Notes Toward a Structuralist Theory of Mental Representation

Gerard O'Brien and Jon Opie

Department of Philosophy
University of Adelaide
South Australia 5005

gerard.obrien@adelaide.edu.au
http://arts.adelaide.edu.au/Philosophy/gobrien.htm

jon.opie@adelaide.edu.au
http://arts.adelaide.edu.au/Philosophy/jopie.htm

In H.Clapin, P.Staines & P.Slezak (eds.) *Representation in Mind: New Approaches to Mental Representation* (Elsevier, Oxford; 2004), pp. 1-20.

1. Introduction

Any creature that must move around in its environment to find nutrients and mates, in order to survive and reproduce, faces the problem of sensorimotor control. A solution to this problem requires an on-board control mechanism that can shape the creature's behaviour so as to render it "appropriate" to the conditions that obtain. There are at least three ways in which such a control mechanism can work, and Nature has exploited them all. The first and most basic way is for a creature to bump into the things in its environment, and then, depending on what has been encountered, seek to modify its behaviour accordingly. Such an approach is risky, however, since some things in the environment are distinctly unfriendly. A second and better way, therefore, is for a creature to exploit ambient forms of energy that carry information about the distal structure of the environment. This is an improvement on the first method since it enables the creature to respond to the surroundings without actually bumping into anything. Nonetheless, this second method also has its limitations, one of which is that the information conveyed by such ambient energy is often impoverished, ambiguous and intermittent.

Once the trick of exploiting the information carried by ambient forms of energy has been mastered, however, a third kind of control mechanism becomes available. Instead of responding directly to the information impacting on its sensory surfaces, a creature can use it to construct internal models of the environment—on-board states that "stand in for" or "represent" external objects, relations and states of affairs. These internal representations, rather than the information-laden signals from which they were constructed, can then be pressed into service to shape behaviour. Such a decoupling of behaviour from direct environmental control confers great benefits, since internal models of the environment can have a stability and definiteness that is lacking in the signals that impact on a creature's sensory surfaces. They also enable a creature to respond to features of the world that are not immediately present, to use past experiences to shape present behaviour, to plan for the future, and, in creatures such as ourselves, to be sensitive to very abstract features of the world.

The appearance of creatures with the capacity to internally represent their environment thus constitutes a very significant branch-point in evolutionary history. It amounts to nothing less than the emergence of minds on this planet. Minds are Nature's most sophisticated solution to the problem of sensorimotor control: neural mechanisms that shape behaviour by constructing and processing representations of the environment and the body.

One of the first tasks confronting a science of the mind is to explain how nervous systems can be in the representing business in the first place—how brain states can be about aspects of the world. It is a commonplace in the philosophy of mind that a theory of mental representation must be *naturalistic*, in the sense that it must explain mental representation without appealing to properties that are either non-physical or antecendently representational.[1] What is not quite so commonplace, however, is the further injunction that such a theory must explain mental representation in a fashion consistent with its causal role in shaping appropriate behaviour. Yet this is precisely the message that issues from the simple evolutionary tale we have just told. We shall call this the *causal constraint* on a theory of mental representation.

Remarkably, even though the well-known causal, functional, and teleosemantic theories of mental representation are all naturalistic, they *all* violate the causal constraint. Despite their internal differences, these theories ultimately treat mental content in terms of the appropriate behaviour that cognitive subjects are capable of exhibiting towards their environments.[2] And any theory that treats mental content in terms of intelligent behaviour is in principle unable to explain how mental representation is causally responsible for such behaviour.[3]

The worry expressed in the previous paragraph is not new, of course. It represents just one (admittedly, somewhat unusual) way of formulating the much debated problem of *mental causation*—the problem of explaining how the specifically *representational* properties of brain states can be causally potent over and above their *physical* properties.[4] The standard response to this problem in the philosophy of mind is to accept that representational properties are causally inert, but to argue that there is enough room between explanation and causation for representational properties to be *explanatorily relevant* despite their inertness.[5] In their heart of hearts, however, most philosophers know that this response is deeply unsatisfactory. Our aim, in adopting the present formulation, is to avoid further fancy footwork on the differences between explanation and causation. What is really needed in the philosophy of mind is a completely

[1] See, e.g., Cummins, 1989, pp.127-29, 1996, pp.3-4; Dretske, 1981, p.xi; Field, 1978, p.78; Fodor, 1987, p.97-8; Lloyd, 1989, pp.19-20; Millikan, 1984, p.87; and Von Eckardt, 1993, pp.234-9.

[2] This is a slightly different way of expressing one of the main conclusions Robert Cummins draws in his book Representations, Targets, and Attitudes (1996). Cummins convincingly argues that all of the main contenders in the contemporary philosophy of mind are what he calls "use" theories, in that they ultimately attempt to explain mental content in terms of the use to which representations are put by a cognitive system. Our characterisation follows from Cummins', once one observes that the use of a representation by a cognitive system is ultimately to be unpacked in terms of the role it plays in causing the system to behave appropriately.

[3] Cummins' way of putting this point is to say that use theories cannot account for the explanatory appeals that cognitive science makes to mental representation (1996, especially pp.47-51). The best way to see this, he thinks, is by observing that use theories cannot do justice to *mis*representation. According to Cummins, representational error occurs when a cognitive system uses a representation *incorrectly*. But this requires a robust distinction between

> how a representation is used, and what it means. Since use theories explicitly deny this distinction, they undermine the notion of representational error and with it the explanatory importance of representation (1996, p.47).

Again, however, once one notes that cognitive science invokes misrepresentation (representational error) in order to account for misbehaviour (inappropriate behaviour) it can be seen that, at base, it is their violation of the causal constraint that renders use theories incompatible with the explanatory appeals that cognitive science makes to mental representation.

[4] In the language made familiar by this debate, the worry is that by explaining mental representation in terms of the intelligent behaviour that cognitive creatures are capable of exhibiting, all of the currently fashionable theories entail that the representational properties of brain states fail to supervene on their intrinsic physical properties. This failure entails in turn that the representational properties of brain states do not determine their causal powers.

