THE PROBLEM OF MISREPRESENTATION

MEETS

CONNECTIONIST REPRESENTATIONS

A thesis submitted for the degree
of Master of Philosophy

MASON CASH

1995

CONTENTS

CHAPTER ONE

FALSITY IN MENTAL REPRESENTATION

Theories of semantics try to explain the relationship between a mental[1] representation and the thing it represents; to explain, for instance, how my **coffee** representation represents coffee. (Here and in the rest of this thesis, I use the convention of writing the label for a representation in bold type.) In many traditional theories of semantics, the relationship between my **coffee** representation and coffee is usually explained by recourse to causal relations between coffee and this representation. But attempts at explanations along these lines have many problems, among them the problem that it is difficult to find a plausible way of accounting for the fact that representations are able to misrepresent–or have false content. Sometimes I can think "that's coffee" when what's actually in the cup being handed to me is tea. Getting this fact to sit happily with accounts of the relation between my **coffee** representation and coffee hasn't been an easy task. Traditional approaches to this problem haven't had a lot of success so far in explaining how a representation can misrepresent. In this thesis I aim to avoid the problems with these traditional approaches, and find a causally-based, biologically realistic way to explain semantic relations between mental representations and objects in the world, which is also capable of explaining misrepresentation.

The best place to start such an endeavour is to examine what the problem of representation and misrepresentation is, and the general tactics used in traditional attempts to solve this problem. This will illustrate why misrepresentation appears to be so intractable. Through such an examination we can get a close look at the traditional approaches, and their assumptions about what representations are, what sorts of things they represent, and how they can represent what they represent. We can also get a good view of the unquestioned assumptions these traditional theories are based on. This will give us a good place to start. I'm going to argue that if we want to achieve our

---

1 I am using 'mental' here, and in the rest of this paper in the sense of 'neurological'. I do not mean anything along the lines of 'non-physical'.

aim of a biologically realistic theory of semantics which shows how representations can misrepresent, we'll need an approach to the problem which does not take these assumptions as foundations. In this thesis I aim to construct an account which isn't based on these assumptions.

1.1 *The "Crude Causal Theory": Why misrepresentation is allegedly impossible.*

The first thing to do then, is to set out exactly what the problem is. The relationship between a representation and the objects it represents is usually explained causally. That is, representation represents whatever objects cause its activation. More precisely, a representation represents those objects which *can* cause its activation, or which *reliably* cause its activation, or which causes its activation in a *law-like* manner (these are all equivalent to this basic theory). The following example,[2] will give a good illustration. Say a person, let's call her Diedre, has a representation **kangaroo**, which she has been trained to activate in situations where a kangaroo is present and not to activate in situations where a kangaroo is not present. The result is that Diedre's **kangaroo** representation is activated whenever Diedre comes into contact with (or perceives) a kangaroo. Thus since **kangaroo** is activated by kangaroos, it represents kangaroos. So in general:

- If X situations cause the activation of representation **R**, **R** represents Xs .

Fodor[3] calls this the "Crude Causal Theory". Figure 1.1 illustrates this view: a representation represents whatever object can cause its activation.

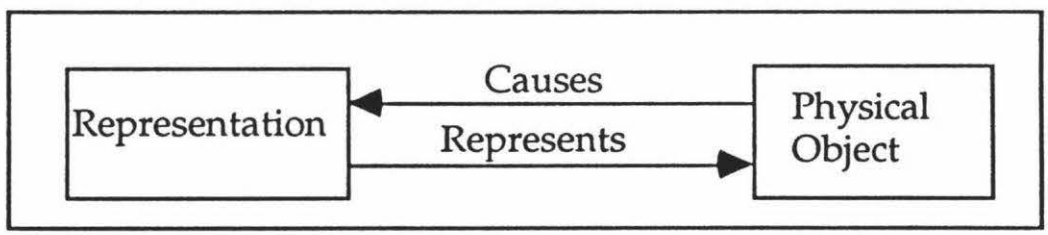


Figure 1.1: Crude Causal Theory's account of representation.

The problem with this Crude Causal Theory, however is that it makes misrepresentation impossible. Imagine that one day Diedre perceives a wallaby, and this also activates Diedre's **kangaroo** representation. In such a situation we

2 This example is stolen and adapted from Kim Sterelney (1990) p122.
3 Fodor (1990)

would like be able to say that the wallaby is misrepresented as a kangaroo, and the representation has the false content *'that's a kangaroo'*. But unfortunately this won't work. The Crude Causal Theory's central tenet is that a representation represents whatever object can cause its activation. So if a wallaby can also cause **kangaroo** to be activated, then **kangaroo** does not represent kangaroos only. It must represent wallabies as well– at least those wallabies which can cause **kangaroo** to be activated. Fodor calls this the "disjunction problem". According to Fodor, **kangaroo** represents (the "disjunctive" class) either a kangaroo or a wallaby, for which I'll use the notation <kangaroo or wallaby>. The problem is that such disjunctive representations cannot have false content. When a wallaby activates Diedre's **kangaroo** representation, her representation doesn't have a false content; it has the true (disjunctive) content *'that's either a kangaroo or it's a wallaby.'*

- If **R** represents the disjunction <X or Y>, then a Y-caused activation of **R** does not have a false content.

