

# **Model-Based Science and the Representational Theory of Mind**

**Peter Godfrey-Smith**

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

(Version M, December 2004. Under review.)

## **1. Introduction**

## **2. Model-based science**

## **3. Structure of the "basic representationalist model"**

## **4. The roles of representationalist description**

## **5. How can this be news?**

## **1. Introduction**

Over the past 30 years, one topic much discussed in the philosophy of mind and philosophy of psychology has been the status of "the representational theory of mind," or "RTM." As usually conceived, the representational theory holds that the mind operates (in part) by creating, storing, and using internal representations of objects and events in the world.

This is generally seen as an empirical theory, but one surrounded by an unusual number of philosophical problems. So a philosophical literature has grown up around the following questions: Exactly what sort of claims are made by the RTM; what does it commit us to? What sort of data should be seen as evidence for the theory? Can the semantic properties of inner representational states be understood in a scientific way? How might mental representations and their semantic properties figure in the explanation of behavior?<sup>1</sup>

The present paper aims to present a different way of looking at this family of issues in the philosophy of mind, by making use of some ideas from the philosophy of science. The main point of the paper is easily stated. Suppose we use the term "representationalism" in a philosophically neutral way, to refer to the commitment to mental representation that is seen in much contemporary psychology and cognitive science.<sup>2</sup> Philosophers generally talk of representationalism as a "theory", and this view is so familiar that it has given rise to the standard acronym, "RTM," which was used above. I suggest that it might be more accurate and useful to see representationalism as a commitment to a model of the mind instead.<sup>3</sup>

When I make this claim I have in mind a specific sense of the term "model," and a specific set of contrasts between "theories" and "models." The term "model" is used in a wide variety of ways in both science and philosophy of science. Some common usages of "model" and "theory" will not generate the contrasts I have in mind. Sometimes the terms are used synonymously, or a "model" is seen as a deliberately simplified theory. In other work, the logician's concept of model is used to give a highly general analysis of the structure of all scientific theories. On yet other occasions, however, the term "model" is used in connection with a particular scientific strategy. In this sense, we can talk of "model-based science," in a way that contrasts with other kinds of scientific work. I will argue that representationalism is best seen as a case of model-based science. This puts many of the standard philosophical questions surrounding representationalism into a new light.

I add some disclaimers before proceeding. First, this is not a paper about the status of folk psychology (belief-desire psychology), or a folk-psychological casting of representationalist ideas. In addition, it will not say much about the relation between representation and computation. Third, I will not be discussing theories of the specific semantic contents of representations. Instead, in this paper the focus is on the idea of mental representation itself, and its role within the sciences of the mind.<sup>4</sup>

## 2. Model-based science

This section will sketch the ideas in the philosophy of science that are used in the rest of the paper. The aim is to describe a particular kind of scientific work, a particular strategy for organizing theoretical ideas. I will call this model-based science (or occasionally, model-based theorizing).

My sketch of these ideas derives from one prominent strand in recent philosophy of science. But it is important to distinguish the strand that I am following from some near-relatives. One literature that focuses on "models" is a family of views known as the "semantic view of theories" (Suppes 1960, Suppe 1977, Van Fraassen 1980, Lloyd 1988). These views aim to give a general account of the structure of scientific theories, to replace the obsolete "syntactic" account associated with logical empiricism. Several senses of "model" have been used in this literature (Downes 1992), some similar to, and some quite different from, the sense employed in the present paper. And more importantly, the aim of those discussions has been to give a fully general account of theories, rather than an account of how some, but not all, fields and scientists operate. In that sense, the "semantic view of theories" is not employed in this paper.<sup>5</sup> But some philosophical work in and around this tradition makes available a different option (Cartwright 1983, Wimsatt 1987, Giere 1988, Downes 1992, Morgan and Morrison 1999, Weisberg 2003 and forthcoming.)

According to this alternative view, there is a variety of approaches to theorizing in science. One of these approaches is to organize theoretical ideas via models, in one sense of that term. Model-based science operates by constructing and exploring hypothetical, usually simplified, systems that are intended to have some relevant resemblance relation to some real-world "target" system that we are trying to understand.

A useful account of the structure of this kind of work can be found in Giere (1988) – though Giere, again, is too inclined to see his description as a fully general account of scientific theorizing. For Giere, theoretical description of world involves a two-step process. The first step is the specification, by means of language or some other symbolic medium, of an imaginary system. I will call these "model systems". The second step is the presentation and discussion of claims of resemblance between the model system and some aspect of the real world.<sup>6</sup>

This two-step account makes possible a contrast with what we can call a more "direct" strategy. Here the scientist tries to describe, as directly as possible, what are taken to be the components of the target system, and how they operate. When following the more direct approach we simply try to say how things are, what is going on. Such work can still be abstract, and can omit many details in an attempt to get to fundamental properties. But there is no side-track through imaginary or hypothetical systems, and hence no need to worry about resemblance relations. Weisberg (forthcoming) calls this approach to theorizing "abstract direct representation."<sup>7</sup> This is a somewhat distilled version of the more familiar and traditional view of scientific theorizing that philosophers of mind have often assumed in their discussions of representationalism, and that I contrast with model-based science.