[5] See, e.g., Baker, 1993; Block, 1989; Dretske, 1987, 1988, 1990; Fodor, 1986, 1989; Heil & Mele, 1991; Jackson and Pettit, 1990a, 1990b; and LePore & Loewer, 1989. Even Cummins was tempted to develop this kind of response in his earlier work (see, e.g., 1989, pp.129-36).

different approach to mental representation—one that doesn't violate the causal constraint, and hence one for which the problem of mental causation doesn't even arise.

The ambitious task we undertake in this paper is to sketch the outlines of a naturalistic theory of mental representation that is consistent with the simple evolutionary story told above. We call this a *structuralist theory of mental representation* for reasons that will become apparent as we proceed.

2. Mental Representation: A Triadic Analysis

What we are all after is a naturalistic theory of mental representation, one that explains mental representation without recourse to the non-physical or antecedently representational. To this requirement we've just added another: such a theory must explain mental representation in a way that is consistent with its causal role in shaping appropriate behaviour. Where are we to find a theory that satisfies these twin demands?

Perhaps we can make some headway by examining representation as it exists in those public objects—words, sentences, paintings, photographs, sculptures, maps, and so forth—with which we are all familiar. By investigating how such public objects operate as representations, we may gain some insight into the nature of mental representation. This is a strategy very effectively deployed by Barbara Von Eckardt in her book *What is Cognitive Science?* (1993). Adapting the work of Charles Sanders Peirce, Von Eckardt analyses non-mental representation as a triadic relation involving a *representing vehicle*, a *represented object* and an *interpretation* (1993, pp.145-9).[6] The representing vehicle is the physical object (e.g., spoken or written word, painting, map, sculpture, etc.) that is about something. The represented object is the object, property, relation or state of affairs the vehicle is about. And the interpretation is the cognitive effect in the subject for whom the vehicle operates as a representation, such that this subject is brought into some appropriate relationship to the vehicle's object. Usually this cognitive effect is understood in terms of the subject *thinking* about the object in question.

What happens when we apply this triadic analysis to the special, and presumably foundational, case of *mental* representation? Given our commitment to naturalism, and hence to a rejection of non-physical properties, the vehicles of mental representation must be understood as brain states of some kind. As for the represented objects, the same analysis we considered above applies: these are the objects, properties, relations, and states of affairs that mental vehicles are about. But the story about the third relatum—namely, interpretation—is different in the case of mental representation. If we apply the account we adopted in the non-mental case, and treat interpretation in terms of a cognitive subject *thinking* about a represented object, we violate the naturalism constraint by interjecting a state that is already representational.[7] What we require, therefore, is another account of interpretation—one that doesn't appeal to further mental representation. This is a complex matter, and one that we can't do justice to here (see Von Eckardt, 1993, pp.281-302 for a much fuller discussion). But, to cut a long story short, the only account that would seem to be available is one that treats interpretation in terms of the modification of a cognitive subject's *behavioural dispositions*. This acts to block the threatened regress since, presumably, it is possible to unpack such behavioural dispositions without invoking further mental representing vehicles. Not any old dispositions will do, however. The process of interpretation, remember, must bring the subject into some appropriate relationship

[6] Von Eckardt actually uses the terms *representation bearer*, *representational object* and *interpretant* to describe the three relata implicated in representation. We prefer our terminology because it is more consistent with the literature on mental representation and with terminology we have employed elsewhere (e.g., see O'Brien & Opie 1999a).

[7] This is why *mental* representation is foundational: any explanation of non-mental representation must ultimately appeal to mental representation to account for interpretation. A theory of representation in general thus waits upon a completed theory of mental representation.

with the represented object. Consequently, interpretation must modify a subject's behavioural dispositions *towards the vehicle's represented object*.

The trouble with this more naturalistic account of interpretation is that it seems to make it impossible to reconcile the triadic analysis of mental representation with the causal constraint. If mental representation incorporates interpretation, and if interpretation concerns modifications to a cognitive subject's behavioural dispositions, then mental representation isn't independent of the subject's behaviour (appropriate or otherwise) and hence can't be causally responsible for it.

All is not lost, however. Von Eckardt observes that the triadicity of representation in general, and mental representation in particular, can be analysed into two dyadic component relations: one between representing vehicle and represented object (which she calls the *content grounding* relation); the other between vehicle and interpretation (1993, pp.149-158).[8] This suggests that any theory of mental representation must be made up of (at least) two parts: one that explains how the content of mental vehicles is grounded, and a second that explains how they are interpreted.[9] The distinction between mental content and mental interpretation is important because it shows us how the triadicity of mental representation can be rendered compatible with the causal constraint. The crucial point is this. A theory of mental representation satisfies the causal constraint as long as it explains how mental *content* is causally responsible for making a cognitive subject's behaviour appropriate to its environment. And a theory of mental representation can do this if it holds that mental representing vehicles possess the capacity to effect mental interpretations *in virtue of* the grounding relations they bear to their representing objects.

Putting all this together, the important question would seem to be: What grounding relations might obtain between mental representing vehicles and their represented objects such that the former are capable of disposing cognitive subjects to behave appropriately towards the latter? Again taking her cue from the case of non-mental representation, Von Eckardt observes that when we consider the many forms of public representation with which we are familiar, there would seem to be three types of ground: *resemblance* (e.g., a portrait represents a person in virtue of resembling them), *causation* (e.g., a knock at the door represents the presence of someone in virtue of a causal relation between them), and *convention* (e.g., the word 'cat' represents the property of being a cat in virtue of a convention of the English language).

