# Chapter 1

# Identifying the Problem and Other Preliminaries

## *Two Problems about Representation*

We should be careful to distinguish two problems about mental representation. The first—the Problem of Representations (plural)—is a theoretical problem in empirical science. Although we know that states of and processes in the nervous system play the role of representations in biological systems, it is an open question just which states and processes are involved in which activities, and how. Moreover, it is an open question how these states or processes should be characterized. For example, orthodox computationalism holds that mental representations are realized as symbolic data structures, but there is considerable controversy among orthodox computationalists as to what kinds of data structures are involved in various processes. Connectionists (see, e.g., Rumelhart et al. 1986), on the other hand, hold that mental representations are realized as activation levels of ensembles of simple processors, and/or as the strengths of the connections among such processors. The problem to which these approaches offer competing responses is that of discovering a way of characterizing representations that will allow us to understand both their physical instantiations and their systematic roles in mental processes.

The second problem—the Problem of *Representation* (singular)—is, at least as I understand it, a paradigmatic problem in the philosophy of science. To a large extent, empirical theories of cognition can and do take the notion of mental content as an explanatory primitive. But this is a kind of explanatory loan

(Dennett 1978): If it turns out that the notion of mental representation cannot be given a satisfactory explication— if, in particular, no account of the nature of the (mental) representation relation can be given that is consistent with the empirical theory that assumes it—then, at least in this respect, that empirical theory must be regarded as ill founded, and hence as a less than adequate response to the drive for the kind of thorough intellectual understanding that motivates scientific theory in the first place.

We can get a better idea of what these two problems are, and how they are related, by surveying in very general terms the various answers that have been tendered to each of them.

*The Problem of Representations*

It is surprising that only four answers have been suggested concerning the sorts of things that can be mental representations. I am not certain that this list of ours is exhaustive, but every proposal I know of fits pretty clearly into one of these four. It doesn't really matter much; my topic is the nature of representation, not what sorts of things do the representational work of the mind. I survey the alternatives here mainly to help to put the main problem in some context.

*Mind-stuff inFORMed* An important scholastic theory holds that in perception the immaterial mind becomes inFORMed by the same FORMS that inFORM the thing perceived. The background metaphysics assumes that knowable or perceivable things are a combination of matter and FORM: the *stuff* and its properties. There are two basically different kinds of *stuff:* mental stuff and physical stuff. When physical stuff is inFORMed by redness and sphericity, the result is a physical red ball. When mental stuff is inFORMed by redness and sphericity, the result is an idea of a red ball—or, perhaps better, the result is a red ball *as mental object* (i.e., as idea) rather than a red ball *as material object*. According to this theory, when you perceive a red ball, the very same FORMS that make the physical object of your perception red and spherical make your idea red and spherical. But of course a red ball *in*



Figure 1.1
Aristotle mentally representing Graycat with a ball.

*idea* is a very different thing than red ball *in matter*. A red ball *in idea* doesn't take up physical space, though it does take up *mental* space.

The basic idea behind this theory is that to know something is, in a pretty straightforward sense, to *be* it. You know the red ball when you see it because *you* have what *it* has: redness and sphericity. Your mind literally *is* just what the physical stuff is, because to *be* red and spherical is just to be inFORMed by redness and sphericity. This doctrine seems to make the notion of mental representation perfectly transparent: The idea represents the red ball, and it represents it as red and as spherical because the idea *is* red and spherical and the redness and sphericity come from the physical ball. To represent the world is to have a model of it in (on?) your mind—a model made of different stuff, as models usually are, but a model just the same. It we draw a picture, we, as theorists, can just *see* what represents what—e.g., the thing on the left part of the thought represents the cat, and the thing on the right part of the thought represents the ball. According to this theory, representation is evidently founded on similarity (shared properties)—a similarity the theorist can just *see*. Of course, the thinker can't just see it, as Berkeley and Hume eventually pointed out, but that is an epistemological problem at most. The fact that we can't see the alleged similarity between our own mental representations and what they represent (or see the representations at all, for that matter) doesn't show that it isn't similarity that underwrites representation; it only emphasizes the trivial fact that we can't hope to infer the way the world is from prior knowledge of the fact that we have it represented correctly.

*Images* The favorite theory of Berkeley and Hume was that mental representations are images. Except for dropping the Aristotelian jargon, however, this is just the same theory over again; the "picture" in both cases is just the same. Images were frequently said to be red and spherical, though with some uneasiness. The scholastic metaphysics was gone, but the basic idea was the same: Images represent things in virtue of resembling them—i.e., in virtue of sharing properties with them (though, of course, a sphere in the mind—i.e., as it exists as an image—takes up no



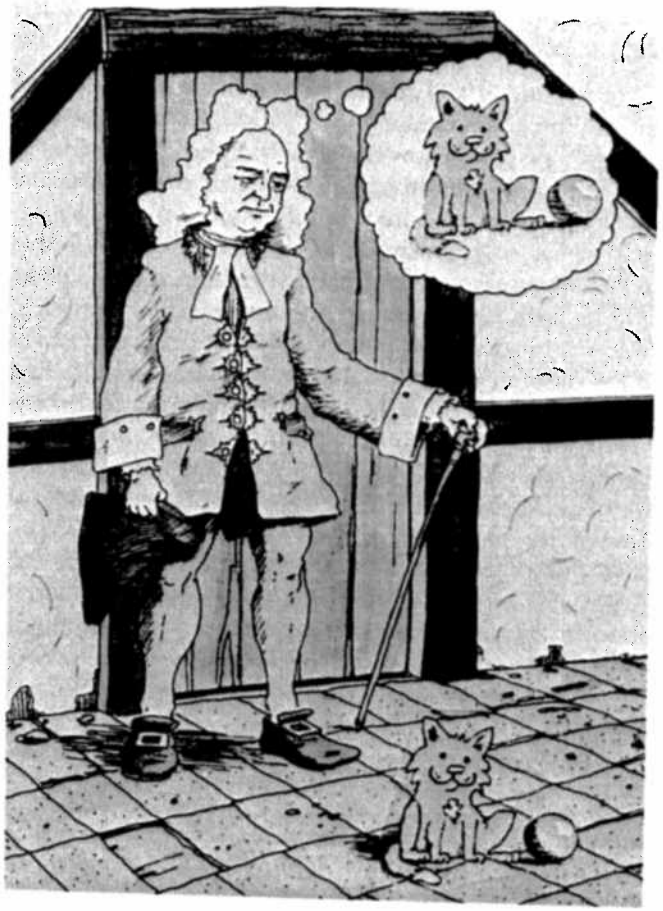

Figure 1.2
Berkeley's mental representations look just like Aristotle's.

physical space, only mental space; it occupies a portion of the visual field, for example).

*Symbols* Haugeland (1985) credits Hobbes with being the first to have an inkling that mental representations might be language-like symbols. This is now the orthodox position, insofar as there is such a thing. The main thing to realize at this stage is just that if mental representations are symbols, then mental representation cannot be founded on similarity; symbols don't resemble the things they represent. The great advantage of symbols as representations is that they can be the inputs and outputs of *computations*. Putting these two things together gives us a quick account of the possibility of thought about abstractions. When you calculate, you think about numbers by manipulating symbols. The symbols don't resemble the numbers, of course (what would resemble a number?), but they are readily manipulated.

Connectionists also hold that mental representations are symbols, but they deny that these symbols are data structures (i.e., objects of computation). In orthodox computational theory the objects of computation are identical with the objects of semantic interpretation, but in connectionist models (at least in those using truly distributed representation) this is not the case.[1] Connectionists also typically deny that mental symbols are language-like. This is not surprising; given that the symbols are not the objects of computation, there would be no obvious way to exploit a language-like syntactic structure in the symbols anyway.

*(Actual) neurophysiological states* The crucial claim here is that mental representations cannot be identified at any level more abstract than actual neurophysiology. Mental representation, on this view, is a biological phenomenon essentially. Mental representations cannot be realized in, say, a digital computer, no matter how "brain-like" its architecture happens to be at some nonbiological level of description.

Like symbols, neurophysiological states cannot represent things in virtue of resembling them. Advocates of symbols or neurophysiological states must ground representation in something other than similarity.



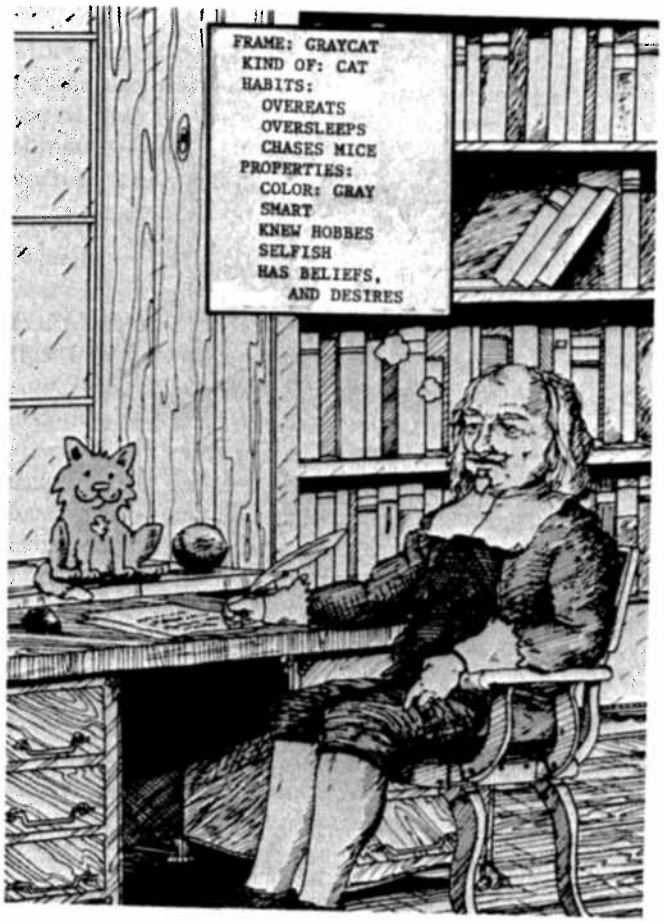


Figure 1.3
Hobbes representing Graycat.

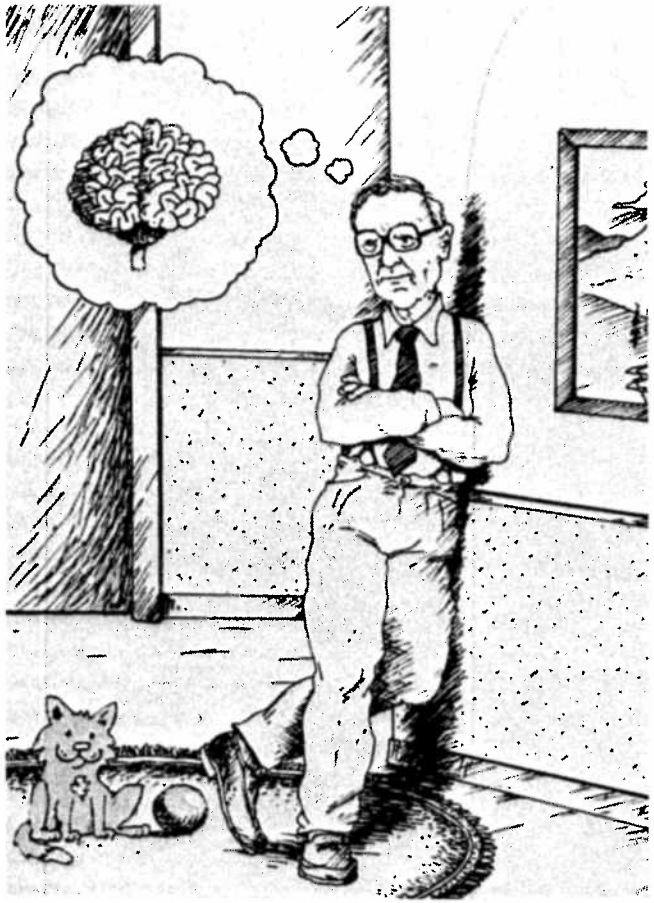

Figure 1.4
Hebb mentally representing Graycat.



*The Problem of Representation*

More surprising than the dearth of candidates to play the role of mental representations is the dearth of suggestions concerning the nature of representation itself. There are, I think, only four: similarity, covariance, adaptational role, and functional role. Each of these will be the subject of a chapter. For now, I will supply only brief intuitive sketches.

*Similarity* The thought that representation is grounded in similarity is what drives the idea that mental representations are in-FORMed mind stuff, or images. The crucial intuition, I think, is this: If you are going to think about things in the world, you need something to go proxy for those things in thought. You cannot, of course, literally turn over cats or the body politic in your mind; all you can turn over is ideas. But this, it seems, will be no help unless ideas are *like* the cats or the body politic: How could having an idea of a cat help you know about cats unless the idea is like the cat? I could say, "OK, this salt shaker represents the pitcher, and the pepper shaker represents the batter." But wouldn't *pictures* be much better—especially moving pictures, such as those in Rod Carew's batting instruction video?