The upshot is that there is no way Diedre can mis-represent anything. Anything that can cause the activation of Diedre's **kangaroo** representation, will automatically have to be included in the disjunction of things it represents. Consequently, this representation can never be activated by something other than the things it represents.

However, we do want a semantic theory to allow it to be possible for representations to misrepresent, to have false content. Falsity is an important semantic notion. A semantic theory which doesn't allow representations to have false contents can't be a complete semantic theory.

1.2 *Moving on from the Crude Causal Theory*

The traditional way of getting around this problem is to refine our definition of the class of things a representation represents. We do this by denying that a representation represents *whatever* can cause its activation. Instead we set aside some special circumstances and say that a representation represents whatever causes its activation *in these special circumstances*. Thus the representation is capable of misrepresentation when activated by something other than the things which caused its activation in those special circumstances.

One way of specifying these special circumstances which define the sort of thing a representation represents, is to use the causal relations between objects and the representation *at a certain time*. For example, the period in which a

concept is being formed, or what is sometimes termed the "learning period."[4] The basic idea is that a representation's content, which specifies the things which that representation does and does not represent, is formed during the learning period. The learning period establishes that representation **R** represents a certain type of objects: those which cause its activation during the learning period. It's the teacher's responsibility to make sure that a wide enough sample of objects is used in training so that Xs and only Xs cause the activation of **R**. Because of this training, **R** comes to represent Xs. So in general:

- Since **R** is activated in X situations and only in X situations during the learning period, **R** represents Xs .

This move denies the idea that *anything* which causes the representation's activation is something the representation represents. Some things which cause the representation's activation *after* this learning period could be misrepresented rather than represented.

- After the learning period, if Y were to happen (Y≠X), and Y activates **R**, then the **R** so activated would have the false content that X is the case.

For example, if a wallaby causes **kangaroo** to be activated (after the learning period), then **kangaroo** *misrepresents* the wallaby. **Kangaroo** has the false content *that's a kangaroo* when really what's there is a wallaby.

1.3 The "Counterfactuals" Objection.

Although this story appears to have merit at first glance, such a solution is hopeless (especially according to Fodor). The problem is that because of the nature of causation the learning period can't be insulated against misrepresentation. A causal theory of representational content must be governed by natural causal laws, and a natural causal law must include counterfactuals. However, the learning period story defies counterfactuals, and thus defies natural causal laws. Let me explain. A natural causal law does not merely relate causes and effects by stating that *when* C (the cause) happens then E (the effect) *does* happen. It states more generally that *if* C were to happen then E *would* happen. For instance, the causal law regarding the effects of gravity doesn't merely state that *when* I let go of this otherwise unsupported object it

---

4 This point of view is due to Dretske (1981) . The exposition of it is Fodor's (1987) , and the criticism of it which follows is Fodor's (1990) Crude Causal Theory's response to this idea, rather than any *honest* criticism. This is given as an illustration of CRUDE CAUSAL THEORY and its assumptions and limitations rather than an illustration of the limitations of Dretske's learning period theory.

does fall; it's more general than that. It encompasses the counterfactual, *if* I were to let it go (even if I don't), then it *would* fall. So:

(1) If the statement "Y causes R after the learning period" is true, then
(2) "Y can cause R after the learning period" is true. This means that
(3) "Y can cause R" is true, and thus the counterfactual
(4) "If Y were to happen, then Y would cause R" is true. Thus
(5) "If Y were to happen (during the learning period), then Y would cause R" is also true. And so
(6) "If Y had happened during the learning period (even if it didn't), then Y would have caused R" is also true.