According to Von Eckardt, convention is an inappropriate ground for mental representation, since it violates the naturalism constraint (1993, p.206). She never makes explicit her reason for drawing this conclusion, however. If a vehicle is related to its object by convention, the cognitive subject must deploy a *rule* that specifies how the vehicle is to be interpreted. In the case of non-mental representation, where for example the vehicle is a word in a natural language, the application of such a rule is a cognitive achievement that must be explained in terms of processes defined over mental representing vehicles. Perhaps this is why Von Eckardt thinks that convention cannot be a naturalistic ground of *mental* representation. However, as Daniel Dennett has been fond of reminding us over the years, at least some of the rules that govern the behaviour of a cognitive system must be deployable without implicating further representation, on pain of an infinite regress.[10] And computer science has even shown us

[8] Von Eckardt admits that some might object to manoeuvre, on the grounds that if representation is triadic then it can't be properly analysed into two dyadic component relations in this way. She responds by noting that if representation is genuinely triadic then it will turn out that these subrelations are not *purely* dyadic, and hence it will be impossible to explicate either of them without reference to the third relatum (1993, p.149).

[9] According to Von Eckardt, most of the proposals in the literature that purport to be theories of mental *representation* are best understood as theories of mental *content*. Cummins is another theorist who thinks that mental representation decomposes into a number of different elements, each of which requires a distinct theory (1996, pp.20-1)

[10] Dennett attributes this idea to Ryle:

how this can be done: a computational device can tacitly embody a set of primitive instructions in virtue of the way it is constructed.[11] So it is not obvious that convention does fail the test of naturalism.[12]

Even so, it *is* obvious that convention violates the causal constraint. Unlike resemblance and causation, the existence of a conventional ground doesn't entail an *objective* relation between a vehicle and its represented object (see Von Eckardt, 1993, p.149). This connection is forged instead by a rule that specifies how the vehicle is to be interpreted. In other words, convention places the whole burden of representation on the shoulders of interpretation: it is the process of *being interpreted* that confers content on a mental representing vehicle. But recall that in the case of mental representation, interpretation is a matter of modifications to a cognitive subject's behavioural dispositions. Any theory of mental representation that invokes convention, therefore, ultimately treats a mental vehicle's content in terms of the cogniser's behavioural dispositions towards the represented object. Consequently, such a theory is inconsistent with the claim that mental content is causally responsible for such dispositions.[13]

This would appear to leave us with resemblance and causation as potential grounds of mental representation. Of these two, causation has been more popular in the recent philosophy of mind, as it forms the foundation of a number of well-known and much discussed proposals.[14] Yet despite their popularity, causal theories of mental representation, like their conventional counterparts, fail to satisfy the causal constraint. The problem this time is not that the existence of a causal ground fails to entail an objective relation between a representing vehicle and its represented object. The problem is that the connection forged by this objective relation has no influence on the vehicle's intrinsic properties, and hence on its causal powers. Consequently,

> This is what Ryle was getting at when he claimed that explicitly proving things (on blackboards and so forth) depended on the agent's having a lot of knowhow, which could not itself be explained in terms of the explicit representation in the agent of any rules or recipes, because to be able to manipulate those rules and recipes there has to be an inner agent with the knowhow to handle those explicit items – and that would lead to an infinite regress. At the bottom, Ryle saw, there has to be a system that merely has the knowhow….The knowhow has to be built into the system in some fashion that does not require it to be represented (explicitly) in the system. (Dennett, 1982, p.218)

See also Dennett, 1978, pp. 119-126.

[11] The Turing machine can be used to illustrate the point. The causal operation of a Turing machine is entirely determined by the tokens written on the machine's tape together with the configuration of the machine's read/write head. One of the wondrous features of a Turing machine is that computational manipulation rules can be explicitly written down on the machine's tape; this of course is the basis of stored program digital computers and the possibility of a Universal Turing machine (one which can emulate the behaviour of any other Turing machine). But not all of a system's manipulation rules can be explicitly represented in this fashion. At the very least, there must be a set of primitive processes or operations built into the system. These reside in the machine's read/write head—it is so constructed that it behaves *as if* it were following a set of primitive computational instructions.

[12] Indeed, Dennett's own theory of mental representation is based on convention—in this case conventions laid down in our brains by our evolutionary history (1987, especially pp.287-321). This is why Dennett rejects the distinction between *original* and *derived* intentionality: mental phenomena have no greater representational status than the words on this page (1980). Interestingly, Von Eckardt doesn't discuss Dennett's proposal.

[13] Dennett, whose theory of mental representation does invoke convention (see the previous footnote), is quite sanguine about this consequence. For example, he writes:

> There is a strong by tacit undercurrent of conviction…to the effect that only by being rendered explicit…can an item of information play a role. The idea, apparently, is that in order to have an effect, in order to throw its weight around, as it were, an item of information must weigh something, must have a physical embodiment…. I suspect, on the contrary, that this is almost backwards. [Representing vehicles]…are by themselves quite inert as information bearers….They become information-bearers only when given roles in larger systems. (1982, p.217)

[14] See, e.g., Fodor, 1987, 1990; Dretske, 1981; Stampe, 1977, 1986. Teleosemantic theories such as Millikan's (1984, 1989) and the later Dretske's (1987) seem to employ a mixture of causal and conventional grounds.

while there might be causal relations between represented objects and representing vehicles, the latter don't acquire the capacity to bring about interpretations in virtue of these relations.[15] In this respect, vehicles whose content is grounded by causation are in precisely the same position as those grounded by convention, in that the full burden of representation falls on their interpretation. For exactly the same reasons as before, therefore, any theory of mental representation that invokes a causal ground will inevitably treat mental contents in terms of a cognitive subject's behavioural dispositions, and in so doing transgress the causal constraint.[16]

3. Resemblance as the Ground of Mental Representation

Neither convention nor causation can satisfy the two constraints on a theory of mental representation that we have set. It remains, then, to consider resemblance. Resemblance would appear to be an appropriate ground of such public representing vehicles as paintings, maps, sculptures and scale models.[17] What we consider in this section is whether resemblance is an appropriate ground of mental representation.

Resemblance is a fairly unconstrained relationship, because objects or systems of objects can resemble each other in a huge variety of ways, and to various different degrees. However, one might hope to make some progress by starting with simple cases of resemblance, examining their possible significance for mental representation, and then turning to more complex cases. Let us begin, then, with resemblance between concrete objects.[18] The most straightforward kind of resemblance in this case involves the sharing of one or more physical properties. Thus, two objects might be of the same colour, or mass, have the same length, the same density, the same electric charge, or they might be equal along a number of physical dimensions simultaneously. We will call this kind of relationship *physical* or *first-order resemblance*.[19] A representing vehicle and its object resemble each other at first order if they share physical properties, that is, if they are equal in some respects. For example, a colour chip—a small piece of card coated with coloured ink—is useful to interior designers precisely because it has the same colour as paint that might be used to decorate a room. First-order resemblance is a promising grounding relation because it depends on a representing vehicle's intrinsic properties, unlike convention and causation.