*Covariance* The idea that representation is grounded in covariance or causation is most naturally motivated by reflecting on vision research.[2] How do we decide, for example, that a certain neural structure in the visual cortex of a frog is a motion detector? Roughly, we notice that a certain characteristic activity in the structure covaries with the presence of moving objects in the frog's field of vision. Given this fact, it seems natural to suppose that what *makes* that structure a motion detector is just the fact that it fires when there is motion in the frog's field of vision. What else could it be? So the fact that the firing of the structure in question represents the occurrence of motion in the frog's visual field is just constituted by the covariance between the firings and the motions represented. If you are attracted to covariance theories, you aren't going to think much of the idea that representations are images, because the similarities images promise to deliver are going to be irrelevant.

*Adaptational role* The idea that representation is grounded in adaptational role is most easily understood as a reaction to certain problems facing covariance theories. The orientation of a bee dance represents the location of flowers to spectator bees, but it doesn't covary with the location of flowers any better than it covaries with lots of things it doesn't represent, e.g., the absence of an insecticide cloud in the indicated direction. Millikan (1984) points out that we take "flowers over there" to be the content of the dance, even if flowers are not often "over there" (and hence there is no substantial covariance), because the cases in which spectators have found flowers (hence food) "over there" account for the continued replication of the dance and the characteristic response it evokes in spectators.

*Functional or computational role* This is just functionalism applied to mental representations. Functionalism says that a mental state is what it is in virtue of its functional role. It is functional roles that individuate mental states. But mental representations are, by definition, individuated by their contents. Hence, content must depend on functional role.[3]

## *Meanings and Meaningfulness*

When we ask what it is in virtue of which something (a mental state, a stop sign, a linguistic utterance) has a meaning or has semantic content, there are two quite different things we may have in mind. We may be asking what it is in virtue of which things of the sort in question have any meaning at all, or we may be asking what it is in virtue of which some particular thing or type of thing has some particular meaning. Although it is rather obvious that a theory that answers the first sort of question (a theory of meaningfulness) needn't provide answers to question of the second sort, it is not so obvious that a theory that provides answers to questions of the second sort (a theory of meaning) must also be a theory of meaningfulness. All the theories I will examine in this book are intended primarily as theories of meaning, not as theories of meaningfulness; but each of them entails, in an obvious way, a theory of meaningfulness. I shall try to make this explicit and, when appropriate, to be clear about whether the theory is being expressed and evaluated as a theory of meaning or as a theory of meaningfulness.



Theories of meaning, in the sense just staked out, should be sharply distinguished from theories that, as it were, distribute meanings (or some other semantic property) over the things that have them. For example, it is perfectly possible to articulate a theory that specifies a truth condition for every sentence in a language but that is entirely neutral concerning what it is in virtue of which a sentence has any truth condition at all, or in virtue of which it has the particular truth condition it happens to have. Tarski's theory of truth is, notoriously, just such a theory—truth is defined in terms of satisfaction, and satisfaction is defined recursively. The theory is completely silent about what satisfaction is. If we ask "In virtue of what is '$X_1$ is a cat' satisfied by every sequence beginning with a cat?" the theory gives no answer (see Field 1971 and Cummins 1975a). Linguists and psychologists want to know which things have which meanings, and why. Philosophers want to know what it is to have a meaning. With any luck, good philosophy might help with the "why" part of the question asked by linguists and psychologists. By my lights, that is really the only thing that could make it good philosophy.[4]

## *"Content"*

When we suppose a system to harbor cognitive representations, we are supposing that the system harbors states, or perhaps even objects, that are semantically individuated. Thus, the central question about mental representation is this: What is it for a mental state to have a semantic property? Equivalently, what makes a state (or an object) in a cognitive system a *representation?*

When we ask what it is for a cognitive state to have a semantic property, there are a number of different things on which we might choose to focus. What is it for a cognitive state to have a truth condition? What is it for a cognitive state to be about something, or to refer to something, or to be true of something?[5] What is it for a cognitive state to be an intentional state (i.e., to

have intentional properties)? The (very) recent tendency in philosophy has been to see all these questions as depending on two prior questions: What is it for a cognitive state to have a content? What is it for such a state to have some specified content, e.g., the content *that Brutus had flat feet* or the content *square?* This, I think, is a useful way to proceed—not because the notion of content is especially clear or simple, but because "content" can function in philosophical investigation as a kind of generic term for whatever it is that underwrites semantic and intentional properties generally. There is little to be gained, and there is a non-negligible risk of bias, if we begin by focusing in a fussy way on semantic or intentional concepts borrowed from theoretical or common-sense discourse about language and the attitudes—concepts that may not apply in any straightforward way to the problem of characterizing the representations assumed by contemporary cognitive theory. In what follows, when I write of the semantics of cognitive systems, or of representations, I mean to address these still poorly defined questions of "content." Since I shall be examining various "theories of content," there is no point in trying to say in advance what "content" means; let the theories speak for themselves. Meanwhile, our intuitive grasp of the thing will have to do.

*Methodology*

It is commonplace for philosophers to address the question of mental representation in abstraction from any particular scientific theory or theoretical framework. I regard this as a mistake. Mental representation is a theoretical assumption, not a commonplace of ordinary discourse. To suppose that "common-sense psychology" ("folk psychology"), orthodox computationalism, connectionism, neuroscience, and so on all make use of the same notion of representation seems naive. Moreover, to understand the notion of mental representation that grounds some particular theoretical framework, one must understand the explanatory role that framework assigns to mental representation. It is precisely because mental representation has different explanatory roles in "folk psychology," orthodox computational-



ism, connectionism, and neuroscience that it is naive to suppose that each makes use of the same notion of mental representation. We must not, then, ask simply (and naively) "What is the nature of mental representation?"; this is a hopelessly unconstrained question. Instead, we must pick a theoretical framework and ask what explanatory role mental representation plays in that framework and what the representation relation must be if that explanatory role is to be well grounded. Our question should be "What must we suppose about the nature of mental representation if orthodox computational theories (or connectionist theories, or whatever) of cognition are to turn out to be true and explanatory?" As I understand this question, it is a question in the philosophy of science exactly analogous to the following question in the philosophy of physics: What must we suppose the nature of space to be (substance? property? relation?) if General Relativity is to turn out to be true and explanatory?

The bulk of this book is an attempt to evaluate existing accounts of the nature of mental representation in the context of computational theories of cognition. By computational theories of cognition I mean *orthodox* computational theories—theories that assume that cognitive systems are automatic interpreted formal systems in the sense of Haugeland (1981, 1985), i.e., that cognition is disciplined symbol manipulation.[6] In the final chapter, I will consider briefly how things might look in a connectionist context.

Computational theories assume that mental representations are symbolic data structures as these are understood in computer science. This is the computationalist answer to the Problem of *Representations.* Although the instantiation of symbolic data structures in the brain is problematic, orthodox computationalism has demonstrated the physical instantiability of such structures and has made considerable progress toward demonstrating that at least some cognitive processes can be understood as symbol manipulation. But, like all theoretical frameworks in cognitive science, orthodox computationalism is silent about the nature of representation itself; it is entirely agnostic concerning what it is for a data structure to have semantic properties. Nevertheless, certain possibilities are ruled out by the empirical assumptions of the theory, as we will see.

I will need a short, convenient way to refer to what I have been calling orthodox computationalism; I'll call it the CTC, for the computational theory of cognition.

*Representation and Intentionality*

This preliminary issue of the explanatory role of mental representation in some particular theoretical framework would not be troubling if mental representation were a *commonplace* rather than a (variously) theoretically motivated *hypothesis*. Most philosophers aren't troubled; they think mental representation *is* a commonplace. They think this because they assume that the problem of mental representation is just the problem of intentionality—i.e., that representational content is intentional content. As I use the term, a system with intentionality is just a system with ordinary propositional attitudes (belief, desire, and so on). Thus construed, intentionality is a commonplace, and hence so is intentional content. So the assumption I want to scout is the assumption that the problem of mental representation is just the problem of what attaches beliefs and desires to their contents.[7]

Although it is evidently a mistake to identify intentionality with representation, there is a widely bruited philosophical theory, mainly due to Fodor, that forges a close connection between intentional contents and representational contents. I call this theory the *representational theory of intentionality* (RTI). The RTI holds that intentional states inherit their contents from representations that are their constituents. The familiar ur-theory goes like this: To have a belief is to have a representation in one's belief box—a box distinguished from the desire box by its function, i.e., by which processes can put things in and take things out. (Belief-box contents are available as premises in inference; desire-box contents are available as goals, i.e., conditions whose satisfaction ends processing cycles.) My belief that U.S. policy in Central America is folly is *about* Central America because the relevant representation in my belief box represents Central America.[8] The RTI has some nice features. Most notably, it captures the two attributes of the propositional attitudes to which we allude when we call them by that name: that they have



propositional contents and that believing involves taking a different "attitude" toward a proposition than desiring. But in spite of its nice features, the RTI is no truism; it is a controversial and powerful empirical theory.

If you accept the RTI explicitly, you will, of course, want a theory of mental representation that attaches intentional contents—the contents of propositional attitudes—to representational states. You will also want a theory of mental representation like this if you are merely sloppy about the difference between mental representation and the attitudes. I think this particular bit of sloppiness is pretty common in a lot of recent philosophical discussion of mental representation, but it doesn't really matter; anyone who assumes, for whatever reason, that a theory of mental representation must give us intentional contents (e.g., objects of belief) is making a very large assumption, an assumption that isn't motivated by an examination of the role representation plays in any current empirical theories. After all, it isn't belief of any stripe that most theoretical appeals to mental representation are designed to capture. Just think of psycholinguistics, which got all this representation talk started. The data structures of your favorite parser are not even *prima facie* candidates for belief contents. This is nonaccidentally related to the fact that the CTC, as we will see in chapter 8, makes use of a notion of representation that is at home in computational systems generally, not just in cognitive systems and certainly not just in intentional systems. If we begin our investigation of mental representation by focusing on intentional states, we will miss what is most distinctive about representation as invoked by the CTC. We certainly do not want to assume, therefore, that the contents of beliefs as ordinarily attributed are the contents of any representations in a computational system. We need to keep open the possibility that, e.g., belief attribution, though a legitimate case of semantic characterization, is not a semantic characterization of any representation in the believer (Dennett 1978; Stalnaker 1984; Cummins 1987).

The fact that current philosophers who are interested in mental representation do not follow the methodological path that I recommended in the last section is explained to some extent by

the prevalence of the assumption (often bolstered by the RTI) that the problem of mental representation is to explain how intentional contents (the contents of belief, desire, etc.) get attached to mental states. This assumption puts very strong constraints on the theory of mental representation. In fact, the constraints are so strong—so hard to satisfy—that one is never tempted to look elsewhere for something to constrain the problem; the last thing one needs is another constraint. Thus, you will never be moved to ask after such things as the explanatory role of representation in, say, John Anderson's ACT* (1983). Conversely, once you abandon (or at least question) the idea that the theory of mental representation must yield contents for intentional states, you *need* a few constraints, and the explanatory structure of a theory that invokes the notion of representation is the natural place to look.

### *Inexplicit Content* [9]

The attribution of intentional states (beliefs and desires) is not the only kind of semantic characterization of cognitive systems that must be distinguished from explanatory appeals to representational states. A computational system can also be semantically characterized in virtue of features of its structure. Here are some examples.

*Content implicit in the state of control* A word processor's search routine tries to match the character currently being read against the second character of the target only if the character read last matched the first character of the target. If it is now trying to match the second character, the current state of control carries the information that the first character matched the last character read; however, the system creates no data structure with this content. Nowhere is that information explicitly represented.

*Content implicit in the domain* I give you instructions for getting to my house from yours, all in such terms as "go left after three intersections" and "turn right at the first stop sign after the barn." Perhaps I even include things like "Make a left down the alley with the blue Chevy van parked in it," because I know you will



be coming after 5 o'clock and I know that the van is always parked there after that time. I rely on this in the same way I rely on the barn's staying put. Now, if you (or anything else) execute this program, you will get to my house. In the process, you never create a representation of the form "Cummins lives at location L"; yet, given the terrain, a system executing this program does "know where Cummins lives."

*Content implicit in the form of representation* Most of us don't know how to multiply (or even add) roman numerals. "XXII times LXIV" has the same *meaning* as "22 times 64," but the partial-products algorithm we all learned in school exploits information that is implicit in the second *form* but not present in the first—e.g., that shifting a column to the left amounts to multiplying by 10. This is the famous problem of knowledge representation in artificial intelligence: find a form that makes more efficient or psychologically realistic algorithms possible.

*Content implicit in the medium of representation* Are the two parts of figure 1.5 the same? If you had each one on a transparency, you could simply put one over the other and rotate them relative to each other to see if they would match. But this works only because of two properties of the *medium* (i.e., the transparencies): They are transparent, and they are rigid in the plane of the figures. When you rotate them, the information about the relative spatial relations of parts of a figure to other parts is implicit in the medium; its rigidity carries the information that these relations remain constant. A different medium might not carry this information, and you would then have to represent it explicitly.