That is, if Diedre's perceiving a wallaby *can* cause **kangaroo** to be activated after the learning period, then if Diedre had perceived a wallaby during the learning period, even though this didn't happen, this also would have caused **kangaroo** to be activated. So if we allow counterfactuals, which we have to do because of the nature of causation, we're forced to conclude that the content established during the learning period isn't plain *kangaroo* after all, but must be *'either a kangaroo or a wallaby'*. Indeed, the content of **kangaroo** isn't even *'either a kangaroo or a wallaby'*, but *'either a kangaroo or a wallaby or anything else which can cause this representation's activation after the learning period.'*

It's in the nature of causal laws that if (1) is true, then all the numbered statements above are true. Basically (5) and (6) stipulate that there is nothing especially sacred about the learning period. Whatever could cause **kangaroo**'s activation after the learning period, would also cause its activation during the learning period. Thus since a wallaby could cause **kangaroo**'s activation after the learning period, a wallaby would cause **kangaroo**'s activation if it were presented during the learning period. And this is enough to include wallabies in the disjunction of things the representation represents. The point is that Diedre hasn't been trained to differentiate a kangaroo from a wallaby, and thus either would activate her **kangaroo** representation. And since **kangaroo** represents whatever *did or would* cause its activation during the learning period, the correlation established during the learning period is not between **kangaroo** and only kangaroos. The correlation is still between **kangaroo** and the disjunction <either a kangaroo or a wallaby>.

If we think about the learning period in this way, the idea appears doomed. Training can't form a representation with a content guaranteed to be correct only when activated in certain situations. If the representation represents whatever activates it or would activate it during the learning period,

then there can't be "wild" activations of a representation, ie. representations which have false contents; not even after the learning period. Diedre's **kangaroo** representation still represents the disjunction <kangaroo or wallaby>, despite her having been trained only on kangaroos. A wallaby can't cause a "wild" activation of **kangaroo**, so it can't cause a representation to have false content either.

This means we still can't get the notion of falsity to be a part of our semantic theory. But falsity remains an important semantic notion we need to account for. So let's look a bit closer at the assumptions made in the above accounts, and see if challenging them can get us anywhere.

1.4 Reject counterfactuals as irrelevant to this account of causation.

The crucial phrase is "*If we include counterfactuals,* the correlation established during learning period is not between **R** and X, but between **R** and the disjunction <X or Y>". Perhaps we could reject counterfactuals. We could perhaps maintain that counterfactuals are irrelevant to the sort of causation we are dealing with here.

Another way to put this worry, is to say that in order for the content of Diedre's **kangaroo** representation to be disjunctive, and to have the content *that's either a kangaroo or it's a wallaby,* surely Diedre has to be *aware* that **kangaroo** represents wallabies as well as kangaroos. And in order for this to be the case it seems that she must have *encountered* wallabies before.

We might well ask: how can Diedre be shown that "were a wallaby presented to you (which it hasn't), you would be tempted to call that a kangaroo too"? Or more to the point, how does one show Diedre this, without showing her a wallaby? And if Diedre has never seen a wallaby, how can her representation represent this potential cause which hasn't happened as well as its actual causes? How could the wallaby bit get into the content of her representation if there's never been a wallaby in her perceptual history to cause this? Surely a causal theory of semantics only needs to have representations founded on the things that actually have caused them.[5]

Look at the example again. If we include counterfactuals, then because a

---

5 Kim Sterelney questions the rejection of Dretske's theory on these grounds also. He asks:

> "...why is Dretske required to count these merely possible contingencies as undermining the claim that, in the learning period, the connection between stimulus and concept is nomic? A correlation does not fail to be reliable just because it is logically possible for it to fail, or even if it is nomically possible for it to fail. If that is necessary for reliability, then no physical device is reliable." Sterelney (1990) : p122.

wallaby would cause **kangaroo** to be activated (even though it hasn't), **kangaroo** also represents wallabies. Thus although Diedre has never seen a wallaby, her **kangaroo** representation has the disjunctive content *'that's either a kangaroo or a wallaby'*. So if a wallaby did cause the activation of **kangaroo**, then **kangaroo** would have a true content, even though there's never been a wallaby in Diedre's perceptual history to cause this. This seems, on the face of it, more than a little weird.

There's an "internal" side to this concern too. Is it not a little odd to say that Diedre has a representation half of whose content she is unaware of? After all, it's her representation. So shouldn't she know what its content is? It's as if someone could say to Diedre "Didn't you know that *this* is a part of the content of your representation too?" Maybe Diedre isn't an authority on what things can cause the activation of her representation, but surely she should be an authority on what her representation's content is.

In contrast, suppose Diedre *had* encountered a wallaby before, and had not been corrected. In this case it would seem to be quite acceptable to say that her representation had the disjunctive content *'either a kangaroo or a wallaby'*, because both kangaroos and wallabies have caused the activation of her **kangaroo** representation.

How much counterfactuals should worry us seems to depend on our interpretation of the sort of causal theory a causal account of representation really requires. The Crude Causal Theory defines the fundamental assumption of causal theories: representations represent the things which can cause their activation. But it doesn't seem necessary to claim that representations represent what *would* cause their activation, merely that they represent what *has* activated then so far. There seems a vast difference between (a)"Representations represent the things which *would* cause their activation" and (b)"Symbols represent the things which *have* caused their activation". On the face of it, (b) seems a much more sensible causal foundation for representation.