For our purposes, many key features of model-based science stem from the role of the resemblance relations between model systems and targets. Resemblance relations are often seen as suspicious in philosophy, because resemblance is seen as a context-sensitive, vague, quirky sort of relation. All those suspicious features of resemblance relations are indeed part of the picture here. But rather than being problematic, I see them as highly functional within model-based science. A distinctive feature of model-based work is that one can develop and explore a model system that has useable similarities to a target while being unclear, indefinite, and changeable about exactly which features of the model are supposed to resemble features of the target, and unclear or changeable about the degree of resemblance intended. The indirect route by which model-based science seeks to represent the world gives rise to a kind of flexibility in the theoretical constructs that result. It is a source of both strengths and weaknesses in model-based science.

For example, one scientist might regard a given model simply as an input-output device, as a predictive tool. Another scientist might use the same model to deal with the same target system, but treat the model as a faithful map of the inner workings of the target system. Both scientists are hoping for a resemblance between model and target, but they are looking for very different kinds of resemblance. Between these extremes there are many intermediate kinds of resemblance that might be sought; a third scientist might see the model as resembling the target with respect to basic structural features, but not with respect to details. There can be constant shifting here, and a good deal of vagueness.

(In later sections I will, for shorthand purposes, sometimes talk of "high fidelity" versus "low fidelity" applications of models, but this should not be taken to suggest that most uses are determinately one or the other, or that there is a single scale.)

Note that flexibility is the emphasis here, not wholesale instrumentalism. Models of this kind can be applied and understood in a strongly realist way; they can be seen as accurately capturing the unobservable inner workings of a complex system. In these cases, models can yield genuine understanding of how the target system works. They can be the basis for explanations of why it does what it does. But the explanatory power of a model, in relation to a given target, is affected by the standards of resemblance that are in play. If the model is treated purely as an input-output device, for example, it cannot be used to explain the workings of the system in the same rich sense.

Models can be used in these diverse ways within theoretical science, but models themselves do not make claims about the way the world is. Models do not have, or determine, a definite set of truth conditions. Rather than making such claims on its own, a model can be employed to make claims about how things are, via specification of a target system and a suitable notion of resemblance. But, as I have been emphasizing, there is often a lot of slippage in the resemblance relation, even when people are trying to be self-conscious about it. So even then, the claims being made about the world are often vague ones. Theories, in the sense usually employed in the philosophy of mind, have semantic contents and make reasonably definite commitments about how things are. Models, on their own, do not.

As a consequence of all these facts, familiar philosophical questions about the reference of terms and the reality of theoretical entities present themselves differently when we are dealing with model-based science. In model-based science there are straightforward, direct semantic relations between the scientist's language and entities within the model. But the relations between the scientist's language and particular aspects of the real-world target system are indirect, and mediated by resemblances between the two whole systems. The term-by-term or entity-by-entity questions that are often asked in discussions of scientific realism become (even more) problematic in model-based science.

That completes my sketch of the ideas from the philosophy of science that will be used in this paper. But before moving on I add three comments and qualifications.

First, I note the role of a kind of shorthand form of description, used in this section and below, that is linked to difficult issues about models themselves.

The simplest case to think about, which I sketched above, is the case where a single imaginary system is specified in some detail and compared to a more complex real-world target. But talk of "a" model in model-based science is usually talk of something that seems to be more abstract and general, not a single imaginary system but apparently a large class of them. When biologists talk of "the" one-locus model of natural selection, they do not have in mind a single imaginary population. Lots of imaginary populations with different fitness relationships, and so, on qualify as cases of "the" one-locus model.<sup>8</sup>

So the short-hand that I have in mind is the practice of talking of "a" model system when something more general is meant, a collection of such imaginary systems, constructed according to a common pattern. The larger issues that arise here concern what model systems really are – how to handle their ontological status and individuation.

According to the view I prefer, a model is often a large class of imaginary structures, unified by abstract similarities, often built from a common stock of elements or ingredients. So model systems themselves are regarded here as imagined concrete things, where "concrete" is opposed to abstract. Model systems are imaginary physical, biological, or social systems, or collections (large and small) of these things. The contrast I am making is with a view of models as purely abstract mathematical objects.<sup>9</sup> However, like many others, I am not sure how best to treat the ontological status of model systems themselves. The details do not, I hope, matter to the main arguments of this paper.

The second is a more minor point. Often, a model that is used to deal with an unknown system has its origin in a familiar or well-understood system. This will, as it happens, be the case in my treatment of representationalism in the next section. But this drawing on the familiar is not essential to all model-based science (an assumption that has hampered some earlier discussions).

Thirdly, it is a difficult question whether there is a reasonably sharp distinction between the two approaches to theorizing discussed in this section, or something more

vague and continuous. The resolution of this issue depends on difficult questions about scientific language and practice. I will proceed here as if the distinction was reasonably sharp. This itself is probably an idealization.