I am sure these examples don't exhaust the cases in which content can be attributed to a computational system in the absence of any explicit representation having the content in question. I have listed them here only to emphasize the fact that represented content isn't all the content there is. There is also inexplicit content of various kinds, and if nothing like the RTI is true there is also intentional content.[10]

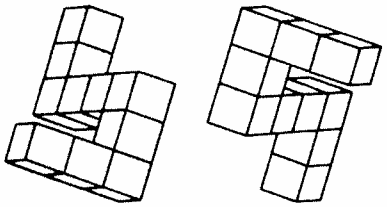

Figure 1.5
Are these the same figure?

*Representation and the Language of Thought*

Representation is often identified with what is really only one kind of representation: quasi-linguistic representation of the sort featured in Fodor's book *The Language of Thought* (1975). But it is at least possible that cognitive states might involve representations of some sort without involving quasi-linguistic formulas type-identified by their status in an internal code with a recursive syntax. In this book, when I mean language-like representations—sentences, or their constituents, written in a brain or in some other physical medium—I will make that explicit. In this connection, it is important to keep in mind that representations may well have propositional contents even though those representations are not language-like, for I take it that an essential feature of the Language-of-Thought Hypothesis—the hypothesis that mental representations are language-like—is that mental representations have a syntax comparable to that of a natural or an artificial language. But it is perfectly obvious that a symbol can have a propositional content—can have a proposition as its proper interpretation—even though it has no syntax and is not part of a language-like system of symbols. Paul Revere's lanterns are a simple case in point.

*Cognition and the Mental*

As is no doubt obvious by now, the use of the word "mental" in the title is misleading, for I will be talking about cognitive systems rather than minds. Some cognitive systems are not



minds (not, at least, as we know minds ostensively), and many aspects of mentality are not cognitive. Cognitive science is founded on the empirical assumption that cognition (hence the study of cognitive systems) is a natural and relatively autonomous domain of inquiry. I shall simply accept this assumption, but a few brief comments are in order.

When we run through mental phenomena as we know them from the human case, many seem inessential in that something could be a mind without exhibiting them. For example, it seems plausible to suppose that a creature could have a mind without having emotions, as is supposed to be the case with Star Trek's Mr. Spock. Descartes held that the essence of mind is thought, Locke that it is the capacity for thought. A system that could do nothing but think might be a rather colorless mind by human standards, but there seems to be something to the traditional idea that such a system would nevertheless *be* a mind. On the other hand, a system that could not think but could feel, have emotions, and so on does not seem to qualify as a mind. If this is right, then what cognitive science proposes is not, after all, very novel; it is just the idea that thinking (and/or the capacity for thought) is the essence of mind and can be studied independently of other mental phenomena.

It is important to be clear about what this hypothesis does and does not accomplish in the way of creating scientific elbow room. It *does* make it possible for the cognitive scientist to ignore (provisionally, at least) such mental phenomena as moods, emotions, sensations, and—most important—consciousness. The hypothesis that cognition is a relatively autonomous domain does not, however, entitle the cognitive scientist to ignore either human psychology or neuroscience. Human beings are the best and only uncontroversial example of cognitive systems we have to study. To try to study cognition without paying attention to how humans cognize would be like trying to study genetics without bothering about biochemistry; some progress is possible, but not a great deal.

Most objections to materialist theories of mind proceed by trying to establish either that a purely physical system could not

be a cognitive system or that a purely physical system could not be conscious. A materialist theory of *cognition* requires a response to the first sort of argument. But materialists, protected by the empirical hypothesis that cognition is separable from mentality generally, can afford to put off responding to the charge that a purely physical system could not be conscious. Perhaps consciousness isn't essential to mind in the way that cognition is.[11] This does not make the problem of consciousness go away, but it does make it, provisionally, someone else's problem.

Since my concern is with thought and not with mental processes generally, it would help to have a term that, unlike "mental representation," suggests only representations that play a role in thought or cognition. "Cognitive representation" isn't too bad; however, for stylistic reasons I will generally stick to the traditional "mental representations." Our questions will be "What is it for a mental whatnot to be a representation (i.e., to have a content)?" and "What is it for a mental representation, a whatnot with a content, to have some particular content rather than some other particular content?"

# Chapter 2

# Mental Representation and Meaning

In this chapter I will take a brief look at the relation between mental representation and meaning generally. Before assessing claims about what it is for a mental state to have a content, it is useful to have some idea of how an account of mental meaning might fit into an account of meaning generally.

## *Original Meaning*

The meaningfulness of some things is often thought to be prior to or more fundamental than the meaningfulness of others. Haugeland (1985, p. 27) writes

> The basic question is: How can thought parcels *mean* anything? The analogy with spoken or written symbols is no help here, since the meanings of these are already derivative from the meanings of thoughts. That is, the meaningfulness of words depends on the prior meaningfulness of our thinking: if the sound (or sequence of letters) "horse" happens to mean a certain kind of animal, that's only because we (English speakers) mean that by it. Now obviously the meaningfulness of thoughts themselves cannot be explained in the same way; for that would be to say that the meanings of our thoughts derive from the meanings of our thoughts, which is circular. Hence some independent account is required.

In this passage, Haugeland expresses the widespread view that meaningfulness generally depends on the meaningfulness of mental states. Mental states, according to this view, have *original meaning*, whereas the meaningfulness of other things (and per-

haps their particular meanings as well) is *derived*, in that they are meaningful, and perhaps have the meanings they have, only because of the meaningfulness and meanings of mental states.

*Neo-Gricean Theories*

Since the pioneering work of Grice (1957), the idea that meaning generally depends on intentionality has come to form the core of a sophisticated theory of meaning and communication. (See especially Schiffer 1982; Bennett 1975; Lewis 1969; Cummins 1979.) Neo-Gricean accounts of meaning proceed in two phases. In phase one, what a speaker (or, more generally, a user, a "meaner") means by some particular performance is explained in terms of the speaker's intentions. According to neo-Gricean accounts of meaning, the intentions with which we deploy a representation determine what we mean by it, and the beliefs others (and ourselves, especially at later times) have about our communicative intentions constitute their (or our) understanding of it. Phase two of the neo-Gricean account explains conventional (e.g., linguistic) meaning by appealing to a shared plan—a convention in Lewis' (1969) sense—for the communicative use of a representational type: *R* means *M* because users of *R* are parties to a convention whereby those who deploy it mean *M* by it. In short, representations have meanings only because their users mean various things by them, and meaning something by a representation is a matter of deploying it with the right intentions. Thus, the semantic properties of representations are derived from the intentionality of their users—either directly, or indirectly via the existence of a convention governing their communicative uses.

Could a neo-Gricean theory apply to mental representations as well as to such nonmental representations as linguistic symbols and stop signs? Neo-Griceans hold that meaning ultimately depends on the communicative intentions of communicating agents. A neo-Gricean theory of mental representation, then, would have to hold that someone or something uses mental representations with the intention of communicating something to someone or something. But a person does not use mental representations with the intention to communicate anything to anyone; indeed, mental representations of the sort standardly featured in the CTC—e.g., a $2\frac{1}{2}$-d sketch or a phonemic representation of a heard utterance—are not used intentionally (or even consciously) at all. Thus, the "communicating agents" required by the theory would have to be subsystems—"sub-personal agents," as Dennett (1978) calls them, or *pro tempore* homunculi (see also Lycan 1981, 1987). These agents would have to have communicative intentions and beliefs in order to mean something by the mental representations they use and in order to enter into conventions governing the communicative uses of those representations.



But this is surely implausible; there is no reason to think that our subpersonal systems (assuming there are such things) *have* beliefs and intentions. Although it is often supposed that subsystems *use* representations in some sense, it is not at all plausible to suppose that they use representations intentionally. Ordinary belief and intention are mysterious enough. We make no explanatory progress by relying on the unexplained and implausible idea that subsystems have communicative intentions and beliefs.[1]

Neo-Gricean theories of meaning can be seen as a species of theory that reduces meaning generally to intentionality. Whereas neo-Gricean theories focus on communicative intentions, there is a tradition, going back to Berkeley and including the later Wittgenstein, that holds that the meaning of a representation is a function of its intended use, where this is construed more broadly than communicative use. The same points just made about neo-Gricean theories apply to the genus generally. They are unpromising as theories of mental representation because they require subpersonal agents with intentions to use mental representations. Thus, "intended-use" theories provide us with no help in explaining mental representation.[2]

*Intended-Use Theories without Intentionality*

The objection to intended-use theories of mental representation is that they implausibly require subpersonal intentional agents.

This objection could be got around if it were possible to get nonintentional states of some kind to play the role that intentions and beliefs play in intended-use theories. Maybe the nested GOALS and PLANS of AI could be made to do the trick.[3] This may strike some as an attractive idea in any case, since the beliefs and embedded intentions required by Gricean analyses are a bit implausible if construed as ordinary beliefs and intentions; certainly people are seldom if ever conscious of having the required intentions and beliefs.

I can't stop to evaluate this idea here, but it is worth pointing out one source of difficulty. It is no accident that Gricean analyses appeal to beliefs and intentions, for these have the same sort of "wide content" (Putnam 1975) as the linguistic and other representations whose contents these analyses seek to explain. If you think that "water" in your mouth means $H_2O$ and not XYZ (the lookalike stuff on Twin Earth), and if you advocate an intended-use theory of linguistic meaning, then you will want your linguistic meanings to be grounded in mental states that have wide content too. Ordinary beliefs and intentions fit the bill, or so it is often claimed,[4] but it isn't at all clear that the data structures of the CTC can be made to fit the bill (and, as we will see in chapters 8 and 10, they probably cannot).

A more plausible line for intended-use theorists is to reduce nonmental meaning to intentionality, and then to either attempt directly to explain intentionality in some naturalistic way or attempt to reduce intentionality to mental representation and try to deal with *that* naturalistically. It is as part of this last strategy that the RTI especially recommends itself to many: Reduce nonmental meaning to intentionality, and then employ the RTI to reduce intentionality to mental representation. But we need to keep in mind that mental representation as supplied by such theoretical frameworks as the CTC may not be able to bear the burden.

### *Symmetrical Theories of Meaning*

The above quotation from Haugeland envisages an asymmetrical treatment of meaning, i.e., a treatment that accords priority



("originality") to mental meaning. But it is possible to hold that mental and nonmental representation are basically the same. (See, for example, Block 1986 and Millikan 1984.) Theories of this kind must reject the Gricean idea that nonmental representation is grounded in intentionality, for if mental and nonmental representation are the same animal, then mental representation will be grounded in intentionality too, and that, as we saw, is implausible at best. Those who advocate a symmetrical treatment of representation will therefore want to hold either that intentionality and representation are simply independent or that intentionality depends somehow on representation. I am not sure that any one currently adopts the first line. Among those who adopt the second line, two different camps can be discerned: Those who, like Quine (1960) and Davidson (1975), hold that belief and desire are somehow parasitic on language, and those who, like Fodor, seek to ground intentionality in mental representation.

### *Grounding Intentionality in Mental Representation*

There are two basic strategies:

*"Localism"* The idea here is to think of each intentional state as grounded in a corresponding mental representation. One can adopt the RTI and then try to attach intentional contents—the contents of beliefs and desires—to mental representations, or one can adopt a modified version of the RTI according to which intentional contents ("wide contents") are the result of subjecting representational contents ("narrow contents") to some further nonpsychological constraint not required for mere representation.

*"Globalism"* The idea here is to adopt a conception according to which one's intentional states are grounded in one's total nonintentional psychological state plus, perhaps, some nonpsychological condition. Dennett holds a view like this, as does (I think) Stalnaker (1984).

*Conclusion*

Philosophy has a lot of roles ready and waiting for mental representation to step into. But whether it can play any of these roles, and if so, which ones, depends on what mental representation *is*. But this question, I contend, can be answered only by examining the scientific theories or frameworks that invoke mental representation as part of their explanatory apparatus. Since there are a number of different frameworks in the running in cognitive science today, we are not likely to get a univocal answer. We won't get a*ny* answer until we focus on some particular framework and start slogging. The remainder of this book tries to get some of the slogging done by evaluating various philosophical accounts of mental representation to see whether any of them will ground the explanatory role assigned to that concept by orthodox computationalism (i.e., the CTC).

We are now ready to turn to the main questions: What is it for a mental representation to have a content, and what determines what content it has? In the context of the CTC, this is equivalent to asking what makes a data structure a *representation*, and what determines what it represents. And let us just remind ourselves once more that folk psychology and the ordinary language of intentional characterization are NOT the topics.

# Chapter 3
# Similarity

*Some Whiggish History*

Several developments in the seventeenth century combined to make the idea that representation is founded on similarity seem difficult to maintain. One of these was the Copernican revolution. Ptolemaics, one supposes, imagined the motions of the planets as they modeled or drew them, and so did their Copernican opponents. But each party imagined matters so differently than the other that, at most, one could possibly have had in mind something similar to the real state of affairs. But then one party or the other (or both) must not have been thinking of the motions of the planets at all! Yet surely the dispute was about the motions of the plants. One party or the other—or perhaps both—*misrepresented* the motions of the planets.

We encounter here for the first time what will be a recurring theme in this book: the difficulty of accounting for *misrepresentation*. The difficulty arises in connection with the similarity view because it seems to make truly radical misrepresentation impossible. Ptolemaic pictures of the planetary motions weren't at all similar to the actual motions, and this seemed to force the conclusion that they were not pictures of the planetary motions but pictures of something else (other Ptolemaic pictures and models?) or of nothing at all. The similarity view seems to allow for misrepresentation only when the dissimilarity is relatively small: If r is to represent $x$ rather than $y$, then $r$ had better be more similar to $x$ than $y$; otherwise, similarity can't be the whole story.