However, as I will show in the next section, even this refinement takes us in the wrong direction. It seems feasible to worry about the merits of (b) over (a) only because the picture of a disjunctive representation we have been working with is misleading. We need a better picture of the sort of thing these "disjunctive" representations are and what sort of things they represent. When this is clear, it will also be clear that **kangaroo** can (correctly) represent a wallaby, without Diedre ever having seen a wallaby before.

1.5 *The difference between disjunctions and descriptions.*

The problem with the above account is not so much a problem with counterfactuals, but a problem with our account of a disjunctive representation. The way things have been explained is confusing the issue. There are two factors which compound the confusion.

A lot of the confusion is caused by calling Diedre's representation "**kangaroo**". This name is what gives **kangaroo** its inappropriate (but initially plausible) taxonomic flavour. It gives the impression that **kangaroo** should represent kangaroos and kangaroos only. This is just not so. Calling it "**representation #7934**" would have been a lot less leading. Our job then would be to explain how **representation #7934** has the content it has, whatever that content is, rather than assuming it must obviously represent kangaroos and only kangaroos, and then trying to explain how it can have *that* content.

But the confusion mainly comes from describing the representation's content as "disjunctive". Saying that **kangaroo** represents the *disjunction* <kangaroo or wallaby> is seriously misleading. Sure, if Diedre hasn't been trained to distinguish wallabies from kangaroos, then since wallabies are quite similar to kangaroos, a wallaby could activate **kangaroo**. But there is a better way of explaining this, which does not involve "disjunctions".

Let's have a look at a slightly extreme training situation, to over-emphasise this point, and hopefully clear up the confusion. Suppose Diedre is trained to recognise kangaroos by being shown lots of different kangaroos, in lots of situations, in lots of lighting conditions. Let's say that the only animals around in the learning period are kangaroos and walruses (I said it was going to be an extreme example). Diedre is shown the walruses as a contrast, and taught that these are not kangaroos. Thus **kangaroo** is activated by kangaroos, and not activated by the walruses. Because of this training Diedre can say "kangaroo" whenever kangaroos activate her **kangaroo** representation, and she won't say "kangaroo" when confronted by things (the walruses) which don't activate **kangaroo**.

Diedre's training has only established her **kangaroo** representation specifically enough to distinguish between kangaroos and walruses, not between kangaroos and every other beast she will ever encounter (there are other beasts, we just haven't exposed Diedre to them yet). Diedre's training only included kangaroos and walruses, and there is a specific feature common to all the things that Diedre has been trained to use **kangaroo** to represent: they are beasts which get around by hopping on their back legs. As a result, her impression could be

that **kangaroo** represents things which propel themselves about by hopping on their back legs. In this situation Diedre's **kangaroo** representation would not have the content *'that's a kangaroo'*, so that it distinguishes kangaroos from everything else in her post-learning-period world. There is a very important difference between a representation with the content *'that's a kangaroo'* (the content of a taxonomer's representation of a kangaroo, for instance) and one with the content *'that's a beast which gets around by hopping on its back legs'* (the content of the representation of a person trained only on kangaroos and walruses).

This difference makes all the difference. If Diedre's training only included kangaroos and walruses, and thus her impression is that **kangaroo** refers to things which propel themselves about by hopping on their back legs, then all sorts of things would correctly activate her representation. But even if this is so, saying that **kangaroo** has the *disjunctive* content <kangaroo or wallaby or rabbit or frog or toad or hopping spider or grasshopper> is a very rigid, categorical, and probably incomplete, way of specifying its content. A better way is to say that it has the *descriptive* (albeit vague) content *'a beast which gets about by hopping on its back legs'*.

A representation's content should be seen as descriptive, rather than disjunctive. No representation has a content which chops the world up into the nice, neat scientifically defined categories the Crude Causal Theory would like it to.

The content of such a descriptive representation quite clearly depends on the training that established the representation's content. A person's representation is built up very subjectively. Only through her use of the representation–its behavioural manifestations–can anyone else get a clue as to whether the content of Diedre's **kangaroo** representation is similar to that of other people. If Diedre had encountered wallabies, toads, rabbits and frogs and so on, and called these "kangaroos", then the content of **kangaroo** could have been made much more specific by her being corrected by her teachers.[6]

But even if the content of the representation was made more specific, by such extra training Diedre would never say that her representation's content is *disjunctive*. This is a dubiously theory-laden way of describing the content of a representation. A representation's content is not made more specific by having fewer and fewer disjuncts, it's made more specific by my making the description

---

6 So the learning period idea was onto something. Training is very important in establishing a representation's content., but training doesn't have the function of establishing, for instance, that **kangaroo** will have true content only when activated by kangaroos.

less and less vague. So after being corrected about using the label "kangaroo" to refer to a frog, Diedre might agree that her representation's content was too *vague*, or not detailed enough.