[15] For example, the causal relation that obtains between a person and a door knock has no influence on the latter's intrinsic properties, and hence door knocks on their own are quite powerless to effect interpretations.

[16] For a much more detailed consideration of the failure of causal theories of mental representation in this regard, see Cummins, 1996, pp.53-74.

[17] Goodman (1969) developed a famous objection to this approach on the basis that, whereas representation is an *asymmetric* relation, resemblance is *symmetric* (a man resembles his portrait to the same extent that the portrait resembles him, for example). Note, however, that the triadic analysis of representation provides the resources to break this symmetry. In the case of a portrait, what picks it out uniquely as the representing vehicle is the interpretation placed on it by an observer. For a fuller discussion of this response to Goodman's objection see Files 1996.

[18] We discuss *systems* of concrete objects below. Since mental vehicles are presumably brain states of some kind we here restrict our attention to resemblance relations between *concrete* objects. However, we will make some brief remarks about how concrete objects might represent *conceptual* objects such as numbers, theories, and formal systems.

[19] We are here adapting some terminology developed by Shepard and Chipman (1970). They distinguish between first and second-order *isomorphism*. Isomorphism is a very restrictive way of characterising resemblance. We explain both the first-order/second-order distinction and the distinction between isomorphism and weaker forms of resemblance in what follows.

Unfortunately, first-order resemblance, while relevant to certain kinds of public representation, is clearly unsuitable as a general ground of mental representation, since it is incompatible with what we know about the brain. Nothing is more obvious than the fact that our minds are capable of representing features of the world that are not replicable in neural tissue. Moreover, even where the properties actually exemplified by neural tissue are concerned, it is most unlikely that these very often play a role in representing those self-same properties in the world. For this reason, philosophers long ago abandoned first-order resemblance as the basis of a theory of mental representation (see, e.g., Cummins, 1989, p.31).

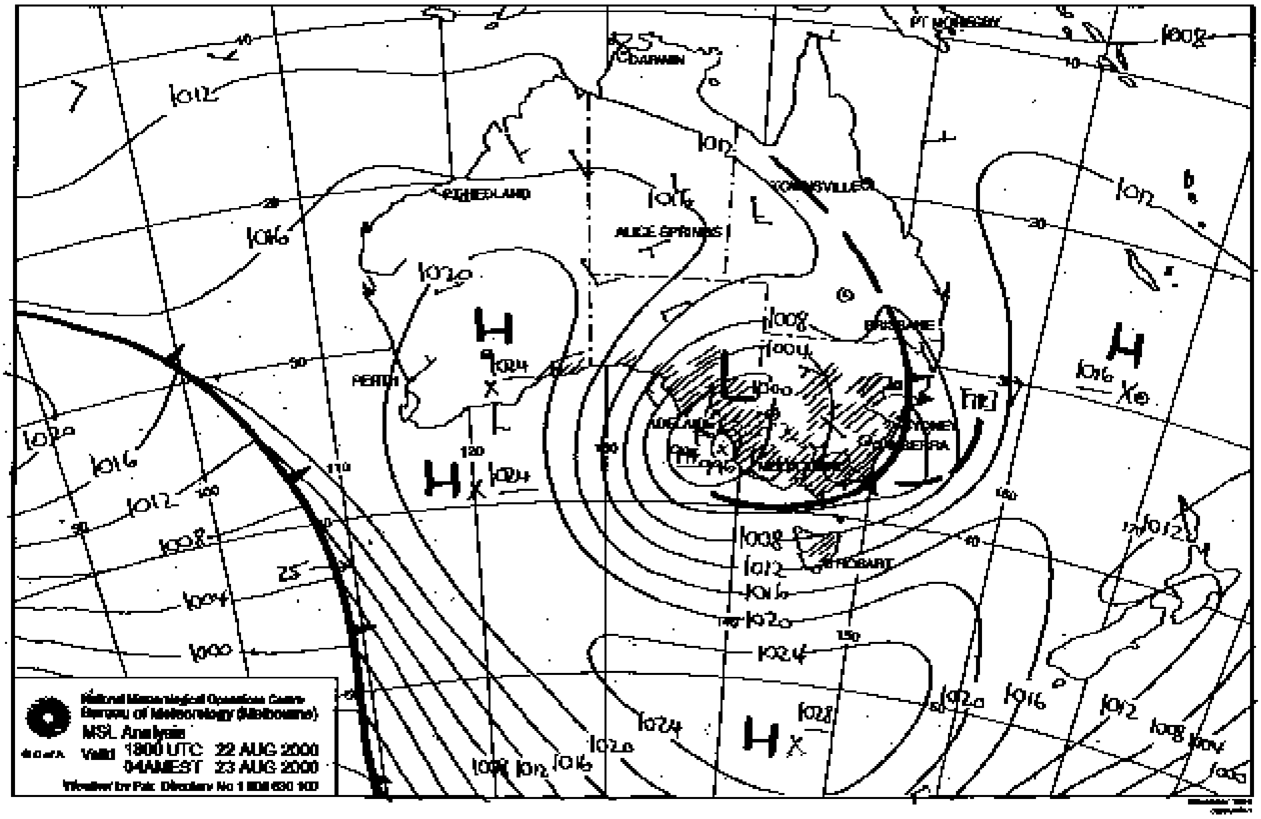

Figure 1. A weather map showing a low pressure cell over South Australia. The isobars around the low are closely spaced, indicating a steep pressure gradient.