A less famous but ultimately more important development

was Galileo's use of geometry to represent nonspatial magnitudes.[1] Consider a body, uniformly accelerated from rest, that travels a fixed time $t$. When time runs out, it will have achieved a velocity $v$. Now consider a body that travels at a uniform velocity $v/2$ for the same time $t$. It turns out that both bodies will cover the same distance. Galileo's proof of this result involves a revolutionary use of geometry. In figure 3.1 the height $BC$ of the triangle/rectangle $EBC/DCBA$ represents the time $t$. The base $EC$ of the triangle $ECB$ represents the terminal velocity ($v$) of the uniformly accelerated object, and hence the base $DC$ of the rectangle represents the constant velocity ($v/2$) of the unaccelerated object. The area of the rectangle $DCBA$ represents the distance traveled by the unaccelerated object ($vt$), and the area of the triangle $ECB$ represents the distance traveled by the accelerated body.[2] Proof of the result reduces to the trivial demonstration that the triangle and the rectangle have the same area.

What is striking about this use of geometry is that lines represent not trajectories or distances but times and velocities. *Areas*, not lines, represent distances. Nowhere is the path of the object through space represented. Similarity evidently gives us no handle on what makes Galileo's diagram a representation of

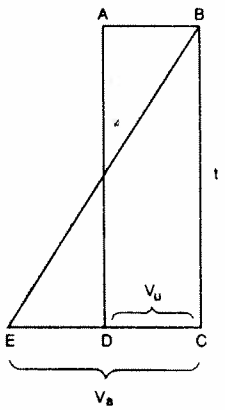


Figure 3.1
Galileo's diagram.



mechanical magnitudes and their relations. What we need is something radically different. The crucial factor seems to be that, given Galileo's interpretation, the laws of geometry discipline the representations and their relations to each other in the same way that the laws of nature discipline the mechanical magnitudes and their interactions.[3] We will return to this important theme in chapter 8.

Descartes put the finishing touches on this story by discovering a way to do geometry with symbols instead of pictures. Descartes' analytic geometry allows us to represent spatial things with equations. Nothing is more obvious than that the Cartesian equation for a sphere doesn't *resemble* a sphere.

As striking as all these examples are, it is possible (just) to dismiss them as cases of nonmental representation on a par with language. After all, it was obvious all along that all representation couldn't be grounded in similarity, since language is an obvious counterexample. There were, of course, half-hearted efforts to see linguistic representation in terms of similarity. But words seldom sound (or look) like what they mean. Still, language and other nonmental cases could be, and generally were, defused by adopting some form of the intended-use theory, leaving original meanings attached to things in the head—images or inFORMed mind stuff—things comfortably dependent on similarity for their status as representations.

For Locke, however, there was at least one scientific development that didn't admit of this otherwise admirable solution: atomism's introduction of the concept of a secondary quality. By Locke's lights, anyway, secondary qualities seem to be explicit cases of mental representation without resemblance (*Essay Concerning Human Understanding*, II, viii). This led Locke to develop an account of mental representation that did not depend on similarity, but on covariance. This idea—an idea that enjoys considerable popularity today—will be the subject of the next chapter.

### *Similarity Critiqued*

Computationalists must dismiss similarity theories of representation out of hand; nothing is more obvious than that data


Figure 3.2
Descartes representing a sphere.



structures don't resemble what they represent. Still, it is worth taking a few pages to rehearse some more general problems with the idea that mental representation is grounded in similarity.

*The Problem of the Brain as Medium*

The most obvious difficulty with the similarity theory is that it seems incompatible with physicalism. If mental representations are physical things, and if representation is grounded in similarity, then there must be physical things in the brain that are similar to (i.e., that share properties with) the things they represent. This problem could be kept at bay only so long as mind-stuff was conceived of as nonphysical. The idea that we could get redness and sphericity in the mind loses its plausibility if this means we have to get it in the brain. When I look at a red ball, a red sphere doesn't appear in my brain. If the ball is a rubber ball, it seems the brain will have to be made of rubber, or at least be elastic. And what about furry tabby cats?

But perhaps we can find a way to get along with less than the real thing. Perhaps something with a kind of *restricted* similarity would do. After all, pictures can represent three-dimensional things without themselves *being* three-dimensional. And isn't pictorial representation—the sort of thing we call "representational art"—grounded in similarity? Of course the nature of the representational medium restricts the kind and degree of similarity that is possible. But that doesn't prevent some representations in that medium from being more similar to some things in the world than others. A cartoon drawing of Sylvester the cat is more similar to Granny than to Tweety Bird, and more similar to Sylvester than to either of those, even though it isn't furry and doesn't chase birds. Cartoon drawings are limited with respect to the kinds of similarity to the world they can exhibit, but they do remarkably well for all that. In principle, anyway, the same point applies to brain processes.

The trouble with this idea is that "restricted" similarity isn't really similarity (actual sharing of properties); it is only "perceived" similarity. When thinking of similarity, it is often useful to ask yourself whether the things said to be similar could literally have the same properties. Cartoon cats cannot resemble cats in

point of furriness, because cartoon cats cannot be furry. Cartoon cats can only *look* furry—*to us*. Cartoon cats manage to represent cats because they *look like cats to us*. Cats and cartoon cats are, up to a point and in certain respects, perceptually equivalent. A cartoon cat in the Sunday comics isn't really similar to a cat in any nonpsychological sense of similarity.[4] The same point applies to brain states: They aren't similar to cats. At best, highly stylized pictures of them might look similar to cats *to us*, in the same way an ink blot or a cloud might look like a cat to us. But this is evidently of no use to the similarity theorist, since perceived similarity is evidently an intentional relation and hence presupposes mental meaning rather than explaining it. Moreover, perceived similarity is useless unless there is something or someone in a position to perceive both the representation and the representandum. But in spite of loose talk of "perceiving" images, it is clear that one does not perceive one's mental representations in the sense in which one perceives cats and red rubber balls.

Of course, a difference in "medium" doesn't rule out all genuine similarity. A clay statue can literally have the same shape and size as a bronze statue. It can even have the same mass. But it cannot have the same mass and the same density, and it cannot have the same melting temperature, and so on. And of course it cannot be made of the same stuff. Once we weed out merely observer-relative "perceived" similarities, it is clear that there isn't a hope of enough genuine similarity's remaining between brain states and the enormous variety of things we represent to underwrite mental representation. When we get down to cases, the idea often doesn't even seem to make sense. After all, what in the brain could literally have the same phonetic properties as a linguistic utterance?[5]

*The Problem of Abstraction*

Even if we ignore the fact that a difference in medium between representation and representandum is bound to rule out all but the most shallow similarities, the doctrine that representation is ultimately grounded in similarity suffers from a serious conceptual defect: Similarity theories cannot deal with abstraction.



To see how this problem arises in a concrete case, suppose that our mental representations are images, and suppose that there is no problem about how images could resemble things in the world. There is still a problem about how images could function as abstract ideas: How could an image of a dog mean any dog whatever, rather than some particular dog (namely the one to which it is most similar)?

As Jonathan Bennett (1971) points out, the problem isn't *completely hopeless*; images can simply be, as it were, silent about certain matters. For example, it is possible to imagine your car without thereby imagining the license plate down to the number and the name of the state. Your image, then, will equally "agree to" any car that differs from yours only in license plate.

The amount of abstraction available from images, however, is limited. We cannot, as Berkeley pointed out in the introduction to the *Principles of Human Knowledge*, imagine a triangle without thereby failing to produce an image that will agree equally to any triangle. Ditto for cats and neckties: Either you imagine stripes or you don't, and either way you're going to miss some of the best cats and ties. So images won't do as abstract ideas—as representations that have, in principle, open-ended extensions.[6]

It doesn't take too much to see that the problem isn't limited to images; anything that is supposed to represent by resemblance is going to suffer the same fate. Indeed, anything *physical* is going to do worse than mental images, because physical things can't simply be "missing" a property; every determinable is going to have some determinate value.[7] Finding a physical object that is equally similar to all cats but more similar to any cat than to any noncat is *conceptually* out of the question. Similarity can't hope to underwrite abstraction, and representation without abstraction is, as Locke pointed out in book III of the *Essay*, not worth bothering about.

The problem of abstract representation is this: How can a representation "agree to" (represent) a whole class of things that differ widely from one another on many dimensions? How, for example, can we represent all and only vegetables? Similarity is no help here, because the brain isn't a vegetable and because

nothing is *only* a vegetable. Anything you happen to pick as a vegetable representation (especially a nonvegetable such as a brain state) will be similar to nonvegetables in a huge number of irrelevant respects. Thus, another way to see the problem of abstraction is this: How do we rule out resemblance in irrelevant respects?

To see how this problem might be solved, consider how simulacra might enter into an account of color recognition. How might we design a system to do the job? As a crude first pass, we might give the system a set of plastic chips of various colors, with color words printed on them. To identify the color of something, the system would find the best match in its supply of chips and display the word. Now, of course this works fine if the system knows to match the *color* of the target to the *color* of the chip. But suppose it is simply a "similarity detector." What is to prevent it from, say, matching its round chips to round targets and its square chips to square targets? After all, it has to *have* such chips if it is going to be able to deal with shape as well as color. A simple solution is to make sure that the only similarities the system can detect are similarities in color. But then what makes the blue chip represent *blue* in this system isn't just the fact that it is blue (and hence similar to blue things); it also depends on the fact that it is used by a device that ignores everything but color. The very same chip, used by a device that ignores everything but shape, represents (say) *round*. Moreover, it is clear on reflection that, even in the color case, the color of the chip is inessential. What is essential is only that something in the system with the word "blue" printed on it should get sent to the display module *when and only when* the system is given a blue target.[8] This is the idea that Locke exploited to develop the core of a theory of representation based on covariance (note the italicized phrase in the previous sentence) rather than on similarity. It is thus no accident that Locke was led to covariance; if you are interested, as Locke was, in the problem of abstraction, there is a natural and compelling route from similarity to covariance. For Locke, the problem of abstraction and the problem posed by secondary qualities lead to the same place.

# Chapter 4

# Covariance I: Locke

## *Plot*

The idea that mental representation is grounded in covariance has recently been worked out by a number of philosophers, most notably Fodor (1987) and Dretske (1981). However, the central thesis—that causal links between mental representations and the world determine the semantic content of mental representations—is widespread. I cannot hope to deal separately with all the important variations on this idea. Instead, I will begin by constructing and criticizing a kind of prototype that I find in book III of Locke's *Essay Concerning Human Understanding*. I think Locke did, in fact, hold something like the theory I will expound, but I don't really care. What I want is a clear and fairly simple version of the sort of theory that founds representation on covariance. The theory I attribute to Locke satisfies this requirement admirably. I am convinced that contemporary versions of covariance theories, including those of Fodor and Dretske, are easily understood and critiqued once we understand the basic flaws in the simple theory I attribute to Locke. The idea, then, is that this chapter will function as a kind of warmup. Getting the basic ideas and moves down pat in this somewhat artificial setting will facilitate discussion of the more sophisticated versions of Fodor and Dretske in the next two chapters.

## *Locke on the Semantics of Mental Representation*

Locke, unlike Berkeley and Hume, saw clearly that representation could not be founded on resemblance. What, then, *does* it rest

on? Locke's answer is that it rests on covariance: Our simple ideas are adequate because they are regular and natural productions in us of external causes. The idea we have when we look at a white thing is an idea of whiteness—a representation of whiteness—because it is the idea white things naturally cause us to have.

Evidently, however, not every case of covariation is a case of representation. Sunburns don't represent exposure to ultraviolet rays. To deal with this problem, Locke had recourse to the following idea: Covariation is representation when the representor (the idea or symbol or whatever) has the right sort of cognitive function. The thing is a representation in virtue of having the right function, and the covariance establishes the specific content.[1]

To see how this works, we need a systematic context—a sketch of a cognitive system—to anchor talk of cognitive functions. (See Cummins 1975a.) To this end, consider Locke's theory of the classificatory use of general words. In book III of the *Essay*, Locke expounds a theory that explains the semantic properties of communicative symbols in terms of the semantic properties of mental representations. For example, on Locke's theory it is the fact that a general word is conventionally associated with a certain abstract idea that gives that term its satisfaction conditions.

Locke was impressed with the tension between two facts: (i) any symbol can have any meaning whatever—words don't fit the world as keys fit locks. (ii) Nevertheless, words can be used incorrectly and falsely. How can (ii) be true, given (i)? How can "cat" be the right word for Graycat, given that the word "cat" doesn't *fit* Graycat any better than any other word? Locke's answer was that when we learn English we learn that, in our language community, the term "cat" is conventionally associated with an abstract idea (concept) that bears a natural, nonconventional semantic relation of *agreement* to all and only the cats. Abstract ideas *do* fit the world as keys fit locks, and words "stand for" abstract ideas in virtue of a purely conventional association.