### **3. Structure of the "basic representationalist model"**

My suggestion is that representationalism in contemporary psychology, and similar disciplines, generally operates as model-based science. I am more confident that this is the case than I am about the analysis I will give of how this works. My proposals in the latter area are more cautious.<sup>10</sup>

In this section, I first describe a structure that I call the "basic representationalist model." Some representationalist work uses little more than this core structure, but the core can also be the basis for various kinds of elaboration – the specification of much more complicated structures that can function as model systems when trying to understand particular cognitive capacities.

In describing the "basic" model, I take as our starting point a familiar everyday sense of "representation," and a particular way of talking about representations. In this sense, a representation is one thing that stands for something else. Another way of putting it is that a representation is a thing whose state is consulted as a guide to something else: the state of X is consulted in dealing with Y. This sense is familiar from ordinary discussion of public representations like maps, graphs, and so on. It is perhaps the central sense of the term "representation" in the everyday domain.

This sense is quite often, though not always, made central in the philosophical literature. For example, I see this as the concept that is at the heart of the treatments of representation in Millikan (1984), Hagueland (1991), and Cummins (1996), though all handle the concept somewhat differently and augment it in various ways.<sup>11</sup> Gallistel (1990, 2001) has expressed similar views from within cognitive science. My approach is to start with that very familiar concept of representation, and place in a different overall philosophical context. Roughly, I treat it as the core of a model rather than as part of a theory in the usual sense.<sup>12</sup>

The core of the model, as I said above, is the idea of an agent's consulting the state of X in guiding behavior towards Y. This seems to involve three separable components. First, there is some representation-producing mechanism or process. Second, there is the representation itself, the thing that is consulted. Third, there is a reader or consumer mechanism, which consults the representation in the guidance of behavior or further processing of some kind.

In contrast with Millikan, I do not treat the roles of the representation-producing mechanism and the reader or consumer in a symmetrical way. As I see the basic model, it is not required that the production of the representation be "adapted," specialized, or purposive, although it may well be. My phrase "representation-producing mechanism or process" above is intended to allow for production by more happenstance events. The role of the reader or consumer, however, is more specific and constrained (as is discussed in the next section). So my sketch of the model incorporates something like the idea that a representation is whatever is read like one, treated as one, consulted as one, even if its production was uncoordinated with this consultation. This asymmetry I see as a piece of somewhat contested folk lore about representation.

Millikan, Gallistel and some others also explicitly build into their concepts of representation a constraint on how it is that X is used to guide behavior towards Y: they claim that this works via an abstract mapping or resemblance relation between X and Y. It is hard to work out how substantive this requirement is, but in any case I do not see this as part of the basic representationalist model.<sup>13</sup> (In cases where such resemblance between X and Y is present, it is natural to call the inner representation a "model" of the external condition it is used to deal with. In this paper, to avoid confusion, I never use the term "model" for an inner structure of that kind. I only talk of models when discussing the scientific status of representationalism itself.)

I call the structure outlined above the "basic representationalist model." Representationalism is the application of this model, and its more elaborate relatives, to the internal workings of the mind.

The basic model, as described above, is very minimal. It is so thin, in fact, that it is hard to see how specific representational contents might be determined. Two factors that we might use to address this issue – constraints on the causal production of

representations, and resemblance relations – have been made inessential. It is not even clear how general questions of "direction of fit" would be settled. These issues are not discussed in this paper. My main points can be made, I hope, without addressing them.

Some representationalist work employs little more than the basic model, and tries to get some explanatory purchase with this simple set of ideas. The early representationalism of the psychologist E. C. Tolman (1948) is an example. Tolman criticized "stimulus-response" models of behavior and, in contrast, proposed that at least some animal behavior results from the construction and consultation of inner "maps" of the environment. For Tolman, the mechanisms behind the construction and use of the maps were treated largely as black boxes, though he was committed to a basically empiricist picture.<sup>14</sup> Tolman's emphasis was on how an inner map-like structure might make possible adaptive behavior in novel circumstances, in a way that stimulus-response mechanisms do not.

More recent representationalist work goes much further, of course. Once we have in place the basic idea of consulting inner representations, this gives rise to the possibility of modelling the ways in which these representational states can be constructed and modified, by means of natural processes. For many, the idea of computation is the key to this question, and the result is a tradition of work on the manipulation of inner representational states by computational processes. This work results in the construction of much more complex representationalist models than the structure outlined above. But the "basic" model is, I suggest, at the heart of these more elaborate models.

So according to my view, the representationalist description of a real-world system is based on a hypothesized resemblance between the internal workings of the system and some version (basic or elaborate) of the representationalist model. I emphasize, again, that the idea that representationalism is a case of model-based science is distinct from the analysis I give of how this works. It would be possible to develop a model-based analysis in which, for example, there is no "basic" or "core" model, but only a set of family resemblances between rather different structures.

#### **4. The roles of representationalist description**

How can we tell whether or not representationalism in psychology operates as model-based science? What sort of evidence is there for my account?