But perhaps resemblance is not yet dead in the water. There is a more abstract kind of resemblance available to us. Consider colour chips again. Interior designers typically use *sets* of chips or colour *charts* to assist them in making design decisions. In other words, they employ a *system* of representations which depends on a mapping of paints onto chips according to their shared colour (their first-order resemblance). A useful side effect of having such a system is that when one wants to compare paints (eg., 2-place comparisons such is "this one is bolder than that one", or 3-place comparisons such as "this one harmonises better with this one than with that one") one can do so by comparing the cards. This is because the system of chips embodies the *same pattern of colour-relations* as the paints. Whenever pairs or triples of paints satisfy particular colour relationships, their ink-coated proxies fall under mathematically identical relations.

Similar remarks apply to surface maps. What makes a map useful is the fact that it preserves various kinds of topographic and metrical information. The way this is accomplished is by so arranging the points on the map that when location A is *closer to* location B than location C, then their proxies (points A, B and C on the map) also stand in these metrical relations; and when location A is *between* locations B and C, then points A, B and C stand in the same (3-place) topographic relation; and so on. The utility of a map thus depends on the existence of a resemblance relation that takes points on the map into locations in the world in such a way that the spatial relations among the locations is preserved in the spatial relations among the points.

We will speak here of *second-order resemblance*.[20] In second-order resemblance, the requirement that representing vehicles share physical properties with their represented objects can be relaxed in favour of one in which the *relations* among a system of representing vehicles mirror the *relations* among their objects. Of course, the second-order resemblance between colour charts and paints is a consequence of the first-order resemblance between individual chips and their referents. And in the case of surface maps, space is used to represent space.[21] But one can typically imagine any number of ways of preserving the pattern of relations of a given system *without* employing first-order resemblance. For example, the height of a column of liquid in a mercury thermometer is used to represent the temperature of any object placed in close contact with it. Here, variations in height correspond to variations in temperature.

Weather maps provide a more compelling example. On such a map, regions in the earth's atmosphere (at some specified elevation) are represented by points, and contiguous regions of equal atmospheric pressure are represented by lines known as "isobars"(see Figure 1). More significantly, the *spacing* of isobars corresponds to atmospheric pressure *gradients*, knowledge of which can be used to predict wind velocity, the movement of fronts, and so on. The representation of pressure gradients by isobar spacing is second-order, because for any two points on the map the *relative* spacing and orientation of isobars in the vicinity of those points corresponds to the *relative* size and direction of pressure gradients at the represented regions. Moreover, since *geometric* relations among lines and points on the map are being used to represent *pressure* relations this is a case of "pure" second-order resemblance—it doesn't depend on an underlying first-order resemblance.

Let us be more precise. Suppose $S_V = (V, \Re_V)$ is a system comprising a set V of objects, and a set $\Re_V$ of relations defined on the members of V.[22] The objects in V may be conceptual or concrete; the relations in $\Re_V$ may be spatial, causal, structural, inferential, and so on. For example, V might be a set of features on a map, with various geometric and part-whole relations defined on them. Or V might be set of well formed formulae in first-order logic falling under relations such as identity and consistency. We will say that there is a *second-order resemblance* between two systems $S_V = (V, \Re_V)$ and $S_O = (O, \Re_O)$ if, for at least *some* objects in V and *some* relations in $\Re_V$, there is a one-to-one mapping from V to O and a one-to-one mapping from $\Re_V$ to $\Re_O$ such that when a relation in $\Re_V$ holds of objects in V, the corresponding relation in $\Re_O$ holds of the corresponding objects in O. In other words, the two systems resemble each other with regard to their abstract relational organisation. As already stressed, resemblance of this kind is independent of first-order resemblance, in the sense that two systems can resemble each other at second-order without sharing properties.

Second-order resemblance comes in weaker and stronger forms. As defined it is relatively weak, but if we insist on a mapping that takes *every* element of V onto some element of O, and, in addition, preserves *all* the relations defined on V, then we get a strong form of resemblance

[20] Bunge (1969), in a useful early discussion of resemblance, draws a distinction between *substantial* and *formal* analogy which is close to our distinction between first and second-order resemblance. Two theorists who have kept the torch of second-order resemblance burning over the years are Palmer (1978) and Shepard (Shepard & Chipman 1970; and Shepard & Metzler 1971). More recently, Blachowicz (1997), Cummins (1996), Gardenfors (1996), Johnson-Laird (1983), O'Brien (1999), and Swoyer (1991), have all sought to apply, though in different ways, the concept of second-order resemblance to mental representation.

[21] Note, however, that this is already a step away from the first-order resemblance employed in the case of colour chips, since each colour chip shares a property with its represented object. Maps, on the other hand, don't preserve *absolute* distance, but only distance *relations*.

[22] The relations in $\Re_V$ must have an arity greater than one. We exclude unary relations (ie., properties) in order to maintain a clear distinction between first-order and second-order resemblance (see the definition above).

known as a homomorphism.[23] Stronger still is an isomorphism, which is a one-to-one relation-preserving mapping such that every element of V corresponds to some element of O, and every element of O corresponds to some element of V.[24] When two systems are isomorphic their relational organisation is identical. In the literature on second-order resemblance the focus is often placed on isomorphism (see, e.g., Cummins, 1996, pp.85-111), but where representation is concerned, the kind of correspondence between systems that is likely to be relevant will generally be weaker than isomorphism. In what follows, therefore, we will tend to avoid this restrictive way of characterising resemblance.