So we can see that the claim I made earlier, in section 1.4, looked sensible but really was quite mistaken. We were concerned, and it seemed right to be concerned, that if Diedre's **kangaroo** representation has the disjunctive content <kangaroo or wallaby> she must have encountered a wallaby before, and *know* her **kangaroo** representation represents wallabies as well as kangaroos. But now that we see the content as *descriptive* rather than disjunctive, we can understand that this is not so. She doesn't need to have encountered a wallaby for the representation with content *'that;s a thing which gets around by hopping'* to be correctly activated by a wallaby. We must realise that the object of training isn't to conclusively establish the content of **kangaroo** so that it distinguishes kangaroos from every other beast Diedre is ever going to encounter. The object of training is to give her representation a content just general enough that she can deal with kangaroos effectively.

Our intuitions are that if Diedre calls something a "kangaroo" when it's a wallaby, then she must *somehow* be misrepresenting the wallaby, because she's put it in the kangaroo category, where it doesn't belong. But in fact our intuitions are wrong, although not for the reasons the Crude Causal Theory uses. **Kangaroo** has the content *'that's a thing which gets around by hopping'*. So if a wallaby activates Diedre's **kangaroo** representation then she does *not* misrepresent the wallaby. She represents the wallaby as a beast that gets around by hopping, which is true of the wallaby. When Diedre meets a wallaby for the first time, it *would* activate **kangaroo**, and she would be quite right in what she *means* by saying "that's a kangaroo". What she means (i.e. the content of her representation) is that this is a beast which gets around by hopping on its back legs, which is true. But there is a difference between what she says and what she means. What she means is true, but what she says is false; that's not a kangaroo, it's a wallaby. But the fact is she has not mis-represented the wallaby; we could perhaps say that she has mis-labelled it. It is just that the representation which she associates with the word "kangaroo" is too vague.

So if we see representations as having (more or less vague) descriptive contents, rather than disjunctive contents, we can see how it's possible for a wallaby to cause **kangaroo** to be activated; in which case the representation has the true content *'that's a beast which gets about by hopping'*. The representation's content is vague enough that it covers both wallabies and kangaroos. And this would be so both during and after the learning period,

whether or not Diedre has ever encountered a wallaby before. So we can't reject counterfactuals. If a wallaby could cause **kangaroo** to be activated after the learning period, then it would cause **kangaroo** to be activated during the learning period. Counterfactuals do matter in the causally-based relations between a representation and what it represents.

Unfortunately then, we still haven't found an account of semantics in which a representation's content can be false; lack of falsity is still a problem here. For, so far, even a descriptive representation can't have a false content. Anything which would activate Diedre's **kangaroo** representation does so because the descriptive content *'that's a beast which gets about by hopping'* is true of it. So even though characterising representations descriptively rather than disjunctively gives us a more plausible perspective on *why* these representations can't mis-represent, we still can't account for false content. All situations in which the representation is activated are situations in which the representation has a true content. We need to dig even deeper to find an account of mental representation in which falsity can play a part. There is one kind of bona fide mis-representation which hasn't been introduced so far. My suspicion is that a lot of what's happening here, the feeling of intractability about the problem, is because this sort of example hasn't been included yet. We need a richer diet of examples to get a better look at what misrepresentation is really all about.

1.6 Use examples which really do exemplify misrepresentation.

The "disjunction problem" states that *anything* which causes or would cause the activation of a representation is to be included in the disjunction of things the representation represents. I translated this as more of a *vagueness* problem. Some representations are vague, so that they apply to more than one similar thing. However, there *are* some cases in which very detailed and specific representations are activated by something which is later discovered not to be at all accurately represented by this representation. This sort of example, which is rare in the traditional literature, *does* provide an example of genuine misrepresentation.

For example: I see someone from a distance walking down the street away from me, and I recognise this person as being Diedre; the walk is right, the clothes look like Diedre's typical apparel, and the hairstyle is right too. Thus my **Diedre** representation is activated, and has the content *that's Diedre.* But as I go running up to greet her, I embarrassingly realise when I see this woman up

close that she is not Diedre.