Locke has given us, or can be construed as having provided, a



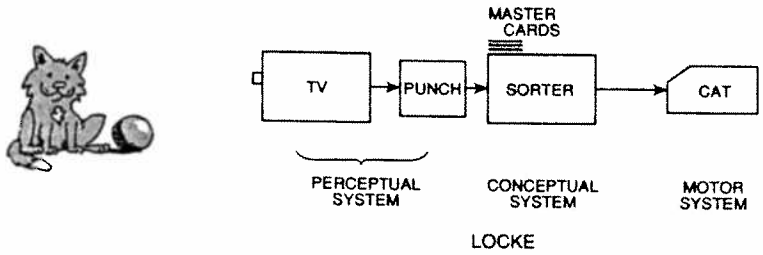


Figure 4.1
The LOCKE machine recognizing a cat.

computational account of the classificatory use of general terms. This becomes obvious if we imagine a concrete instantiation. Consider, then, the mechanical device LOCKE (figure 4.1). LOCKE is equipped with a TV camera hooked up to some input modules (in the sense of Fodor 1983), which in turn are hooked up to a card punch. When the TV camera is pointed at something, a punch card called a concrete idea of sense or a *percept* is produced. Percepts are fed into a sorter, which compares them with a stack of master cards called abstract ideas or *concepts*. When a percept matches a concept—i.e., when the percept contains at least all the holds the concept contains—LOCKE displays the term written on the back of the concept. Any word can be written on the back of any concept; that is a matter of convention. But once the words are printed on the concepts, everything else is a matter of physics. Concepts, of course, can have control functions other than the one just described, and percepts needn't be visual. Moreover, concepts are made from percepts, according to Locke. But enough; what we have will do for the purpose at hand.[2]

Given this sketch of a part of the human cognitive system, we can put the notion of covariance to work to define representation. What makes a given concept the cat-concept is the fact that it is the thing that matches percepts of cats. What makes something a percept of a cat is just that it has some features (some pattern of punches) that percepts come to have in the system when, only when, and because the system is in perceptual contact with a cat. Cats cause Locke's perceptual system to generate percepts with

a characteristic punch pattern. When it finds (or constructs) a master card that matches that pattern, it writes 'cat' on the back, because that is the pattern that identifies the presence of cats to the system and hence the pattern wanted as the meaning of 'cat'. (It does this, we may suppose, by a kind of trial and error, trying various words on various cards until it is able to substantially avoid error messages from its peers.) If there is a pattern of punches that shows up on percept cards when, only when, and because the TV camera is pointed at a cat, then that pattern, wherever it occurs in the system, represents cats. Being a cat representation is being something that is, in perceptual contexts, a litmus test for cat presence. For future reference, the idea is briefly expressed as follows:

(L1) *x* represents *y* in LOCKE $=_{df}$ *x* is a punch pattern that occurs in a percept when, only when, and because LOCKE is confronted by *y* (whiteness, a cat, whatever).

There is, of course, a problem about how to spell out "confronted by" in a non-question-begging way, for the only thing that looks a sure bet is to say that LOCKE is confronted by a cat, or whiteness, just in case a corresponding percept is produced. But a corresponding percept is just one that has the right representational content. I suppose the best strategy is to pass the buck to the psychophysicists, trusting them to identify in some nonintentional way some causal conditions sufficient for percept production. In practice this is likely to be circular, since the only way psychophysicists are going to discover such conditions is by correlating them with the "corresponding" percepts. But in principle (philosopher's friend!) it doesn't have to go that way, as the example of LOCKE shows; we can simply correlate a punch pattern *qua* punch pattern with a set of conditions sufficient to produce it (given proper functioning of LOCKE).

Notice how the theory works: If something with the right role in the system—the right function—covaries with something else, then we have not only representation but also a specific content. Locke's theory begins with the plausible (perhaps inevitable) idea that the things that mediate cat recognition in the system



must be the cat representations. To put this idea to work, we have had to sketch enough of a functional analysis of the recognition system to identify the relevant things: punch patterns on percept cards. This is surely the right way to solve the Problem of Representations. But the theory goes farther; it proceeds to read off a solution to the Problem of *Representation*—*viz.*, L1.[3]

The essential points about the theory, from Locke's point of view, are (i) that it does away with resemblance as the ground of representation and (ii) that it solves the problem of abstraction. Let us take a moment to understand how this is accomplished.

*Resemblance avoided* In discussing a simple color-recognition system in the last chapter, we encountered the following problem: How is the system to avoid matching the round blue chip to round targets instead of blue ones? The obvious solution is to design the system so that it is insensitive to everything but color. But then it is easy to see how to make resemblance drop out of the picture, for what matters is only that the system produce something with 'blue' printed on it in response to blue things. Whether that something is itself blue is quite irrelevant; the causal origin and the functional role of the thing—the fact that it gets produced by blue things and the fact that it drives the "speech" system (and other motor and cognitive systems) appropriate—are what count.

*Abstraction achieved* Once we cease to think of the relation between representation and representandum in terms of similarity and begin to think in terms of covariance, the problem of abstraction has a simple solution. A master card (concept) that has a pattern of punches that occurs in a percept when and only when the system is confronted by blue has something that will match (have the same punch pattern as) every adequate percept of a blue thing, and in that sense will "agree with" (Locke's term) each and every blue thing. No such solution is available to the resemblance theorist, because nothing can resemble all and only the blue things. But something can be the "regular and natural effect" of blue on the system, and hence occur in the system's percepts when and only when blue is present to it.

*Misrepresentation*

The fundamental difficulty facing Lockean theories is to explain how misrepresentation is possible. To see why this is a difficulty, try to describe a case of misrepresentation. Suppose LOCKE is confronted by Graycat but generates a dog-percept (i.e., a percept with the feature *D*). Then it is not true that *D* occurs in a percept when, only when, and because a dog is present, since no dog is present and the current percept has feature *D*. Hence, *D* doesn't represent doghood, and LOCKE has not generated a dog-percept, contrary to hypothesis. LOCKE cannot misrecognize Graycat as a dog—not because LOCKE is so clever, but because misrepresentation is an incoherent notion given L1, the target theory of representation. Since it is possible (indeed inevitable) to sometimes misrecognize cats as dogs, something must be wrong.

Lockeans, I think, have just one way of dealing with this problem: idealization.[4] This can take one of two forms: idealizing away from malfunctions and idealizing away from suboptimal conditions of perceptual recognition.

*Malfunctions and Misrepresentations*

It is tempting to regard misrepresentation as something that arises from malfunction: Perhaps if LOCKE were functioning properly, it wouldn't misrecognize Graycat as a dog. We can exploit this idea by defining representation as follows:

> (L2) *x* represents *y* in LOCKE $=_{df}$ were LOCKE functioning properly, punch pattern *x* would occur in a percept when, only when, and because LOCKE is confronted by *y*.

L2 allows for misrepresentation because it makes having a representational content a modal property of punch patterns—a property a punch patter can have even if LOCKE never succeeds in recognizing something corresponding to that content. Perhaps it always malfunctions when confronted by cats. Nevertheless, it could still be true that *were LOCKE to function properly*, pattern *C would* occur in a percept when, only when, and because the system is confronted by a cat. Given this revision, it isn't actual covariance that matters; it is the covariance that would obtain



were LOCKE functioning properly. Perhaps, like many AI systems, LOCKE seldom functions properly.

Given our focus on the CTC, the trouble with this response to the problem of misrepresentation is that, according to the CTC, the most obvious and everyday cases of perceptual misrepresentation—*viz.*, the illusions—are not cases of malfunction but cases of proper functioning in abnormal circumstances. What happens is that the normal functioning of the system in an abnormal situation results in a misrepresentation. For example, subjects looking into the Ames Room (figure 4.2) misrepresent the relative heights of things in the opposite corners. But the problem isn't that the visual system suddenly breaks down in some way when one looks into the Ames Room; the problem is

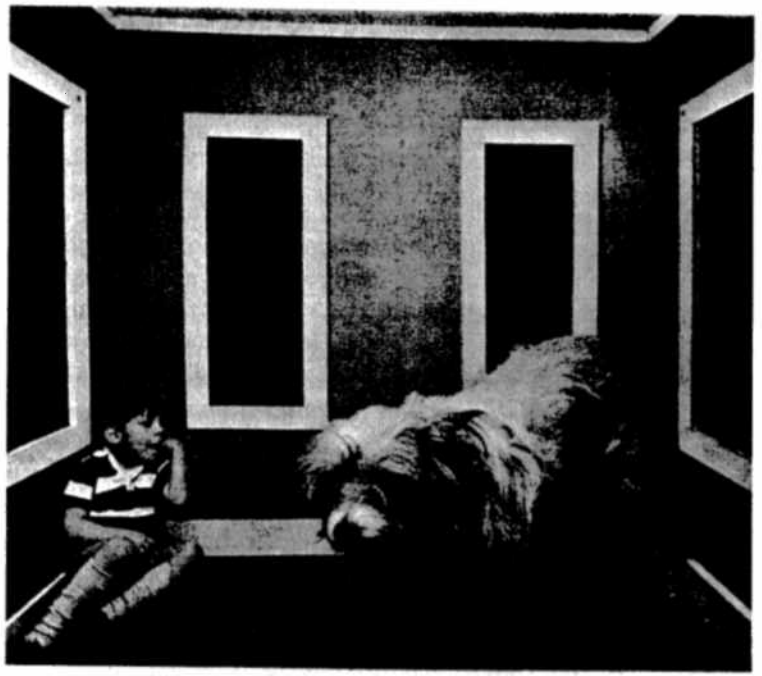

Figure 4.2
The Ames Room. The dog is really much smaller than the child!

rather that the visual system computes the relative heights of the things in the room from, among other things, the assumption that the room has square corners.[5] The same principle holds even more obviously in purely cognitive cases; for example, the detective who draws the most rational conclusion given the available evidence may yet arrest the wrong person. In such a case, normal functioning—even optimal functioning—guarantees misrepresentation if the evidence is inadequate in some way.

If cognition rests on computation, as the CTC assumes, then there is an important respect in which error is essential to a well-designed cognitive system: The computational problems faced by a system with finite resources—especially memory and time—can succeed only by taking short-cuts. Such a system must employ algorithms that rest on fallible assumptions—for example, that objects in space are rigid (Ullman 1979), that corners of a room are "square," that the future will resemble the past in the respects chosen by conceptually salient features, that other agents are rational and in fact know what they are in position to know, or that objects don't come in transparent pairs (hold up your finger in your field of vision and focus on something beyond it), etc.

Traditional epistemology typically attempts to idealize away from resource constraints. Research in AI strongly suggests that this is a false idealization: When you try to add in resource constraints afterward, you always wind up redesigning the system from scratch. Epistemology for God and epistemology for us are two different things. God never had to worry about recognizing tigers *in time to evade them.*

*Ideal Circumstances for Perception*

Assimilating misrepresentation to malfunction, then, yields a concept of misrepresentation that undermines computationalist explanations of misrepresentation. Still, reflection on the critique just rehearsed suggests another cure. The core of that critique is that misrepresentation often occurs as the result of proper (even ideal) functioning in *less-than-ideal circumstances.* Misrepresentation seems (in these cases, anyway) to be due to a departure from



*ideal circumstances.* This suggests that we revise the definition as follows:

> (L3) *x* represents *y* in LOCKE $=_{df}$ were LOCKE functioning properly and circumstances ideal, *x* would occur in a percept when, only when, and because LOCKE is confronted by *y.*

L3 evidently allows for truly radical misrepresentation of the sort imagined in Cartesian Demon scenarios: If all my perceptual states are caused directly by the Demon, then conditions are never ideal. But it is still possible to represent cats, say, because it might still be the case that, *were* conditions ideal, the relevant pattern would occur when, only when, and because a cat is present. I emphasize this point in order to make it clear that L3 (and L2, for that matter) accommodates misrepresentation by going modal and thereby putting meanings in the head.[6]

Not only is this a natural way for the account to bend under pressure from misperception cases; it is really the *only* way it can bend. The *essence* of the position is that something is a representation of a cat in virtue of having some feature that is, in percepts, an effect of cat presence and not of anything else. It has to be something that occurs in percepts because a *cat* is present. If it occurs because something else is present—a clever cat robot, or a dog, a raccoon, or a koala bear—then the account is going to attach the wrong content to the punch pattern in question, with the result that nothing will count as a cat representation. But no occurrence in a perceptual system has a chance of being the effect of cats (or anything else interesting) *exclusively* unless conditions are ideal.[7] Under *real* conditions, error is the price you *must* pay for computational tractability.

The obvious first question that L3 invites is whether it is really possible to assimilate *all* misrepresentation to failures of one sort of idealization or the other, i.e., to improper function or to less-than-ideal "circumstances." My own view is that it is not possible. I will pursue this point shortly. For the moment, I want to pursue a different sort of objection: When combined with a fundamental empirical assumption of the CTC, L3 leads us in a circle and is therefore incompatible with the CTC.

The assumption in question is that cognitive systems manage to get into states that reliably covary with distal features of the

environment because of their representational resources: What the system does is *infer* the distal situation from current data (proximal stimuli, if the problem is perceptual) and *a great deal of knowledge stored as data structures*.