The significance of second-order resemblance for mental representation is this. While it is extremely unlikely that first-order resemblance is the general ground of mental representation (given what we know about the brain) the same does not apply to second-order resemblance. Two systems can share a pattern of relations *without* sharing the physical properties upon which those relations depend. Second-order resemblance is actually a very abstract relationship. It is a mathematical or set-theoretic notion—something which should be apparent from the way it was defined. Essentially nothing about the physical form of the relations defined over a system S_V of representing vehicles is implied by the fact that S_V resembles S_O at second-order; second-order resemblance is a formal relationship, not a substantial or physical one.[25]

It is a little acknowledged fact that one of the more prominent approaches to mental representation in the recent literature exploits second-order resemblance. We have in mind the group of theories that go variously by the names *causal*, *conceptual*, or *functional role semantics*.[26] These *functional role* theories (as we shall call them) share a focus on the *causal* relations that a system of mental representing vehicles enter into; where they differ is in the class of causal relations they take to be significant for mental representation. What informs this causal focus is the idea that a system of vehicles represents a domain of objects when the former *functionally resembles* the latter. A functional resemblance obtains when the pattern of *causal* relations among a set of representing vehicles preserves at least some of the relations among a set of represented objects.[27]

Nonetheless, while it is not always made clear (even by their proponents!) that these functional role theories of mental representation rely on second-order resemblance,[28] it is clear that they violate the causal constraint. The reason for this should now be familiar. Unlike

[23] Bunge describes this kind of resemblance as an "all-some…analogy" (1969, p.17). Swoyer refers to it as an "isomorphic embedding" (1991, p.456). A homomorphism is an *injection*: a one-to-one, all-some mapping, because very element of its domain maps to a unique element in its range, but not every element of the range is the image of some domain element. In other words, a homomorphism maps the *whole* domain onto *part of* the range.

[24] An isomorphism is, therefore, a *bijection*: a one-to-one, all-all (surjective) mapping. Every isomorphism is a homomorphism, and every homomorphism is a weak (some-some) second-order resemblance relation. But there are second-order resemblance relations that are not homomorphisms, and homomorphisms that are not isomorphisms. Second-order resemblance is therefore the most inclusive category, isomorphism the most restrictive.

[25] A consequence of this is that a system of *mental* vehicles (which by assumption is a set of brain states) is not only capable of standing in a relationship of second-order resemblance to concrete or natural systems, but also to abstract systems such as logical formalisms and theories. This is presumably a welcome outcome.

[26] See, e.g., Block, 1986, 1987; Cummins, 1989, pp.87-113; Field 1977, 1978; Harman, 1982; Loar, 1982: McGinn, 1982; and Schiffer, 1987.

[27] Some theorists characterise functional role semantics in terms of a *functional isomorphism* between representing vehicles and represented objects, rather than a functional resemblance (see, e.g., Von Eckardt, 1993, pp.209-14). A functional isomorphism obtains when the pattern of causal relations among the set of representing vehicles is *identical* to the pattern of relations among the set of represented objects. We have already observed, however, that a resemblance relationship weaker than isomorphism may generally be sufficient to ground representation.

[28] Cummins (1989, pp.114-25) and Von Eckardt (1993, pp.209-14) are important exceptions.

convention, the connection between representing vehicles and represented objects forged by this grounding relation (ie., by functional resemblance) is fully objective. But just like causation, it has no impact on the intrinsic properties of mental vehicles, and hence no influence on their causal powers. Specifically, mental representing vehicles don't possess the capacity to enter into causal relations (and thereby affect behaviour) in virtue of this second-order resemblance relation; rather, the resemblance relation *itself* obtains in virtue of the causal roles those vehicles occupy in a cognitive economy. This version of second-order resemblance, just like convention and causation, reduces representation to interpretation. Hence the analysis we applied to conventional and causal theories of mental representation, also applies to functional role theories.[29]

With resemblance, therefore, we seem to be caught in a dilemma. On the one hand, any theory that treats first-order resemblance as the ground of mental representation, although compatible with the causal constraint, is incompatible with naturalism. On the other, while second-order resemblance is capable of forming the basis of a naturalistic theory of mental representation, it seems to give rise to theories that violate the causal constraint.

4. A Structuralist Theory of Mental Representation

Fortunately, there is a way out of this dilemma. In all the recent discussion of functional role semantics in the philosophy of mind, another variety of second-order resemblance has been overlooked: second-order resemblance based on the *physical* relations among a set of representing vehicles. We will say that one system *structurally resembles* another when the physical relations among the objects that comprise the first preserve some aspects of the relational organisation of the objects that comprise the second. Structural resemblance is quite different from functional resemblance. What determines the functional/structural distinction is the way relations in the second system are preserved by the first: by *causal* relations in the case of functional resemblance, by *physical* relations in the case of structural resemblance. In neither case is there any restriction on the kinds of relations allowed in the second system—they can be relations among objects, properties or relations; they can be physical, causal or conceptual.[30]

The *structuralist theory of mental representation* is based on the conjecture that structural resemblance is the general ground of mental representation. This amounts to the claim that it is a relation of structural resemblance between mental representing vehicles and their objects that disposes cognitive subjects to behave appropriately towards the latter.

Structural resemblance grounds all the various examples of representation discussed in the last section. A surface map preserves spatial relations in the world via spatial relations among map points. Since spatial relations are a species of *physical* relations this clearly qualifies as an instance of representation grounded in structural resemblance. Likewise, the representing power of a mercury thermometer relies on a correspondence between one physical variable (the height of the column of mercury) and another (the temperature of bodies in contact with the thermometer). And in weather maps the relative spacing of isobars is employed to represent relative pressure gradients. In each of these cases we can identify a system of vehicles whose physical relations ground a (non-mental) representing relationship.

As yet we don't know enough about the brain to identify the structural properties and consequent resemblance relations that might ground mental representation. However, in our view, connectionism has taken important steps in that direction. In this context connectionist

[29] For a much more detailed consideration of the failure of functional role theories of mental representation in this regard, see Cummins, 1996, pp.29-51.

[30] Note that it follows from this that both functional and structural resemblance can be *asymmetric* relations: one system can functionally/structurally resemble a second, without the converse obtaining.

networks are to be understood as *idealised* models of real neural networks, which, although unrealistic in certain respects, capture what may well be the key structural features whereby the brain represents its world (see O'Brien 1998 and Opie 1998). As an example consider Cottrell's face-recognition network (see Churchland 1995, pp.38-55 for discussion). This network has a three layer feed-forward architecture: a 64x64 input array, fully connected to a hidden layer of 80 units, which in turn is fully connected to an output layer comprising 8 units. Each unit in the input layer can take on one of 256 distinct activation values, so it is ideal for encoding discretised grey-scale images of faces and other objects. After squashing through the hidden layer these input patterns trigger three units in the output layer that code for face/non-face status and gender of subject, and five which encode arbitrary 5-bit names for each of 11 different individuals. Cottrell got good performance out of the network after training it on a corpus of 64 images of 11 different faces, plus 13 images of non-face scenes. He found that the network was: i) 100% accurate on the training set with respect to faceness, gender and identity (name); ii) 98% accurate in the identification of *novel* photos of people featured in the training set; and iii) when presented with entirely novel scenes and faces, 100% correct on whether or not it was confronting a human face, and around 80% correct on gender.