This sort of situation is where mis-representation truly finds its home. And this sort of situation still needs to be accounted for; the way we've been describing things so far hasn't explained this sort of case. It's certainly not that my **Diedre** representation is disjunctive; it's not a representation whose content is <Diedre or this complete stranger>, or *that's either Diedre or a complete stranger*. And taking my representation's content descriptively, it's not that its content is too vague or badly-formed. The content of my **Diedre** representation is quite specific. It's at least specific enough to distinguish Diedre from the stranger; I know Diedre well, and can recognise that the stranger isn't Diedre when I see the stranger up close and from a better viewing angle. The problem here is not a problem with specifying the content of my representation. The problem is that I'm getting imperfect or incomplete information about my environment. Similar examples of genuine misrepresentation are those of the person who sees a possum up a tree in the dark and thinks it's a cat, the person who sees a cardboard cut-out cow in a paddock and takes it to be a real cow, and the myopic person who sees (without his glasses) his jersey crumpled up on a chair and believes it's the cat.

It's important to notice that activating a representation involves some sort of recognition–a connection is made between the environmental information my senses pick up, and some aspect of my representation. When I thought the stranger was Diedre, the visual information I was picking up matched some of the visual aspects of my **Diedre** representation. But in this case the environmental information that my senses picked up wasn't complete. I was looking at the stranger from a distance, and she had her back to me. If I'd had more complete information to go on–if I'd seen the stranger from close up or had seen her face, for instance–then my **Diedre** representation would not have been activated. So what happened in this case is that the environmental information picked up by my sense organs activated a representation that wouldn't have been activated if the sensory information was of better quality, or had been more complete.

Two points can be made here:

- There are two types of example used in traditional accounts of misrepresentation, examples using representations like Diedre's **kangaroo** representation which are vague, and ones using representations like my **Diedre** representation, which are detailed, specific representations activated inappropriately.
- The senses play the crucial role here–ignoring the role of the senses in

perception is one of the major deficiencies in traditional accounts. And to a large extent it's because they don't acknowledge the role of the senses that they don't see that there are two types of example here.

I'll deal with these points in reverse order. I'll spend some time filling out the importance of the role of the senses in activating representations. After that I'll come back to discuss the examples used to illustrate accounts of misrepresentation,. Because they don't distinguish these two types, a significant proportion of the examples that are used are simply of the wrong sort. They often use vague representations which don't display genuine misrepresentation.

1.7 The role of the senses in perception and representation.

Realising that I don't always (or maybe ever) have access to the complete facts of the way the world is, is one of the major keys to solving the problem of representation and misrepresentation. I don't (and I can't) represent the way the world really is. Rather, I represent the way my senses portray my environment. My representations are activated by the environmental information picked up by my senses. My representations are not activated by objects.

The representations we've been dealing with so far have been incapable of misrepresentation because they have been based upon a perspective in which *physical objects* cause my representation's activation. The Crude Causal Theory assumes that there is no (relevant) intermediary between objects and our representations of objects. The Crude Causal Theory's version of a representation is one which portrays the world as it "really is". Because representations represent the *things which cause their activation*, all the Crude Causal Theory's representations are veridical by definition.

I believe a perspective in which there is an intermediary between my representations and the world makes a lot more sense. The intermediary is my senses. All I really have access to are my sense-organs' outputs, and the way my senses portray the world to me. The properties of my sense-organs' outputs are what cause my representations' activation. But having said this, I don't believe that we *first* perceive this intermediary, and then "infer" the state of our environments from this. The senses *causally* mediate between objects and our representations, but there is no *cognitive* mediation here.[7] (I'll explain why this

---

7 See Ben-Zeev (1988) and Bradshaw (1991) for discussions of the difference between causal and cognitive mediation.

is so in the next chapter.)

The senses' mediation makes all the difference. Misrepresentation occurs when my sense-organs don't *accurately* portray the state of the world. This can happen because the information they pick up is of poor quality due to bad lighting, or because I'm not wearing my glasses. Or it can happen because this information is incomplete, due to bad viewing angle, like seeing the stranger who looked like Diedre *from behind*, for instance. In such situations my sense organs' outputs could activate representations they would not activate if I had access to better quality or more complete information. And it is in precisely such situations, the representation which is activated can have false content.

In order to explain how representations can misrepresent, and have false content, then we need to revise the traditional notion of the way representations are activated. We need an account in which the causes of my representations' activation are not physical objects, but the outputs of my sense organs. On such an account my representations do not represent the objects which caused their activation, because *objects* don't cause their activation at all. The outputs of the sense organs cause the activation of representations. Only with this perspective can representations have false content.[8]

When we put the sense organs in the picture, the diagram becomes:

---

8 On good days I'd *almost* be prepared to give Fodor some credit in not holding the sense organs to be transparent. He does promote a "Slightly Less Crude Causal Theory of Content", in which a sort of foundationalism (inference from sensory information) applies: "The causal chain runs from horses in the world to horsey looks in the world to psychophysical concepts in the belief box to 'hores' in the belief box."Fodor (1987) : p122. Or to put it Granny's way : "...having a HORSE concept requires that you be able to have certain experiences; and that you be prepared to take your having those experiences to be evidence for the presence of horses; and, indeed, that you can sometimes be *right* in taking your having those experiences to be evidence of horses." (also p122) I think Fodor's Granny has a better version.