To see why this sorts ill with L3, we need to scrutinize this business of ideal circumstances. According to the CTC, what is likely to be involved? Under *what* conditions is the system likely to produce percepts with features that reliably covary with some *distal* feature? The CTC has it (indeed, this was the fundamental claim of the so-called cognitive revolution) that cognitive systems are able to get into states that reliably covary with distal features because of their stored knowledge.[8] For LOCKE, what this means is that, in addition to good lighting and that sort of thing, the perceptual system is going to have to have access to a rich fund of *knowledge about what sorts of distal features are likely to produce which sorts of signals at the output end of the TV camera. The idea (*THE* idea) is that the system is able to reason from the TV-output (transduced proximal stimuli) and its fund of *knowledge to a conclusion about the responsible distal feature.[9] The system, in fact, executes a program that has access to a representation of the transduced proximal stimuli and to all this *knowledge, and that program computes a representation of the distal feature. That representation, in turn, drives the card-punch, which produces a percept. This, at any rate, is the story the CTC directs us to tell. Thus (to echo Fodor), if we are after the notion of representation that underwrites computational explanations in psychology, we had better take this story seriously.

For present purposes, we can sum up the implications of the story as follows: *If the percept is to be adequate, the mediating *knowledge had better be adequate too.* Of course, the transduced proximal stimulus has got to be high-grade as well. That will require a properly functioning TV camera, and good light, and appropriate distances and angles, and so on. But all that won't be nearly enough. A big part of what must be the case if the occurrence of $x$ in a percept is to covary with the occurrence of $y$ in the environment is that the mediating *knowledge must be there and must be adequate. No matter how good the stimulus



and how well the mechanism functions, suboptimal mediating *knowledge—a pack of lies, for instance—is going to make it impossible for LOCKE to produce percepts with features that reliably occur when, only when, and because there is whiteness, or a cat, present. But it follows from this that the relevant notion of "ideal circumstances" to which F3 appeals is, in large part, a matter of the system's having the right *knowledge—i.e., the right representations; representations with the right content. And *that* means we cannot fill out L3 without making liberal use of the very notion that L3 is supposed to explain.

It is worth belaboring this point a bit. A computational system of the sort favored by the CTC has no serious hope of arriving at the truth about even very common *perceptual* matters without the help of a formidable background of *knowledge. It is fundamental to the computational approach to perception that perceptual systems must make use of a very considerable and sophisticated base of *knowledge about the world, including its own "specifications," in order to construct reliable percepts. Language perception is the most celebrated case, but any perceptual system that solves the problem of perceptual constancy is essentially the same; a central claim of the CTC is that the only hope of mapping proximal stimuli onto distal stimuli is to use *knowledge of how proximal stimuli are generated to arrive at a "best hypothesis" concerning the distal situation. To define representation in terms of the optimal functioning of such a system is to presuppose the very notion one is trying to define, for such systems are *specified*, in large and essential part, by the tacit *knowledge they embody (i.e., by their representational resources). According to the CTC, perceptual and other cognitive systems are able to generate reliable indicators of distal features because of their cognitive resources—that is, because they are representational systems. If you define representation in terms of the ideal behavior of a certain kind of system, you must be prepared to specify the kind of system you have in mind. But the CTC holds that there is no way to specify a system that has a hope of reliably indicating the sort of facts we are capable of representing without making liberal use of the notion of representation.

That, to repeat, is what the cognitive revolution and the defeat of behaviorism was originally all about. A program isn't enough; to understand such things as speech perception you need to specify the relevant *knowledge (data) structures. Indeed, it is hardly exaggerating to say that, from a CTC perspective, the problem of (say) speech perception just *is* the problem of discovering what *knowledge is required, and in what form it must exist to mediate the required inferences.[10]

It might seem that the Lockean doesn't owe us an account of ideal circumstances.[11] The Lockean says, in effect: "Being a representation—having a content—is essentially a matter of having the right sort of function. Which content a representation has is determined by what its tokening in the system would covary with under ideal conditions. Thus, what you do to ascribe content is point to the right sort of thingamabob—a punch pattern in a percept card, say—and ask what would covary with the occurrence of that thingamabob if circumstances were ideal. Why isn't that clear enough?"

It is clear enough as far as it goes, but it doesn't go very far. We might concede a kind of formal correctness to the definition, but it has no explanatory value except insofar as we have some conception of what is meant by ideal circumstances. To see this—to see that the explanatory value of L3 depends on what conception of ideal circumstances one has—just consider the default conception one *does* have, viz., that conditions are ideal when they are such as to guarantee (or maximize the chances of) success. On this conception, circumstances are ideal for perceiving (say) cats only if the system, when confronted by a cat, produces (or is maximally likely to produce) a representation with the content CAT. This understanding of ideal circumstances plainly renders L3 circular. So, evidently, if L3 is to tell us anything useful, we must bring some other conception of ideal circumstances to bear. Moreover, it must be a conception that does not depend on a prior understanding of the notion being defined, or of *any* other semantic/intentional concept, since Lockeans typically propose to use mental representation to explain all that other stuff.



What can this conception be? It cannot be the default conception, as we have just seen. And, as we saw earlier, it can't be the one that falls out of the CTC either, for that conception relies heavily (as does everything that falls out of that theory) on the very notion of representation we are trying to explicate. My own view is that these exhaust the plausible alternatives; hence my claim that L3 leads us in a circle when combined with the CTC. The Lockean wants to explain representational content in *S* by reference to the covariance that would emerge if things were NICE FOR *S*. This helps only if we understand what it is for things to be NICE FOR *S*. The difficulty is that the CTC gives us formulations of what it is to be NICE FOR *S* that make use of the very notion of representational content that the Lockean is trying to define.

Of course, Lockeans won't give up that easily. They have, I think, two more cards to play. One is a kind of semantic reductionism, and the other depends on the notion of inexplicit mental content (i.e., mental content that is not the content of some representation). These don't represent *plausible* alternatives, but that remains to be argued. Let us take them in turn.

*Semantic Reductionism*

The situation is this: The Lockean needs to tell us under what conditions LOCKE will be able to punch a certain pattern—the *c*-pattern, let's call it—into a percept when and only when confronted by a cat. Under normal conditions, LOCKE will not be able to do this. It is no mean feat, after all. LOCKE needs all the help it can get. Computationalist theories all agree about what sort of help LOCKE needs: lots of the right *knowledge. But if Lockeans go that route, they render their account circular.

To avoid being circular, Lockeans must specify ideal conditions in a way that does not presuppose content assignments to states of the cognitive system. They cannot, therefore, appeal to all that *knowledge. Thus, it is natural for a Lockean to ask what can be achieved *without it*. What sort of perceptual successes can one expect the system to achieve *in complete *ignorance*, as it were?

The inevitable move is some version of reductionism. We "begin" with simple perceptual features. A simple perceptual

feature is, by definition, the representation of something the properly functioning system cannot be mistaken about (given the right lighting and so on) precisely because it is a representation whose construction is immune to influences from whatever *knowledge a system might have. Simple perceptual features are, in fact, direct correlates of transduced proximal stimuli; they represent properties that can be transduced.[12] For these cases, L3 works as it stands. We then move on to "complex features." Constructing these *does*, of course, require the mediation of *knowledge, but that is OK because we have explicitly provided for some (or something out of which it can be built) by providing for simple perceptual features. And so on.

This reply avoids the objection, but at a considerable two-part price:

> (i) There have to *be* simple perceptual features, i.e., perceptual features that represent properties that can be transduced.
>
> (ii) Percepts the construction of which requires mediation by *knowledge must require only such *knowledge as can ultimately be expressed solely in terms of representations of simple features. The punch pattern for CAT must be a superposition of punch patterns that represent simple perceptual properties.

It is worth emphasizing that (ii) must be interpreted in a strongly reductionist way. Under ideal conditions, the system must be infallible. Confronting whiteness must be nomically sufficient and necessary for the occurrence of the w-feature in percepts. Hence, the transduced proximal stimulus, plus *knowledge, plus nonpsychological laws of nature must *entail* (not just make highly probable; not just reliably indicate) that there is whiteness out there. Remember the "when and only when" in L3. "When": If a cat occurs and the *c*-pattern doesn't occur, then the possibility exists that only orange cats, or only Graycat, excite the *c*-pattern. "Only when": If the *c*-pattern occurs sometimes when it is a dog out there, then there is no principled reason not to say that the *c*-pattern represents CAT-or-DOG. Thus, the concept CAT must *reduce to* concepts that apply to simple perceptual properties—i.e., to proximal stimuli.



Good luck. The literature since Descartes is littered with bankrupt programs that found this price too high.[13] If you want to get all your content out of representations of simple perceptual properties, you are welcome to try; however, you would do well to keep in mind that this strategy has a dismal track record. That is good enough for me; I don't propose to rake it all up again.

*Inexplicit Content: An Alternative Reply*

As we saw in chapter 1, natural and artificial information-processing systems can be semantically characterized—characterized, in fact, in terms of propositional contents—even though the propositional content in question is not explicitly represented in the system. I call the object of such characterization *inexplicit content* to distinguish it from content that is explicitly represented in the system.

Inexplicit content is "in" the system without being represented in it. It is thus open to a Lockean to claim (with little plausibility, as we will see) that a cognitive system doesn't require *knowledge to mediate perception. It does require content of a sort, of course, but nothing *explicitly represented* is required. The relevant facts about the system are facts to be specified in terms of inexplicit content. Since inexplicit content is not represented content, a definition of representation that presupposes inexplicit content is not circular or regressive. This reply blocks the critique just leveled against L3, for it demonstrates that in specifying ideal conditions for perception we *can* presuppose contentful background states of LOCKE so long as the presupposed content is *inexplicit*.

Empirically, this is not a very plausible idea, as I said a moment ago. Such things as the rigidity and continuity assumptions exploited in vision (Marr 1972; Ullman 1979) may well be implicit in the architecture of the visual system in some way (Pylyshyn 1984, p. 215).[14] Much of the information that a perceptual system brings to bear on a particular perceptual problem is unlearned and fixed. But much is not. Language perception is a good case in point. The ability to perceive the phonemes, words, phrases, structures, etc. of one's language is, to a large extent, acquired.

Foreign speech sounds like rapid, continuous, unorganized noise, but this changes drastically as you learn the language. Now, the CTC accounts for learning—as opposed to other kinds of psychological change (maturation, trauma, disease)—as the result of the acquisition of new *knowledge. Changes in architecture (program) don't count as learning, for they are not computationally driven. Thus, if acquiring a new language is learning (as it seems to be), it is not, according to the CTC, a matter of acquiring a new architecture, and hence it is not something to be explained in terms of changes in inexplicit content. The perceptual skills involved in understanding speech are therefore mediated to a significant extent by *knowledge. Much the same goes, I suspect, for other domains. The cases of perception mediated only by inexplicit content probably do not go very far beyond the cases of simple transduction.

But it doesn't really matter; even if we concede that perception is mediated only by inexplicit content and not by explicit representations, we will have saved the letter but not the spirit of Lockean covariation theories. Lockean theories are supposed to explicate what it is for a cognitive system—its states, processes, or whatever—to have semantic properties. The assumption is that cognitive representations are the fundamental bearers of such properties. If Lockean approaches are construed so as to presuppose inexplicit content, they fail to address the fundamental problem they are designed to solve: the problem of what it is or something mental to have a semantic property.

*Covariation and Inexplicit Content*

But perhaps we can work out a Lockean approach to the problem of inexplicit content. If so, and if we can get around the fact that perception mediating *knowledge is often learned and hence not inexplicit, it could still be maintained that mental content is ultimately grounded in covariance.

Inexplicit content is part of what Pylyshyn (1984) calls the biologically fixed functional architecture.[15] It isn't something that comes and goes in the system, at least not as the result of cognitive factors. It is, therefore, essential to a particular cognitive system; change the inexplicit content descriptions and you



have described a different cognitive system (though perhaps one that is realized in the same biological system). Given this, if we are going to make use of the idea of covariation we are going to have to trade on the idea that a certain kind of functional architecture occurs when and only when the world exhibits a certain feature, or when and only when a certain condition obtains.

This is plainly going to fail for artificial computational systems, for we are constantly building systems whose architectures embody horribly false assumptions. Every logical bug is a case in point. What is more serious, every program that falls victim to the frame problem or fails to capture the flexibility of human reasoning is a case in point. Every time we build a system that fails in some way because it is programmed wrong (rather than merely misinformed), we instantiate an architecture that embodies false assumptions. It is, to say the least, difficult to avoid this. That, in part, is what makes AI a challenging empirical discipline.

I think we should be impressed by the obvious hopelessness of a covariance account of inexplicit content in artificial systems, for it seems clear that anyone who accepts the CTC must suppose that appeals to representation have just the same explanatory role in artificial systems as in natural ones. That, in fact, is one way of stating a fundamental assumption of computationalism. Thus, if an account of representation doesn't work for artificial systems—if, in fact, it is patently silly for such systems—then it isn't an account of the concept of representation that underlies the CTC.

This, by my lights, is enough to kill Lockean accounts of inexplicit content in the context of the CTC stone dead. Nevertheless, I am going to ignore the problem raised by artificial systems and push forward with the discussion of natural systems, because I think something interesting emerges.