One hope I have is that once this set of ideas has been put on the table, those who are familiar with the science will immediately see the appeal. Much of philosophy of mind has operated for decades with a very simple set of assumptions about relevant parts of philosophy of science. (The assumption that the main goal of science is the discovery of "law-like generalizations," is another aspect of this situation.) Much of the theoretical side of psychology, and much of the rest of cognitive science, works explicitly as model-based science. The currency used in most theoretical discussion is the model.

I do not suggest that the frequency with which psychologists say "model" is a sure indicator of the proportion of model-based science in their field, however. As was noted earlier, the term "model" is used in diverse ways, and conversely, one can engage in model-based science without overtly speaking "modelese." It is a subtle issue whether and to what extent a field engages in model-based science, and whether a particular theoretical idea is being employed in a way that exemplifies this scientific strategy.

So to assess my claims we must see whether the framework used in this paper makes good sense of the mindset and practice in psychology and related sciences, whether it seems to yield an accurate account of how the science works.

One mark of model-based science that I have been emphasizing here is a certain kind of flexibility in the employment of theoretical ideas. Scientists can agree on which model to use while disagreeing in both obvious and very subtle ways about the kind of resemblance the model is supposed to have to its target. This kind of flexibility is constantly seen in representationalist cognitive science.

I will give a quick example here, and a more detailed and central one towards the end of the section. My quick example is the Bayesian model of belief change and learning, as used in psychology. The Bayesianism model is set up in terms of representations of possible states of the environment, and it models the updating of degrees of belief in response to incoming evidence. For some, Bayesianism is seen as a hypothesis about the inner mechanics of reasoning. For others, it is a sort of idealized theory of overall competence, not a structure to look for in the nuts and bolts of the mind

(Tenenbaum and Griffiths 2001).<sup>15</sup> Some say it is a normative model, but still treat it as a model that gives us some empirical purchase on how minds work. For others still (perhaps for most), Bayesianism is something that we have reason to think is a good model, though it is unclear what sort of mapping between the model and inner mechanisms to envisage. More bluntly, people are fairly sure it's a good model, but not at all sure what it's a good model of.

This combination – some degree of consensus about models but less consensus about their status – is common. In day-to-day work, a basic form of representationalism is employed as a framework within which empirically assessable alternatives are expressed. Within the framework, people can contrast and investigate hypotheses about what is represented, what information is used by various kinds of cognitive processes. If the foundational status of this kind of talk is challenged, the response is often that the framework has been seen to guide work in a fruitful way, and we do not have a better option. Such a defence is compatible with considerable flexibility on the richness of the resemblance relation that people are expecting to find, when neuroscience gives us a lower-level story.

Scientists could, in principle, handle a scientific theory of the more traditional kind in a "flexible" way as well. But in the case of a theory that is supposed to have definite content, only some ways of handling the theory will be in accordance with what the theory actually says. The other uses will be derivative or deviant. Van Fraassen's (1980) account of how a theory might be accepted without being believed is a well-known example of how this kind of special handling of a theory might work. In the case of representationalism, I suggest that there is no evidence of a distinction between a "normal" handling of representationalism that accords with the theory's real commitments, and deviant handlings that involve special attitudes and stances. Instead, there is a large range of acceptable construals, surrounding the models that organize day-to-day work.

My analysis also makes sense of a striking sociological fact recently emphasized by Matthew Barrett (2004). It is common for textbooks and other works in psychology and cognitive science to open with an endorsement of representationalism about the mind. But this initial endorsement is rarely accompanied by a serious attempt to say what

internal representations are. If a definition or explanation is given, it will be unhelpful and, for a philosopher, circular or chronically slippery.<sup>16</sup>

So what are all these authors assuming? Are they being coy, or confused? The idea of a model derived from public representation use that can be flexibly applied is helpful here. I think the authors are assuming not that we already know exactly what mental representations are like, but that we know what structures the idea of mental representation is being modelled on, and that we will pick up, in the course of the discussion, what sort of possible resemblance between internal processing and familiar cases of representation use is being envisaged by the author. The author in this situation does not want to lay down a genuine definition of mental representation ahead of time, unless it is a definition that builds in lots of slippage. The author wants flexibility to be retained. The result has been puzzlement on the part of philosophers. Philosophers want to know exactly what this notion of mental representation is supposed to be, but scientists can seem strangely insouciant about the issue.

Barrett's own diagnosis of psychologists' reticence about properly explaining what representations are supposed to be is different from mine. He thinks that a representationalist commitment functions much of the time as little more than an expression of anti-behaviorism.<sup>17</sup> Beyond this rejection, "representationalist" work has little in common.

As a matter of sociological fact, it is hard to deny that these loose and weak uses of representationalist description exist. And it seems clear that at the present time we are very far from what some philosophers imagined would happen as cognitive science developed – that the concept of representation would settle into a definite and unified scientific usage, as well-grounded in empirical work as the concept "oxygen" is. The question is exactly how far we are from that situation. I would say that present-day work exhibits a mixture of roles for representationalist description, with plenty of minor differences across fields and subfields. Some representational talk is mostly, as Barrett says, "ornamental." Some is used to indicate the presence of complex adaptive coordination between an organism's states and its environment, regardless of how that coordination is achieved. But there is also, I suggest, a fairly unified pattern of more serious use of the concept of representation, and this pattern bears the marks of model-

based science. Representational work is done in the style of model-building, and the models are unified by a core structure discussed in this paper, the basic representationalist model.

