reflect and refract light, and that this, along with other factors, does create rainbows. Whether or not physicists decide to speak this way is then up to them.

CONFESSION AND CONCLUSION

Before I conclude I have a confession to make. I may have just sinned against ordinary language at least once, and perhaps twice. In the process of forming and refining my theory of propositional attitudes, I have identified four classes of objects: representations, functions, employed representations, and employed functions. Representations and functions are defined without reference to use in order to avoid the problems of explanatory vacuity as it appears in use theories of both representations and functions. Employed representations and employed functions, on the other hand, are to be identified by cognitive scientists who postulate that certain representations and functions are employed by certain cognitive systems in order to explain how those systems form propositional attitudes, and ultimately, the behavior of those systems.

However, I anticipate that devotees of ordinary language philosophy will object that the word “function” as it is ordinarily used really means what I mean by “employed function.” They may also object that “representation” really means what I mean by “employed representation.” In

other words, it may be that people ordinarily use the words "function" and "representation" in such a way that the notion of an unused function or representation is simply incoherent.

Now if this objection is meant to return us to use theories of representation and function, then it cannot stand, as we have seen that this leads to the two problems of explanatory vacuity. However, it is possible for the ordinary language objection to be made in a way that does not have this consequence. What the arguments Cummins and I have offered really entail is the conditional claim that if representations and functions are defined in terms of use, then those notions cannot be used to produce non-trivial explanations in cognitive science or functional analysis. Consequently, if one really wants to retain the idea that the words "representation" and "function" should be defined in terms of use, one can do so as long as one is willing to deny that appeals by cognitive scientists to these notions produce non-trivial explanations. In that case, in order to produce non-trivial explanations one would still need to refer to the intrinsically defined notions that I call "representation" and "function," but would presumably name them something else.

I am not myself a particularly devout follower of the ordinary language movement. Consequently I may not be a very good judge of my own sinfulness in this matter. When I consult my own, rather weak, intuitions I admit that I find my own use of "function" a bit awkward. This is perhaps partially because the concept I pick out with this term is basically the concept of a causal disposition, which already has a perfectly good term, namely "causal disposition." Intuitively, "function" is supposed to be a more informative, and hence more restrictive, term, and it seems somewhat natural to restrict the term "function" to employed causal dispositions. However, I have resisted the temptation to use "causal disposition" in place of "function," and "function" in place of "employed function" primarily because I wanted to preserve the symmetry between functions and employed functions on the one hand and representations and employed representations on the other. This symmetry is rhetorically useful because it emphasizes the fact that in both cases, in order to use either representations or functions in their explanations, cognitive scientists must posit that certain representations and/or functions are used. Furthermore, though I admit to finding my use of "function" less than

intuitively satisfying, I am less troubled by my use of "representation." Unlike the case of functions, the concept I pick out with the term "representation" does not have an intuitively natural alternative name. It would be a bit of a mouthful to call them "ordered isomorph/mapping rule pairs," so instead, we would have to invent a term: "potential representation" might do.

In the end, I have decided that my own way of naming these four concepts is, on balance, the best alternative. Since my pair of representation terms seems as intuitive as any pair is going to be, and since there is value in having function terms that parallel them, I am willing to accept a certain degree of discomfort about the intuitiveness of the function terms. If others should find their intuitions tugging them in a different direction, so that one or both of these pairs of terms should be replaced with a different pair, I have no strong objections, so long as the underlying conceptual structure remains the same.

It would be disingenuous to deny the strong intuitive pull of use theories. However, the value of adhering to this intuition must be weighed against the scientific applicability of the terms "representation" and "function." Cognitive scientists, biologists, and others use assertions

about the availability, applicability, and use of various representations and functions in order to explain the phenomena of their respective fields. If the terms "representation" and "function" are to be defined as use theories define them, all such explanations are vacuous, and hence are of no value to scientists seeking explanations for those phenomena. When philosophers of cognitive science or biology insist upon endorsing a use theory of representations or functions, they undermine the very scientific uses of these terms they wish to explicate.

So, we must resist the admittedly strong intuition behind use theories, and instead define "representation" and "function" as I have done, without reference to use. When we do, we find that there are indeed unused representations and functions: vastly more of them, in fact, than there are used representations and functions. It is easy enough to overlook these unused entities, since scientists normally mention a representation or function only when reference to its use is explanatory. However, although assertions about unused representations and functions don't themselves explain much, their coherence is necessary to the success of explanations involving representational or functional use.

REFERENCES

Akins, K.: 1995, "What is it Like to be Boring and Myopic?" in B. Dahlbom (ed), *Dennett and his Critics*, Blackwell, Oxford, UK, pp. 124-160.