I say I'd *almost* give Fodor credit for taking the sense organs into consideration because Fodor himself, if we ignore Granny's comments as Fodor appears to do, still ignores the role of the sense organs, going from "horsey looks in the world" *straight* to stuff in the belief box. (Unless "a horsey look in the world" is the outputs of the sense organs?? Fodor isn't clear on this.)

And even in later work (especially when discussing Dretske) Fodor appears still committed to the idea that my **horse** representation is activated by *objects*, rather than experiences or "horsey looks" and that the representation therefore could only represent the object which caused its activation. See for example Fodor (1990) : pp40-42, pp 57-64.

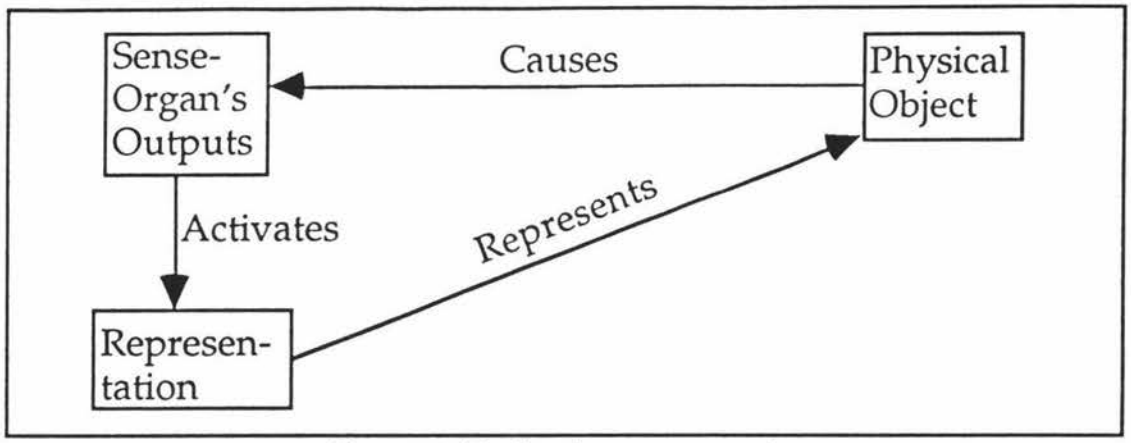


Figure 1.2 Putting the sense organs in.

Thus when I saw the person on the street, and recognised her as Diedre, the stranger herself didn't cause **Diedre**'s activation. Rather, because I didn't see her from close enough, and the viewing angle was not the best, the outputs of my visual sense-organs activated my **Diedre** representation. Because of this I misrepresented this stranger as Diedre. In this situation my representation had the genuinely false content *'that's Diedre'*.

1.8 Which cases properly qualify as examples of misrepresentation?

It seems that many of the main players in the game approach the above question in different ways. Thus often when they think they're scoring points against each other, in reality they're not playing in the same ballpark, they might not even be playing the same game. Ignoring the role of the sense organs for the moment, as these theorists seem to do, the divisions between the positions seem to depend on whether they think a case of misrepresentation can be characterised by:

(i) even though Xs and Ys equally can both cause representation **R** to be activated, **R** *should* only represent Xs, and thus when activated by a Y, **R** misrepresents the Y.
(ii) when representation **R** happens to be activated by something which it shouldn't represent, like a Y for instance, **R** misrepresents the Y.

More accurately, the divisions rest on whether these theorists notice that there is a difference between these two sorts of situations. There is a difference, and it's a very important one.

I believe that version (i), the view held by many theorists, is responsible for misdirecting the debate. But type (i) cases, where a representation can be

activated by two or more different things, but *should* only represent some of these things which can cause its activation, don't exemplify misrepresentation but have vague descriptive contents which apply correctly to Xs and to Ys. And this is so even when we put the sense organs in the picture. Suppose we change (i) to read:

(i') even though Xs and Ys *can* cause sensations which activate representation **R** , **R** *should* only represent Xs, and thus misrepresents when activated by sensations caused by a Y.