If we are going to make use of the idea of covariation, then (as I said above) we are going to have to trade on the idea that a certain kind of functional architecture occurs when and only when the world exhibits a certain feature, or when and only when a certain condition obtains. What this gives us is something like the following (assuming, for now, propositional contents):

(L4) *S* has (embodies?) an inexplicit content with truth condition $C =_{df}$ the sort of functional architecture *S* exhibits occurs (persists?) iff *C* obtains.

Thus, for example, an architecture inexplicitly embodies the rigidity assumption just in case architectures like it occur (persist?) if and only if the rigidity assumption is in fact satisfied.

One would have to be a wildly enthusiastic adaptationist to believe this, even about biological systems. Surely satisfaction of the rigidity assumption isn't sufficient for the occurrence of the relevant architecture. I suppose the assumption is satisfied on Mars, but I'm quite sure the architectures in question don't occur there. Nor is satisfaction of the assumption necessary for the occurrence of the architecture; lots of interesting biological features occur that aren't adaptations. If this happens in the cognitive realm—and I don't see any reason to suppose that it couldn't—then the architecture could occur in environments that don't satisfy the assumption.

To get around this, the Lockean will have to resort to the old idealization trick: Perhaps under ideal evolutionary conditions, . . . .[16] Still, this may look promising; after all, *adaptation* isn't an intentional notion, and the mechanisms responsible for the occurrence of a certain kind of architecture do not depend on the mediation of *knowledge, and that looks like progress. I suppose it is progress, but it is progress down the wrong road.

The problem is that the sort of covariance envisioned by L4 just isn't what is behind inexplicit content. What makes it appropriate to describe the architecture of the visual system in terms of (e.g.) the rigidity assumption is, minimally, that *the system wouldn't work if the assumption didn't hold*. If things seen didn't generally remain more or less rigid under spatial transformation, the system would constantly misrepresent things. *That* is why it makes sense to say that the assumption is, as it were, built into the architecture. It is wired up to operate as if it were reasoning from *knowledge that included the rigidity assumption. The vision program exploits the constraint in that its proper operation presupposes that the constraint is satisfied.



The evolutionary story is plausible only because we know that a system with the architecture in question will work well only if the rigidity assumption is approximately satisfied, for the evolutionary story depends on the idea that such architectures will not survive—will not be replicated over many generations—unless the conditions for their working well are met. This is a pretty dubious idea, even under the assumption of ideal evolutionary conditions (whatever that may come to); but that is not my point. My point is that the evolutionary story assumes that a system with the architecture in question will work well only if the rigidity assumption is satisfied. But if we have assumed that, we have assumed all we need to assume for the relevant inexplicit content; the evolutionary story *presupposes* the inexplicit content attribution! Covariation, and the evolutionary scenario that allows us to trot it out, simply drop out as irrelevant. I don't know if we should count this as a circularity in L4, but I do think it renders L4 intellectually uninformative. It just can't help you understand what it is to have an inexplicit content unless you already have what it takes.

Before we leave this, there are two final points to be made. The evolutionary story depends on the idea that only systems that work well will persist. But, first, systems will occur that do not persist. What of *their* contents? Second, in this context, working well means getting the right percepts constructed, and that clearly presupposes the notion of representational content.

*Idealization and Infallibility*

A number of pages ago, I promised to return to the question whether we can really assume that a cognitive system would be infallible under ideal conditions.

There are well-known philosophical reasons for resisting this assumption. If you take this line, you have to be prepared to *legislate* against alleged cases in which the truth differs from the result of ideal inquiry, and that means you have to adopt some form of verificationist anti-realism. You must, in short, be prepared to say that what isn't ideally detectable isn't *there*, and this looks more like arrogance than serious theory.

One needn't rely on this philosophical line of attack, however, for there is an uncontroversial empirical objection to the assumption. As we saw in our discussion of malfunction, error is the inevitable price of computational tractability. The *knowledge that mediates cognitive inferences is, of necessity, only typically and approximately true. Some bodies aren't rigid. Some rooms aren't square. Some noses *are* concave (see Gregory 1970). What is more, there is no idealizing away from this kind of error. If you want to consider a system with unlimited time and memory, you are going to be considering a system with a completely different functional architecture than the one that operates under real resource constraints.

When you take friction and air resistance away from a pendulum, you still have a pendulum. Furthermore, you have a pendulum whose period depends on its length in just the way in which period depends on length in "normal" pendulums. The independence of the effect of length on period from the effects of friction and air resistance is what makes it proper to idealize away from the latter. But when you reduce the resources required by an infallible program, what you typically get is not a program that performs acceptably but not optimally; what you typically get is a program that fails to perform at all, or one that performs very poorly. Typically, then, this infallible program (if there is one) is just a different program, root and branch, than the one that makes things tractable given limited resources by making simplifying assumptions. Computational work in early vision is a striking example of this general point. There are algorithms that will solve many of the computational problems infallibly, but they require unrealistic resources. Progress was made by turning to algorithms that rely on assumptions that, although they are fallible, hold quite generally in normal environments (Marr 1982; Ullman 1979).

Thus, the idea that one can idealize away from cognitive error is incompatible with a fundamental finding of the CTC. That theory holds all such idealizations to be fallacious on the ground that they violate the requirement that what one idealizes away from must be independent of what is left. According to the CTC, then, an ideal but finite cognitive system operating under ideal conditions will inevitably make lots of mistakes. Since L3 assumes the contrary, L3 is incompatible with the CTC.



*Summary*

It looks, tentatively, as if computationalists cannot understand mental representation in terms of covariation. In a way, we should have seen this coming: We're going to have covariance only when the epistemological conditions are right. Good epistemological conditions are ones that are going to get you correct (or at least rational) results. Conditions like that are bound to require semantic specification. Less obviously, but just as surely, covariance theories presupposes a kind of epistemological idealization that is forbidden by the CTC. In the next two chapters, we will see whether the most prominent contemporary variations on the basic Lockean theme manage to resolve these fundamental difficulties.

# Chapter 5

# Covariance II: Fodor

Jerry Fodor (1987) defends an account of the nature of mental representation that is remarkably similar to the one I have just discussed. The similarity is no accident; it will become clear as we go along that covariationists have a limited number of basic tools in the box. But there is no question that Fodor has added a few that are worth examining.

*Background*

Fodor begins by assuming the Representational Theory of Intentionality (which he calls the Representational Theory of Mind) and the Language-of-Thought Hypothesis (the hypothesis that mental representations are language-like symbols). Given these two assumptions, we can assume further that the problem of mental meaning generally reduces to the problem of understanding what it is for a primitive, nonlogical term of Mentalese to have a content. Given this focus, it will be convenient to have a convention for naming terms of Mentalese. In what follows, I shall write the term in Mentalese supposed to denote horses as |horse|; absolute values seem appropriate.

The basic idea—an idea Fodor calls the crude causal theory—is that symbol tokenings denote their causes and symbol types express the property whose instantiations reliably cause their tokenings.[1] Two problems immediately arise: that some noncats cause |cat|s, and that not all cats cause |cat|s.



*The Disjunction Problem*

Fodor calls the first problem the *disjunction problem*, for reasons that will emerge shortly. Suppose we try to describe a case of misrepresentation. A case of misrepresentation has to be a case like this: (1) Graycat causes a |dog| to occur in *S*; (2) |dog| expresses the property of being a dog in *S*; (3) Graycat is not a dog but a cat (of course). Now, since a cat (or, anyway, Graycat) causes a |dog| to occur in *S*, it follows that what |dog| must express in *S* is the property of being a dog-or-cat, or perhaps being a dog-or-Graycat, contrary to (2). It seems that any reason the crude causal theorist has to think that |dog| misrepresents Graycat as a dog is, for that theorist, a better reason to think that the content of |dog| has been misdescribed. Misrepresentation is always upstaged by a redescription of the alleged content. When the redescription is carried out, there is no misrepresentation. Hence, the crude causal theory implies that there is no misrepresentation.

It is tempting to reply that the causal route from Graycat to |dog| is not reliable. However, we can always make it reliable by describing the case in enough detail: There must be some situation in which Graycat reliably causes a |dog| in S—namely, the situation that obtained when, by hypothesis, Graycat was causally responsible for a |dog| in S. Moreover, there is such a thing as *systematic* misrepresentation: If I systematically misrepresent shrews as mice, this must be a case in which, according to the crude causal theory, shrews reliably cause |mouse|s in me. But there can't *be* such a case, since whatever is reliably caused by shrews is supposed to be a |shrew|.

*Idealization*

The obvious solution to the disjunction problem—one that Fodor himself briefly favored—is to idealize: In S |cat|s express the property of being a cat if, under ideal or optimal conditions, cats would reliably cause |cat|s in *S*. This move is, of course, familiar from our discussion of LOCKE, and it suffers from the same flaws: If the CTC is on the right track, there is no thoroughly naturalistic way to spell out the ideal conditions in question, and

they won't eliminate error anyway.

Fodor in fact favors a different and far more ingenious solution.[2] The account pivots on the following claims: The fact that shrews sometimes cause |mouse|s in me depends on the fact that mice cause |mouse|s in me. On the other hand, the fact that mice cause |mouse|s in me doesn't depend on the fact that shrews sometimes cause |mouse|s in me. Mice look mousey to me, and that mousey look causes a |mouse|. But it is only because shrews also look mousey to me that shrews cause |mouse|s. Thus, if mice didn't cause |mouse|s, shrews wouldn't either. But it needn't work the other way; I could learn to distinguish shrews from mice, in which case mice would cause |mouse|s even though shrews would not.

This applies to the disjunction problem as follows; |mouse|s don't express the property of being a mouse-or-shrew, because the shrew-to-|mouse| connection is *asymmetrically dependent* on the mouse-to-|mouse| connection—the former connection would not exist but for the latter. In the case of genuinely disjunctive concepts, however, *A*-to-|*D*| connections are on a par with *B*-to-|*D*| connections, so |*D*|s express the property of being *A*-or-*B*.

*Objections to Asymmetrical Dependence*

I find this line unconvincing. Consider again the crucial counterfactuals:

(i) If mice didn't cause |mouse|s, shrews wouldn't cause |mouse|s.

(ii) If shrews didn't cause |mouse|s, mice wouldn't cause |mouse|s.

The alleged asymmetry depends on the claim that (i) is true and (ii) false. But is this really right? Shrews cause |mouse|s because they look like mice. Thus, if shrews didn't cause |mouse|s, that might be because (a) shrews didn't look like mice or because (b) mouse-looks didn't cause |mouse|s. If (b) were the culprit, though, mice wouldn't cause |mouse|s either, and that would make (ii) true.

It might seem that we can't blame (b) because the closest world in which shrews don't cause |mouse|s is the one in which (a) holds, not (b), since (b) requires a break in the rather central connection between mouse-looks and |mouse|s, whereas (a) requires only learning to distinguish mice and shrews. But this really isn't very persuasive. Perhaps shrews just look like mice to people, and finding out about shrews just makes them *uncertain* when they see either one. In a case like that, anything that will break the shrew-to-|mouse| connection will break the mouse-to-|mouse| connection as well. Even experts might perform randomly (perhaps the technology isn't adequate), even though they understand the difference perfectly well and can explain it to laypersons. Look what doctors do with diseases, or psychiatrists with psychoses and neuroses!



A variation on this theme suggests that the theory of asymmetrical dependence inverts the explanatory order: |mouse|s are wild when caused by shrews not because the more basic causal connection is with mice, but because |mouse|s express the property of being a mouse—something they might well do even if the dependence were symmetrical. Consider this story: In a certain tribe, all the youngsters are taught that they must catch a mouse for a certain potion the tribe needs. Mice are very rare, but only mice will do. Like all the other children, Broomhilda is taught how to catch a mouse (but not how to make the potion; only the medicine woman knows that). She is taught this by practicing on shrews. She has never seen a mouse, and she wouldn't recognize one if she saw one. Perhaps a mouse hasn't been seen in generations. Broomhilda knows there is a difference, however, for she knows at least this: Mice work in the potion, and shrews don't. Since the whole point of the training is to catch a mouse, the shrew-to-|*S*| connection (|*S*| is Broomhilda's internal representation) wouldn't exist but for the mouse-to-|*S*| connection. |*S*|s are, as Millikan (1984) would say, reproduced in Broomhilda because of the connection with mice. But also, given the way things are learned, the connection between |*S*|s and mice wouldn't exist if it were not for the connection between shrews and |*S*|s. There is no saying which connection is more fundamental. Hence, the asymmetrical-dependence doctrine must hold that |*S*| expresses the property

of being a shrew-or-mouse. But it doesn't. It expresses the property of being a mouse, and *that* is why |S|s occasioned by shrews are wild.

A determined defender of asymmetrical dependence might avoid this criticism by claiming that scenarios like the ones just rehearsed that break down the asymmetry between (i) and (ii) are scenarios in which |mouse| (or |S|) is no longer a primitive term of Mentalese. But I don't think this will do, for it is pretty obvious that you can cook up similar scenarios for, say, |puce|.

Fodor's own reply is that the asymmetrical-dependence condition must apply *synchronically:* No matter how |mouse| and |shrew| are learned, current dispositions make the mouse-to-|mouse| connection primary. This strikes me as rather *ad hoc,* but let's see where it leads.