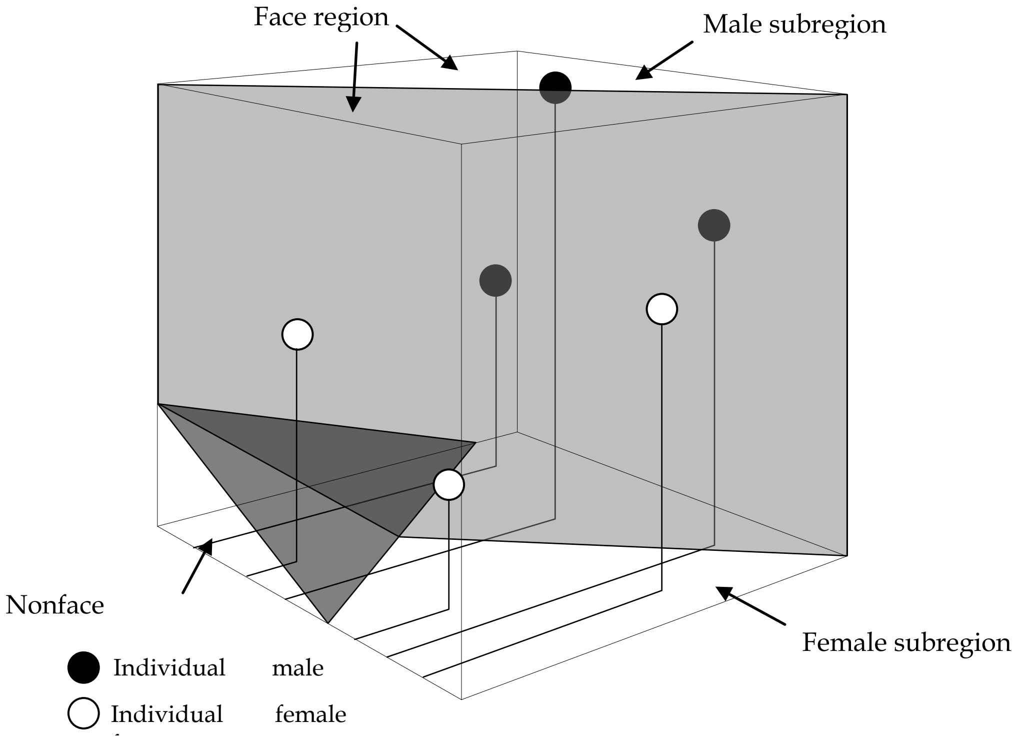

Figure 2. The hierarchy of learned partitions across the hidden unit activation space of Cottrell's face recognition network (after Churchland 1995, p.49).

What is significant about the face-recognition network, for our purposes, is the way it codes faces at the hidden layer. Cluster analysis reveals that the network partitions its hidden unit activation space into face/non-face regions; within the face region into male/female regions; and then into smaller sub-regions corresponding to the cluster of patterns associated with each subject (see Figure 2). What this suggests is that the network supports a structural resemblance between activation patterns and the domain of human faces. Within the face region of activation space each point is an abstract (because compressed) representation of a face. Faces that are similar are represented by points that are close together in the space (this particularly applies to different images of a single face), whereas dissimilar faces are coded by points that are correspondingly further apart. So the relations among faces which give rise to our judgments of similarity, gender, and so on, are preserved in the distance relations in hidden unit activation space. This space, it should be remembered, is a mathematical space used by theorists to *represent* the set of activation patterns a network is capable of producing over its hidden layer. Activation patterns themselves are physical objects (patterns of neural firing if realised in a brain or a brain-like network), thus distance relations in activation space actually codify *physical* relations among activation states. Consequently, the set of activation patterns generated across any implementation of Cottrell's face-recognition network constitutes a system of representing vehicles whose physical relations capture (at least some of) the relations among human faces.

All of this would be moot if this structural resemblance were causally inert. But in fact the structural resemblance embodied in Cottrell's network is arguably what powers both its computational and its behavioural capacities. Structural resemblance thus appears to ground representation in any PDP-style computational system. Accordingly, if the brain is a network of PDP networks, as connectionists urge, then mental content is governed by the structural properties of neural networks, and semantic properties are finally brought home to roost in the mental representing vehicles themselves. Naturally, this is what any advocate of the structuralist theory of mental representation would hope to find. Standard approaches take the intrinsic properties of representing vehicles to be inconsequential, except in so far as these properties are consistent with the causal relations in which these vehicles are caught up. The physical relations among the vehicles are essentially arbitrary so far as their semantic properties are concerned. On the other hand, if the grounding relation for mental representation is a structural resemblance between the system of representing vehicles and the represented objects, then the intrinsic properties of the vehicles run the show. The semantic relations among the vehicles are none other than their physical relations, because it is the latter that generate the required grounding.[31]

According to the structuralist theory of mental representation we can make sense of the power of our inner models to safely guide us through the world if we suppose two things:

1) that the system(s) of mental representing vehicles in our heads stand in an objective relation of *second-order resemblance* to their represented objects;

2) that this resemblance is supported by the *physical* relations among the mental vehicles.

The crucial difference between a theory of mental representation based on structural resemblance and one based on physical (first-order) resemblance, is that the former is compatible with the implementation of mental representation in the brain.

The crucial difference between a theory of mental representation based on structural resemblance and one based on functional resemblance, is that the former is compatible with the causal constraint. The physical relations among a system of representing vehicles are independent of their causal relations, and hence of the use to which they are put. Indeed, it is the physical relations among representing vehicles which explain their causal powers. In this sense, structural resemblance underwrites and explains the existence of any functional resemblance between representing vehicles and their represented domain. This is just how things should be if the evolutionary story with which we started is on the right track.