I will finish this section with a more detailed discussion of one aspect of the model and its use: the model's structural distinction between inner representations and the readers or users of them. I take this to be a central feature of the model, and one that generates many of its natural descriptive and explanatory applications.

The idea of inner "readers" is the basis for a well-known family of objections to representationalism – allegations of regress and pseudo-explanation. According to these objections, the positing of an inner representation requires the positing of an inner reader, but the reader tends to simply reproduce the cognitive abilities that are supposed to be explained. The result is the illusion of explanation. (See, for example, Wittgenstein 1953, and those influenced by him such as Goldfarb 1992.)

A model-based perspective is helpful here. If representationalism is a model, then high-fidelity applications of it must indeed posit a reader or user of some kind (a "consumer" as Millikan has it). And this reader has to have distinctive properties; it is not just the next domino in a chain. But the reader does not have to be as smart as a whole agent. It can be much simpler.

So we do not face an insoluble problem of regress, but instead there is a kind of tightrope that representationalist explanation has to walk. The smarter the reader is, the closer we come to pseudo-explanation. But a reader that is made too unlike a whole agent is not able to act like a reader at all, and cannot justify the description of the structure it is responding to as a representation of something beyond itself.<sup>18</sup> This tightrope can be successfully walked; it is not impossible. Representationalists, with good reason, often cite the case of bee dances (Gallistel 1990, Millikan 1984). This is a clear and striking example of a non-human (though inter-individual) realization of the structure found in the basic representationalist model. If bees can do it, why not neural networks?

In this discussion of reader mechanisms, I have assumed so far that the basic representationalist model is being applied in a fairly high-fidelity way. The model has a separation between representation and reader; that does not mean that all systems to

which the model can be applied must have this separation. The model can be applied using a lower standard of resemblance.

The debate over "distributed representation" in connectionist cognitive science provides an interesting example here. (See Ramsey 1997 for a detailed discussion.) Connectionist models explain cognitive capacities in terms of the activities of layered networks of simple neuron-like elements, whose patterns of interaction are "trained" via learning algorithms. Connectionist systems seem not to support standard distinctions made within mainstream cognitive science, between data structures and the rules that operate on them. These systems are also very holistic in their operation.

People both inside and outside the connectionist movement have wondered whether connectionism, if successful, would vindicate or undermine representationalism. Do connectionist systems represent their environment, albeit in a different way from "classical" cognitive architectures, or do they deal with their environment in a non-representational manner?

The majority of workers within the connectionist camp have wanted to insist that connectionist networks can indeed be described in representationalist terms.<sup>19</sup> The representations are each "distributed" across many or all of the nodes in the network, rather than being more discrete and localized, but this does not disqualify them from representational status. This sort of move has mostly been accepted by philosophical commentators. Certainly no one has been able to show that there is something in the concept of representation that precludes this view. Some writers, however, think that something has gone awry here. Ramsey (1997), in particular, argues that this notion of "distributed representation" does none of the explanatory work associated with the idea of representation within mainstream cognitive science, and does no other new explanatory work either.<sup>20</sup>

I interpret the case as follows. Classical cognitivist models are applications and elaborations of the basic representationalist model, usually with an emphasis on computational manipulation of the inner representations. The structure of connectionist systems seems very different, especially with respect to features of classical architectures that have natural-looking connections to the basic representationalist model. That raises the question of whether connectionism is anti-representationalist. Many connectionists

respond by retaining the model, as a way of talking about the systems they work on, but doing so by moving to a lower-fidelity construal of how the model relates to its targets.<sup>21</sup> This, on the view defended in this paper, is entirely possible. In a standard connectionist system there is at best a very dubious distinction between that which represents and that which consults the representation. There is no proper separation of these two different kinds of causal role. As Ramsey notes, we can say little more of the features often described in representational terms (the "weights" between units, or the activation profiles of the "hidden units") than that they are involved in the transition from input to output, and that modifications of these features systematically affect the way the system handles inputs.

The model can indeed be applied in this very low-key way. What this does, however, is change the status of descriptions and explanations that might be expressed in representationalist terms. Much of the characteristic "explanatory grammar" associated with the basic representationalist model is dependent on the separation between representation and reader.

Take, for example, the explanation of behavioral failure. When a process is described in terms of a representation's guidance of behavior, this brings with it a categorization of explanatory possibilities. The options for explaining behavioral failure or maladaptive action become roughly: (i) inaccurate representation, (ii) bad processing, (iii) bad luck, in a broad sense, and (iv) some conjunction of those possibilities. In this categorization, inaccurate representation is seen as involving the normal, functional consultation of a representation that has some definite, identifiable mis-match with the target. Bad processing, in contrast, involves the mishandling of a representation that could naturally have generated successful behavior. In that case, the blame lies with the processing, rather than the representation.