The picture Fodor has in mind is shown in figure 5.1. Mice cause mousey looks, which cause |mouse|s. Since shrews look mousey, they also cause mousey looks, thus poaching on the causal route from mice to |mouse|s and producing "wild" |mouse|s. Here are the relevant counterfactuals:

(1) If mice didn't cause |mouse|s, shrews wouldn't either. (T)

(2) If shrews didn't cause |mouse|s, mice wouldn't either. (F)

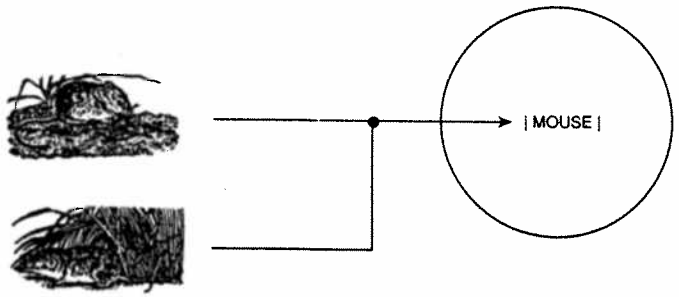


Figure 5.1
A shrew poaches on the mouse-to-|mouse| connection.



As indicated, we have the required asymmetrical dependence when (1) is true and (2) is false.

Start with (2). Figure 5.1 suggests two ways to break the shrew-to-|mouse| connection: (i) Mousey looks don't cause |mouse|s. But then mice won't cause |mouse|s either. Thus, (ii) shrews don't cause mousey looks. Perhaps shrews are extinct. More realistically, perhaps I come to *know something (perhaps tacitly) about how shrews and mice differ, and, as a result, shrews cease to even *look* like mice. But *mice* still look mousey, and hence they cause |mouse|s. So (2) is false, as desired.

Now consider (1). Again figure 5.1 suggests two ways to attack the mouse-to-|mouse| connection: (iii) Mousey looks don't cause |mouse|s. Since, by hypothesis, shrews poach by looking mousey, they will also cease to cause |mouse|s, and (1) is true, as required. Unfortunately, this way of making (1) true makes (2) true, as we just saw, so (iv) mice don't cause mousey looks. Perhaps mice become extinct, or acquire some disfiguring disease. But this won't affect the shrew-to-|mouse| connection, so (1) is false, contrary to requirements.

It looks as though the only way to have (1) true and (2) false is to employ different rules for evaluating them: Use (ii) to evaluate (2) and (iii) to evaluate (1). The possible worlds in which (1) is true and (2) is false are not the same possible worlds. To put it another way, there is no single interpretation that makes (1) true and (2) false. Therefore, (1) and (2) do not jointly express something about the connection between mice and shrews and that between mice and |mouse|s.

One might reply: "Well so what? All that means is that the definition of asymmetrical dependence is a bit messy. You have to say how (1) and (2) are to be (separately) evaluated." I wish I had a knock-down rebuttal to this reply, but I don't (even though I have the feeling there must be one). All I have is this: If you must get this tricky with the counterfactuals, you don't have a philosophical *explanation* any more; at best, you have a technically defensible equivalence between analysandum and analysans. It is hard to believe that the content of |r| is *mouse* rather than *mouse-or-shrew* BECAUSE

> if mice didn't cause | mouse | s because mousey looks didn't then shrews wouldn't cause | mouse | s either, and if shrews didn't cause | mouse | s becauseshrews didn't look mousey then mice would still cause | mouse | s (*ceteris paribus*, of course).

Maybe there *is* a way to make asymmetric dependence work without sacrificing explanation, but enough. As Fodor quite rightly says, the disjunction problem is the lesser of the two problems faced by the covariance theorist. Let's stop counting angels on pinheads and move on to where the action is.

*Omniscience*

> The Crude Causal Theory says, in effect, that a symbol expresses a property if it's nomologically necessary that *all* and *only* instances of the property cause tokenings of the symbol. (Fodor 1987, p. 100)

Lots of cats never cause | cat | s. (Well, to be safe, lots of rocks never cause | rock | s. But I prefer to stick with the cat-and-mouse game.) But why is that a problem for the covariationist? The crude causal theory was expressed thus: "... symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings." It isn't obvious that *this* says, in effect, that *all* instantiations of the property cause tokenings of the symbol. Wherefore this strong and bothersome *all*? Granted, some cats don't cause | cat | s; but so what? Why isn't it enough that nothing else causes | cat | s (or, rather, that nothing else causes | cat | s in the basic way supposedly picked out by asymmetrical dependence)? It is well to get clear about this, because this seems to commit Fodor to the claim that cognitive systems are omniscient, and, as he admits, this is preposterous on its face. "... [P]roblems about the 'all' clause are, in my view," he writes, "very deep." So why is the 'all' clause *there*? Surprisingly, Fodor never answers this question, but the answer is quite simple: If some cats don't cause | s | s, then it seems that the extension of | s | should be the subset of cats that *do* cause | s | s. We need to rule out the possibility that | s | expresses the



property of being, say, a black-and-white cat, or that of being Graycat. The only way the causal theorist can get around this is to insist on genuine covariation: All cats cause | s | s (or, anyway, any cat *would* cause an | s | *if given a fair chance*). But what is it to be *given a fair chance*?

The difficulty, of course, is that according to the CTC there is a fair chance that a cat will cause a | cat | only if the system is prepared to *attend* properly and to make the right *inferences* (or *inferences) on the basis of the right **knowledge*. But this sort of reply is clearly out of bounds; it will render the theory circular. Fodor realizes this but argues that, contrary to appearances, it *is* possible after all for a computationalist to specify causally sufficient conditions for a cat to cause a | cat |, or even for a proton to cause a | proton |, without trafficking in intentional or semantic notions.[3] Here is what he says:

> But though protons typically exert causal control over | proton | s via the activation of intentional mechanisms, a naturalistic semantics doesn't need to specify all that. All it needs is that the causal control should actually obtain, *however* it is mediated. The claim, to put it roughly but rather intuitively, is that it's sufficient for | proton | to express *proton* if there's a reliable correlation between protons and | proton | s, effected by a mechanism whose response is specific to psychophysical traces for which protons are *in fact* causally responsible. And that claim can be made in nonintentional, nonsemantic vocabulary. It just was.
>
> No doubt mechanisms that track nonobservables in the required way typically satisfy intentional characterizations (they're typically inferential) and semantic characterizations (they work because the inferences that they draw are sound). But that's OK because, on the one hand, the semantical/intentional properties of such mechanisms are, as it were, only contingently conditions for their success in tracking protons; and, on the other, what's required for | proton | to express *proton* is only that the tracking actually be successful. For purposes of semantic naturalization, *it's the existence of a reliable mind/world correlation that counts, not the mecha-*

*nisms by which that correlation is effected.*[4] (Fodor 1987, pp. 121–122)

We have seen this move before; it is just the idea, scouted in the chapter 4 above, that the covariationist doesn't really owe us an account of the conditions under which, say, an *arbitrary* cat is guaranteed to produce a |cat|. All we need is (i) some guarantee that the relevant mechanism exists and (ii) a non-question-begging way to *pick out* that mechanism. The first part is plausible enough on general empirical grounds: There must be some circumstances in which cats are sufficient for |cat|s. And for Fodor the second is a cinch: "The mechanism that does the trick" does the trick! This is because all Fodor requires is a "naturalistic" way to pick out the mechanism, i.e., a way of picking out the mechanism without explicit use of intentional or semantic terms:

> What is required to relieve the worry that meaning will resist assimilation into the natural causal order is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form '*R represents S' is true if C* where the vocabulary in which condition *C* is couched contains neither intentional nor semantical expressions. (Fodor 1984a, p. 2)

Fodor says that avoiding semantical and intentional expressions is only a *minimal* requirement, but in fact he takes it to be sufficient:

> The reference to 'mechanisms of belief fixation' perhaps makes this look circular, but it's not. At least not so far. Remember that we're assuming a functional theory of believing (though not, of course, a functional theory of believing that p; . . . .). On this assumption, having a belief is just being in a state with a certain causal role, so—in principle at least—we can pick out the belief states of an organism without resorting to semantical or intentional vocabulary. But then it follows that we can pick out the organism's mechanisms of belief fixation without recourse to semantical or intentional vocabulary: The mechanisms of belief *fixation* are, of course, the ones whose operations eventuate in the organism's having belief. (Fodor 1987, p. 105)



Perhaps we can pick out the mechanisms of belief fixation in "naturalistic" terms, but the CTC holds that we can't understand them or describe them without a healthy dose of representational lingo.

Well, admittedly, one philosopher's (or one scientist's) explanation is another's explanadum, but this seems like cheating to me. We are told that representation rests on a covariance between representation and representandum—between cats and |cat|s, for example. Covariance, in turn, is grounded in a mechanism that, under the right conditions, will produce a |cat| from a cat. According to the CTC, the mechanism in question can be understood only by appeal to inner representations, for the mechanism in question is one of inference from stored *knowledge. It follows that in order to understand the mechanism that the CTC invokes to explain the covariance between cats and |cat|s we must already understand representation and the explanatory role it plays in mental mechanisms. And that, by my lights, is enough to undermine the power of covariance theories to help us to understand the nature of representation in the CTC.

The problem, of course, is that it isn't enough to avoid intentional/semantic vocabulary; you must do it in a way that explains what representation is. It becomes obvious that just avoiding intentional/semantic vocabulary isn't enough when you see how easy it is. The problem, remember, was to say under what conditions cats are sufficient for |cat|s,[5] and to do it in naturalistic vocabulary. But look how easy it is: (i) Find an actual occasion in which a cat does cause a |cat|. Name that occasion *O*. (ii) Consider the mechanism that did the trick on occasion *O* (never mind how this worked, or whether it was peculiar to *O*), and call it *M*. (iii) Construct the desired counterfactual: Were *M* to operate on a cat in circumstances like those that obtained in *O*, a |cat| would result. Nothing to it!

The thing starts to come unraveled when we ask what *O* and *M* are like, for it is a fundamental consequence of the CTC that these must be *understood* inferentially (though, of course, they *can* be

*picked out* naturalistically). The covariationist tells us that there is representation because there is covariance. The CTC tells us that there is covariance because there is representation, and Fodor agrees. But you can't have it both ways without undermining the explanatory power of one of the two doctrines. And since the philosophical problem before us is to explain representation in a way that will underwrite (not undermine) its explanatory role in the CTC, it is the covariationist doctrine that must go.

Here is a kind of analogy that may help clarify how I see the intellectual situation: Suppose someone tells you that the temperature of something depends on the amount of caloric in it. "What is caloric?" you ask. "Well," says your informant, "it is clear what one would like to say: Caloric is the stuff that increases in a thing when you raise its temperature. Of course, that's circular. But I can avoid the circle. Consider the mechanism that operates when you put tap water from the tap marked "C" in a pan on a lighted stove: Caloric is the stuff that mechanism causes to increase in the water." This identifies caloric without explaining it.

*Idealization Again*

We saw in chapter 4 that covariationists require idealization away from all sources of error. We are now in a position to put this point together with the point about circularity. The fact that you can't idealize away from error means that there is no *general* way to pick out a mechanism that will produce a |cat| in response to an arbitrary cat. Thus, the only way to do it is by reference to some *specific* instance or instances in which a cat *does* product a |cat|. We then say for all S that if S were in a situation like *that*, a cat would yield a |cat|. The sense that we no longer have an explanation of representation can be traced to the demonstrative. The account is essentially ostensive. "Representation," it says, "is when you have a case like *that*." Then you give an example or a sketch of what one would be like: "You know. It's like when you think there is a cat there because there *is* one there." There is no substantive way to specify the *C* in "In *C*, any cat would cause a cat in *S*," so the covariationist must, in the end, have recourse to ostension, and must hope you don't notice that there is no principled way to generalize on the example.

# Chapter 6

# Covariance III: Dretske

*The Account in Knowledge and the Flow of Information*

For present purposes, the account of the nature of representation as set out by Fred Dretske in his 1981 book *Knowledge and the Flow of Information* can be boiled down to the following two claims:

> (D1) The semantic content of a cognitive state *M* is a privileged part of its informational content, *viz.*, that informational content of *M* which is nested in no other informational content of *M*.[1]
>
> (D2) A cognitive state *M* of *O* has the proposition *p* as an informational content if the conditional probability that *p* is true, given that *O* is in *M*, is 1.

On this view, informational content is explicitly a matter of covariation between the representing state and the state represented. Indeed, Dretske often glosses D2 as the claim that *M* is a perfect indicator of the truth value of *p*. Perhaps it is worth emphasizing that, on this view, as on Fodor's and Locke's, *M*'s covariation with *p*'s holding isn't merely evidence that *M* has *p* as its informational content; it is constitutive: Representation *is* a special case of covariation on these accounts.

*Misrepresentation*

Notoriously, Dretske's account gives rise to difficulties in explaining the possibility of misrepresentation. It follows from D2 that if *p* is the informational content of *M*, then *p* is true. Hence, by D1, if *p* is the semantic content of *M*, *p* is true. It looks as if there can't be a false representation.