Boorse, C.: 2002, "A Rubuttal on Functions," in A. Ariew, R. Cummins, M. Perlman (eds), *Functions: New Essays in the Philosophy of Psychology and Biology*, Oxford University Press, Oxford, UK, pp. 63-112.

Cummins, R.: 1975, "Functional Analysis" in E. Sober (ed), *Conceptual Issues in Evolutionary Biology*, 2nd edition, MIT Press, Cambridge, Massachusetts, pp. 49-70.

Cummins, R.: 1996, *Representations, Targets and Attitudes*, MIT Press, Cambridge, Massachusetts.

Cummins, R.: 2000, "Reply to Millikan," *Philosophy and Phenomenological Research*, Vol. LX, No. 1, pp. 113-127.

Cummins, R.: 2002, "Neo-Teleology," in A. Ariew, R. Cummins, M. Perlman (eds), *Functions: New Essays in the Philosophy of Psychology and Biology*, Oxford University Press, Oxford, UK, pp. 157-172.

Dennett, D.: 1995, *Darwin's Dangerous Idea*, Simon and Schuster, New York.

Flanagan, O.: 1991, *The Science of the Mind*, MIT Press, Cambridge, Massachusetts.

Fodor, J.: 1987, *Psychosemantics*: *The Problem of Meaning in the Philosophy of Mind*, MIT Press, Cambridge, Massachusetts.

Fodor, J. and Lepore, E.: 1993, "Is Intentional Ascription Intrinsically Normative?" in B. Dahlbom (ed), *Dennett and his Critics*, Blackwell, Oxford, UK, pp. 70-82.

Gould, S.: 1983, "Darwin's Untimely Burial – Again!" in L. Godfrey (ed), *Scientists Confront Creationism*, W. W. Norton and Company, New York, pp. 139-146.

Hacking, I.: 1998, "Experimentation and Scientific Realism," in M. Curd and J. Cover (eds), *Philosophy of Science*: *The Central Issues*, W. W. Norton and Company, New York, pp. 1169-1185.

Hanson, N.: 1998, "Observation," in J. Kourany (ed), *Scientific Knowledge*: *Basic Issues in the Philosophy of Science*, Wadsworth, Belmont, California, pp. 81-99.

Kitcher, P.: 1985, "Darwin's Achievement," in N. Rescher (ed), *Reason and Rationality in Natural Science: A Group of Essays*, University Press of America, Pittsburgh, pp. 127-189.

Kitcher, P.: 1999, "Explanatory Unification," in R. Boyd, P. Gasper, J. Trout (eds), *The Philosophy of Science*, The MIT Press, Cambridge, pp. 329-347.

Maxwell, G.: 1998, "The Ontological Status of Theoretical Entities," in M. Curd and J. Cover (eds), *Philosophy of Science*: *The Central Issues*, W. W. Norton and Company, New York, pp. 1052-1063.

Millikan, R.: 1984, *Language, Thought and Other Biological Categories*, MIT Press, Cambridge, Massachusetts.

Millikan, R.: 1995, "On Mentalese Orthography," in B. Dahlbom (ed), *Dennett and his Critics*, Blackwell, Oxford, UK, pp. 97-123.

Millikan, R.: 1996, "Thoughts Without Laws: Cognitive Science with Content," in *Readings in Language and Mind*, Blackwell Publishers Ltd., Oxford, UK, pp. 305-326.

Millikan, R.: 2000, "Representations, Targets and Attitudes," *Philosophy and Phenomenological Research*, Vol. LX, No. 1, pp. 103-111.

Perlman, M.: 2002, “Pagan Teleology,” in A. Ariew, R. Cummins, M. Perlman (eds), *Functions: New Essays in the Philosophy of Psychology and Biology*, Oxford University Press, Oxford, UK, pp. 263-290.

Searle, J.: 1991, "Minds, Brains, and Programs," in D. Rosenthal (ed), *The Nature of Mind*, Oxford University Press, Oxford, UK, pp. 509-519.

Shepard, R. and Metzler, J.: 1971, "Mental Rotation of Three-Dimensional Objects," *Science*, Vol. 171, pp. 701-703.

Sternberg, S.: 1966, "High-Speed Scanning in Human Memory," *Science*, Vol. 153, pp. 652-654.

van Frassen, B.: 1980, *The Scientific Image*, Clarendon Press, Oxford, UK.

Wittgenstein, L.: 1994a, "Tractatus Logico-Philosophicus," in A. Kenny (ed), *The Wittgenstein Reader*, Basil Blackwell Ltd., Oxford, UK, pp. 1-31.

Wittgenstein, L.: 1994b, "Meaning and Understanding," in A. Kenny (ed), *The Wittgenstein Reader*, Basil Blackwell Ltd., Oxford, UK, pp. 53-66.