It's our contention that of all the possible grounds of mental representation, only structural resemblance can satisfy the twin demands of naturalism and the causal constraint. By our lights, this makes the structuralist theory of mental representation mandatory for those philosophers of mind who think this discipline is answerable to cognitive science. In this paper we have merely offered the skeletal outlines of such a theory. The task of putting some flesh on those bones will have to wait for another day (but see O'Brien & Opie (in prep.)).

[31] Paul Churchland is one neurophilosopher who has long argued for the semantic significance of neuronal activation spaces. See, for example, Churchland 1998, forthcoming.

References

Baker, L. R. 1993. Metaphysics and mental causation. In J. Heil & A. Mele, eds., *Mental Causation*. Oxford University Press.

Blachowicz, J. 1997. Analog representation beyond mental imagery. *Journal of Philosophy* **94**: 55-84

Block, N. 1986. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy* **10**:615-78.

Block, N. 1987. Functional role and truth conditions. *Proceedings of the Aristotelian Society* **61**:157-181.

Block, N. 1989. Can the mind change the world? In G. Boolos, ed, *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Bunge, M. 1969. Analogy, simulation, representation. *Revue-Internationale-de-Philosophie* **23**:16-33

Churchland, P.M. 1995. *The Engine of Reason, the Seat of the Soul*. MIT Press.

Churchland, P.M. 1998. Conceptual similarity across neural and sensory diversity: The Fodor Lepore challenge answered. *The Journal of Philosophy* **95**:5-32

Churchland, P.M. Forthcoming. Neurosemantics: On the mapping of minds and the portrayal of worlds. *The Journal of Philosophy*.

Cummins, R. 1989. *Meaning and Mental Representation*. MIT Press.

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

Dennett, D. 1978. *Brainstorms*. MIT Press.

Dennett, D. 1980. The milk of intentionality. *Behavioral and Brain Sciences* **3**: 428-30.

Dennett, D. 1982. Styles of mental representation. *Proceedings of the Aristotelian Society*, *New Series* **83**: 213-26.

Dennett, D. 1987. *The Intentional Stance*. MIT Press

Dretske, F.1981. *Knowledge and the Flow of Information*. MIT Press

Dretske, F. 1987. The explanatory role of content. In R. Grimm & D. Merrill, eds., *Contents of Thought*. University of Arizona Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Dretske, F. 1990. Does meaning matter? In E. Villanueva, ed., *Information, Semantics, and Epistemology*. Blackwell.

Field, H. 1977. Logic, meaning, and conceptual role. *Journal of Philosophy* **74**:379-409.

Field, H. 1978. Mental representation. *Erkenntnis* **13**: 9-61

Files, C. 1996. Goodman's rejection of resemblance. *British Journal of Aesthetics* **36**: 398-412

Fodor, J. A. 1986. Banish DisContent. In J. Butterfield, ed., *Language, Mind, and Logic*. Cambridge University Press.

Fodor, J.A. 1987. *Psychosemantics*. MIT Press.

Fodor, J.A. 1989. Making mind matter more. *Philosophical Topics* **17**:59-79.

Fodor, J. A. 1990. *A Theory of Content and Other Essays*. MIT Press.

Gardenfors, P. (1996) Mental representation, conceptual spaces and metaphors. *Synthese* **106**: 21-47

Harman, G. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic* **28**:242-56

Heil, J. & Mele, A. 1991. Mental causes. *American Philosophical Quarterly* **28**:61-71.

Jackson, F. & Pettit, P. 1990a. Causation and the philosophy of mind. *Philosophy and Phenomenological Research Supplement* **50**:195-214.

Jackson, F. & Pettit, P. 1990b. Program explanation: A general perspective. *Analysis* **50**:107-17.

Johnson-Laird, P. 1983. *Mental Models*. Harvard University Press.

LePore, E. & Loewer, B. 1989. More on making mind matter. *Philosophical Topics* **17**:175-91.

Lloyd, D. 1989. *Simple Minds*. MIT Press.

Loar, B. 1982. Conceptual role and truth conditions. *Notre Dame Journal of Formal Logic* **23**:272-83.

McGinn, C. 1982. The structure of content. In A. Woodfield, ed., *Thought and Context*. Oxford University Press.

Millikan, R.G. 1984. *Language, Thought and Other Biological Categories*. MIT Press.

O'Brien, G. 1998. The role of implementation in connectionist explanation. *Psycoloquy* **9**(06) http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?9.06

O'Brien, G. 1999. Connectionism, analogicity and mental content. *Acta Analytica* **22**: 111-136

O'Brien, G. & Opie, J. 1999a. A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences* **22**: 127-48

O'Brien, G. & Opie, J. 1999b. Putting content into a vehicle theory of consciousness. *Behavioral and Brain Sciences* **22**: 175-96.

O'Brien, G. & Opie, J. 1999c. Finding a place for experience in the physical-relational structure of the brain. *Behavioral and Brain Sciences* **22**: 966-7

O'Brien, G. & Opie, J. In preparation. Structural resemblance and neural computation.

Opie, J. 1998. Connectionist modelling strategies. *Psycoloquy* **9**(30) http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?9.30

Palmer, S. 1978. Fundamental aspects of cognitive representation. In: E.Rosch & B.Lloyd, eds., *Cognition and Categorization*. Lawrence Erlbaum

Schiffer, S. 1987. *Remnants of Meaning*. MIT Press.

Shepard, R. & Chipman, S. (1970) Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology* **1**: 1-17

Shepard, R. & Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science* **171**:701-3

Stampe, D. 1977. Towards a causal theory of linguistic representation. *Midwest Studies in Philosophy* **2**:42-63.

Stampe, D. 1986. Verificationism and a causal account of meaning. *Synthese* **69**:107-37.

Swoyer, C. 1991. Structural representation and surrogative reasoning. *Synthese* **87**: 449-508

Von Eckardt, B. 1993. *What is Cognitive Science?* MIT Press.