This partition of explanatory possibilities comes from the representationalist model itself, but the status of these distinctions, in a particular case, depends on the construal of the model that is operating. If the model is being used in a very low-fidelity way, then the target system is not being taken to have distinct components whose roles can be associated with these explanatory options, making possible an allocation of blame

of the kind outlined above. Low-fidelity applications of the representationalist model might, in cases like this, sometimes generate explanatory illusions.

This example involving the connectionist notion of distributed representation illustrates in a vivid way something that I think is very common in less marked forms. The representationalist model brings with it a range of explanatory patterns and possibilities. But because it is handled as a model, there is a great deal of "wriggle room" surrounding the description and explanation of cognition in representationalist terms. In some discussions in the science there can be an almost visible ebb and flow in the way the model is applied, due to the perennial appeal of the model's resources and the ongoing uncertainties over its status.

## **5. How can this be news?**

How can representationalism operate as a model in the sense described here without everyone already knowing it, without it being obvious? How can my analysis be news?

First, I emphasize again that assumptions deriving from the philosophy of science seem to have played a definite role within this part of philosophy of mind. Old views about scientific theorizing really do get in the way. And for some readers, insulated or immune from these influences, what I am saying here may be no news at all.

But I will offer a more tendentious hypothesis that bears on this issue as well. When we describe the mind in representationalist terms, in cognitive science and elsewhere, this may involve the deployment of two deep and non-obvious features of our psychology. When I say "our" here, I include the psychology of the scientists and philosophers themselves.

The two possible psychological features I have in mind are: (i) a general facility for what we might call "model-based understanding," and (ii) a set of psychologically deep habits of interpretation and attribution of meaning. By "model-based understanding" I do not envisage a skill that is exactly the same as what we see in model-based science, but something quite close to it. The tendentious hypothesis is that representationalist work within the sciences is guided or constrained by these two deep habits of thought. A scientist might say: "What I mean by mental representation is Z, so my work will proceed

as follows..." where Z is something quite far from the basic representationalist model as described here. But despite saying this, the scientist might often then lapse back into the characteristic framework of the model. There might be a quite reliable tendency to head back to this way of thinking.

My suggestion here can be compared to one made by Paul Griffiths (2002) about the concept of innateness. Griffiths hypothesizes that innateness is a sort of "attractor concept," one that we head back towards even when we would like to avoid it or deflate it, and a concept with a mass of scientifically unhelpful habits of thought surrounding it. A psychologist might say: "By 'innate' I mean X, and no more..." but still be led back towards more traditional ways of thinking about innateness – essentialism, preformationism, and so on.

Griffiths thinks that in the case of innateness, this tendency is bad for science. Indeed, he thinks we should simply avoid the concept of innateness, because we find it very hard not use it in a particular and rich way. The idea of innateness triggers deep habits of folk-theoretic thought that are possibly adaptive in some situations, but are not helpful in present-day science. Perhaps Griffiths is right in the case of innateness. This need not be so in all cases in which scientific thinking interacts with entrenched pre-scientific habits of thought. Suppose representationalism about the mind is the product of the interaction of deep-seated interpretive habits and model-building habits, operating in a scientific setting. This may or may not succeed in picking up on useful structural facts about how the mind works. Even if it succeeds, however, this situation should certainly motivate a rather cautious attitude towards the idea of mental representation.

My overall view can be summarized as follows. The idea of mental representation is handled with the characteristic flexibility seen in model-based science; it functions as a model. So in the sense of "theory" in which theories can be contrasted with models, representationalism is not a theory at all. I am more sure of that than I am of my account of how exactly this works. The "basic representationalist model" is (I conjecture) the core structure around which more elaborate representationalist models can be constructed. The idea of mental representation can be applied in a highly realist way, and can be used to develop detailed causal explanations. But there are also lots of subtle ways of backing

away from these most realistic construals, and much of the time the model is being handled in this more circumspect manner.

\* \* \* \* \*

**Acknowledgment:** This paper was presented at the Australasian Association of Philosophy Conference, 2004. I thank those present at the talk, and also Matthew Barrett, Tania Lombrozo, Bill Ramsey, and Michael Weisberg, for helpful comments and discussions.

#### Notes

<sup>1</sup> The literature is huge. Some landmarks include Fodor (1975), (1981), and (1987), Dretske (1981), (1988), (1996), Dennett (1978), (1987), Millikan (1984), and Cummins (1996). Useful surveys and synoptic discussions (which also defend specific options) include Cummins (1989) and Sterelny (1990).

<sup>2</sup> In this paper I use the term "cognitive science" in a broad way, so that all of (theoretical) psychology is within this category. Some people use the term in much narrower senses.

<sup>3</sup> I see Sellars, interestingly, as a precursor here, via his treatment of "sense-impressions" in sections 60-61 of "Empiricism and the Philosophy of Mind" (1956/1997). There may well be other more recent precursors. Here I do not have in mind people who simply call representationalism a "model," but those who would do so in conjunction with a contrast between models and theories, of the kind central to this paper.

<sup>4</sup> These other topics are discussed in some other papers. For the status of folk psychology, see Godfrey-Smith (forthcoming a). For the relations between the ideas in this paper and the status of semantic properties, from naturalistic point of view, see Godfrey-Smith (forthcoming b).

<sup>5</sup> More specifically, I am not making use of the views usefully summarized under that title in Suppe (1977). The views employed here are in some ways close to what Downes (1992) calls the "deflationary semantic view," though Downes does not treat resemblance relations in the way I do.

<sup>6</sup> It is also possible, of course, for a physical system to function as a model, for the purpose of representing some other system. Some (though not all) uses of physical

models in science follow the same basic pattern described by Giere for the case of theoretical models. These cases will not be important in this paper, however.

<sup>7</sup> Weisberg's paper contains a detailed discussion of the contrast between model-building and abstract direct representation. In Weisberg's paper, the ecologist Volterra is seen as an exemplar of the former kind of science, the chemist Mendeleev of the latter. Richard Levins (1966) is another important and self-conscious proponent of model-building in the present sense.

<sup>8</sup> Uncertainties of this kind can be found within the literature on the "semantic view of theories." See, for example, Suppe (1977) p. 227, note 565.

<sup>9</sup> Proponents of the semantic view of theories often see a model as a state-space and a set of paths through it (Suppe 1977, Van Fraassen 1980, Lloyd 1988), or as a set-theoretic entity (Suppes 1960). This highly abstract approach is linked to the semantic views' search for a fully general account that works uniformly for all theoretical science, even sciences that do not explicitly work with models at all.

<sup>10</sup> In particular, some may want to place more emphasis than I do here on the concept of computation, and on the role of Chomsky's work. There is also no discussion here of information theory in the sense of Shannon. A slightly fuller account is given in Godfrey-Smith (forthcoming b).

<sup>11</sup> This sense of representation is also assumed by some in the anti-representationalist camp in cognitive science; Van Gelder (1995) is a good example.

<sup>12</sup> As should be clear, the important point does not concern the origins of the concept of representation used in cognitive science, but the ongoing role of this concept. I suppose that most might accept that scientists and philosophers are using a concept of representation that is originally derived from public symbol use.

<sup>13</sup> For some more discussion of the role of resemblance in this part of the story, see Godfrey-Smith (forthcoming b).

<sup>14</sup> Tolman's work is discussed in more detail in Godfrey-Smith (2002) and (forthcoming b).

<sup>15</sup> Distinctions of this kind are sometimes explicitly marked with David Marr's three "levels" of theory in psychology (Marr 1982). For some discussion of how traditional concepts of "level" fit into my analysis of model-based science, see Godfrey-Smith (forthcoming c).

<sup>16</sup> See, for example, Roberts (1998, p. 17), and Eysenck and Keane (1995, p. 2). Sometimes an explicit definition will be given that, fortunately, does no work in the discussion: "A representation is a realization of an entity in a different format."

(Haberlandt 1997, p. 5). Gallistel (1990) is an exception, giving an explicit account of representation that is both coherent and does real work for him.

Best (1992) is an interesting case, as he is very explicit about the "metaphorical" basis of the information-processing approach to the mind. The Eysenck and Keane book also emphasizes the "open-ended" nature of the "computational metaphor" (1995, p. 3).

<sup>17</sup> Ramsey (1997), p. 62, also makes this point, as part of a diagnosis of the handling of representationalist ideas within connectionism. See below.

<sup>18</sup> In Godfrey-Smith (forthcoming c) the present handling of regress problems is contrasted with the more familiar way of handling them seen in homuncular functionalism.

Some ways of talking about mental representation within the philosophical literature seem designed to tiptoe along the tightrope while not explicitly acknowledging the need to do so. There is often no discussion of readers, but the causal role of representational states in relation to downstream processes is carefully marked out as distinctive. For example, behavior might be said to be "guided" by an inner state that covaries with some external condition.

<sup>19</sup> Van Gelder (1995) is an exception, though he is writing within a dynamical systems perspective rather than standard connectionism.

For another interesting discussion of the relation between connectionist models and the idea of symbol-manipulation, see Marcus (2001).

<sup>20</sup> According to Ramsey, many connectionists want to retain representationalist language for reasons that are more rhetorical than substantive; they want to avoid any disciplinary association between connectionism and behaviorism, and they want to reduce the appearance of a complete break with mainstream cognitive science.

<sup>21</sup> I follow the common practice here of writing as if these connectionist systems were usually physically real, as opposed to being themselves modelled within a computer that has a classical architecture.

## References

Barrett, M. (2004). Misrepresenting the Mind. PhD Dissertation, Department of Philosophy, Stanford University.

Best, J. B. (1992). Cognitive Psychology. 3rd edition. St Paul: West Publishing.

Cartwright, N. (1983). How the Laws of Physics Lie. Oxford: Oxford University Press.

Cummins, R. (1989). Meaning and Mental Representation. Cambridge MA: MIT Press.

- Cummins, R. (1996). Representations, Targets, and Attitudes. Cambridge MA: MIT Press.
- Dennett, D. C. (1978). Brainstorms. Cambridge MA: MIT Press.
- Dennett, D. C. (1987). The Intentional Stance. Cambridge MA: MIT Press.
- Downes, S. (1992). "The Importance of Models in Theorizing: a Deflationary Semantic View." In D. Hull, M. Forbes and K. Okruhlik (eds.), PSA 1992, Volume 1. East Lansing: Philosophy of Science Association, pp. 142-153.
- Dretske, F. (1981). Knowledge and the Flow of Information. Cambridge MA: MIT Press.
- Dretske, F. (1988). Explaining Behavior. Cambridge MA: MIT Press.
- Dretske, F. (1996). Naturalizing the Mind. Cambridge MA: MIT Press.
- Eysenck, M. W. and M. T. Keane (1995). Cognitive Psychology: A Student's Handbook. 3rd edition. East Sussex: Psychology Press.
- Fodor, J. A. (1975). The Language of Thought. New York: Crowell.
- Fodor, J. A. (1981). Representations. Cambridge: MIT Press.
- Fodor, J. A. (1987). Psychosemantics. Cambridge: MIT Press.
- Gallistel, R. (1990). The Organization of Learning. Cambridge: MIT Press.
- Gallistel, R. (2001). "Mental Representation, Psychology of" in P. Baltes and N. Smelser (eds.) The International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier.
- Giere, R. (1988). Explaining Science: A Cognitive Approach. Chicago: Chicago University Press.
- Godfrey-Smith, P. (2002). "Environmental Complexity and the Evolution of Cognition." In R. Sternberg and J. Kaufman (eds.) The Evolution of Intelligence. Mahwah NJ: Lawrence Erlbaum, pp. 233-249.
- Godfrey-Smith, P. (forthcoming a). "Folk Psychology as a Model."
- Godfrey-Smith, P. (forthcoming b). "Naturalism, Mental Representation, and Teleosemantics." To appear in Teleosemantics, edited by D. Papineau and G. MacDonald.

- Godfrey-Smith, P. (forthcoming c). "Information and Innateness." To appear in P. Caruthers and S. Lawrence (eds.) The Innate Mind, Volume 3: Foundations and the Future.
- Goldfarb, W. (1992). "Wittgenstein on Understanding." In P. French, T. Uehling and W. Wettstein (eds.), Midwest Studies in Philosophy XVII: The Wittgenstein Legacy. Notre Dame: University of Notre Dame Press.
- Griffiths, P. (2002). "What is Innateness?" The Monist 85: 70-85.
- Haberlandt, K. (1997). Cognitive Psychology. 2nd edition. Boston: Allyn and Bacon.
- Hagueland, J. (1991). "Representational Genera," in W. Ramsey, S. Stich and D. Rumelhart (eds.), Philosophy and Connectionist Theory. Hillsdale: Lawrence Erlbaum, pp. 61-89.
- Levins, R. (1966). "The Strategy of Model-Building in Population Biology," American Scientist 54: 423-431.
- Lloyd, E. A. (1988). The Structure and Confirmation of Evolutionary Theory. Boulder: Greenwood Press.
- Marcus, G. F. (2001). The Algebraic Mind: Integrating Connectionism and Cognitive Science. Cambridge MA: MIT Press.
- Millikan, R. (1984). Language, Thought, and Other Biological Categories. Cambridge MA: MIT Press.
- Morgan, M. S. and M. Morrison, M. (1999). "Models as Mediating Instruments." In M. Morgan and M. Morrison, (eds.), Models as Mediators. Cambridge: Cambridge University Press, pp. 10-37.
- Ramsey, W. (1997). "Do Connectionist Representations Earn their Explanatory Keep?" Mind and Language 12 (1997): 34-66.
- Roberts, W. A. (1998). Principles of Animal Cognition. Boston: McGraw Hill.
- Sellars, W. (1956/1997). Empiricism and the Philosophy of Mind. Cambridge MA: Harvard University Press.
- Sterelny, K. (1990). The Representational Theory of Mind: An Introduction. Oxford: Blackwell.
- Stich, S. and T. Warfield (eds.), (1994). Mental Representation: A Reader. Oxford: Blackwell.

Suppe, F. (1977). "The Search for Philosophical Understanding of Scientific Theories." In F. Suppe (ed.) The Structure of Scientific Theories. 2nd edition. Urbana: University of Illinois Press, pp. 3-232.

Suppes, P. (1960). "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences," Synthese 12: 287-301.

Tenenbaum, J. and T. Griffiths (2001). "Generalization, Similarity, and Bayesian Inference." Behavioral and Brain Sciences 24: 629-640.

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

Van Fraassen, B. (1980). The Scientific Image. Oxford: Clarendon Press.

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

Weisberg, M. (2003). When Less is More. PhD Dissertation, Philosophy Department, Stanford University.

Weisberg, M. (forthcoming). "Who is a Modeler?"

Wimsatt, W. C. (1987). "False Models as a Means to Truer Theories." In M. Nitecki & A. Hoffmann (eds.), Neutral Models in Biology. Oxford: Oxford University Press, pp. 23-55.

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