Even then we still can't get representation **R** to misrepresent. The problem is that representations to which (i) applies are vague. *If* Xs and Ys *can* both cause sensations which activate **R**, then **R**'s content isn't specific enough to differentiate between Xs and Ys. In such a type (i) situation, we can't say in any non-*ad hoc.* way that **R** *should* only represent Xs. If this "should" is based on anything, it must be based on the representation's content. The problem is that the representation's content is vague, so that it *correctly* represents both Xs and Ys. So we can't use this representation's content to specify that it "should" represent only Xs and not Ys.

There is an important difference between the activation of *vague* representations like those just mentioned and the *inappropriate* activation of representations, as we find with type (ii) situations. The stranger causing sensations which activated my **Diedre** representation is an example of a type (ii) situation. Suppose we put the sense organs back in version (ii) as well.

(ii') when representation **R** happens to be activated by sensations caused by something which **R** shouldn't represent, like a Y for instance, **R** misrepresents the Y.

I want to insist that type (ii) situations, in which the representation is activated because I get poor quality or incomplete sense-information from an object, are the only places where we'll find genuine misrepresentation.

So there is a difference between type (i) and type (ii) situations. Type (i) situations are ones in which a representation's content is vague, and so its content does correctly apply to the thing which caused the sensations which activated it. Type (ii) situations are ones in which a representation's content is specific enough, but the representation is activated by sensations caused by something to which that content does not apply. Many theorists seem not to notice that there is a difference between type (i) and type (ii) situations. And

because they don't notice the difference, these theorists often use type (i) examples to illustrate their account of misrepresentation. But because these cases are open to the "disjunction problem" objection, they often are criticised because the example does not allow for the possibility of misrepresentation. Unfortunately these theorists are short-changing themselves. The failures are often taken, even by themselves, to be failures of their theories of representation, where the fault is rather with the examples they use. (Appendix One is a discussion of some of the more prevalent examples used in the literature, explaining whether these are type (i) or type (ii) cases. But be warned that it requires concepts I don't develop until Chapter Two.) This shows that if we're going to use examples of misrepresentation, we had better ensure that we use the right sort of examples: type (ii) ones which *do* display misrepresentation. In type (ii) cases my sense organs' outputs can activate representations they would not activate if the sensory information was more complete or of better quality. In such situations a representation will have a perfectly specific content, but this content won't apply to the object which caused the activating sensory outputs. That is, the representation will be incorrectly activated, and will misrepresent the object which caused the activating sensory outputs.

This view of how mental representations can have false content fits perfectly to many familiar situations. As we've seen, it fits my **Diedre** representation being activated by the stranger seen from behind. But take a slightly different example: I'm not wearing my glasses, and see my grey jersey on the chair, and take it to be my grey tabby cat, Madison. Here I misrepresent the jersey as Madison. This time, rather than getting incomplete sense information, the sensory information picked up by my visual perceptual system is noisy, or of bad quality. But it's ridiculous to say that my **Madison** representation is disjunctive, and really represents the disjunction < Madison the cat or my grey jersey (when I'm not wearing my glasses, and I see it from far away)>. And it's equally ridiculous to say that my **Madison** representation is descriptive but vague, so vague that it covers both Madison and grey jerseys too. Indeed, what would a description which equally describes grey cats and grey jerseys seen without my glasses on even look like–a greyish something or other? What has happened in this case is that my senses translated information about my environment imperfectly because I wasn't wearing my glasses. Because of this, the visual information was noisy enough that some aspect of it fitted some aspect of my **Madison** representation. Because of this noisy sensory information my **Madison** representation was activated inappropriately; it would not have been activated if I had been wearing my glasses. Here we have a case where a

representation with a very specific content is activated by sensations caused by something that content doesn't correctly apply to.

We could use two "tests" to check if any example is a type (i) or a type (ii) case. It would be a type (ii) situation, which does exemplify misrepresentation, if either of the following were the case:

- If the environmental information was of better quality or less noisy, the same representation wouldn't be activated.
- If I attempt to get more complete information, to activate the representation through other of its aspects (by looking from a different angle, by listening, smelling, feeling and/or tasting as well as looking, or by looking closely at features not inspected originally) the same representation wouldn't be activated.

The first test implies that if I improve the quality of the sensory information, by turning the lights on, by putting my glasses on, by moving to a distance where the object's features are more distinct, I could check whether this same representation would be activated. If this is a type (ii) situation then the aspect of the sensory information I was receiving would become better quality and I would realise my error; I'd most likely activate a different representation instead. This happened when I went up to pat Madison the cat when I my grey jersey activated this representation. As I moved closer and the sensory information got a little more distinct, I realised that this wasn't Madison the cat. My **Madison** representation was no longer activated.

The second test probably played more of a part in my realising that the stranger wasn't Diedre. (In fact it probably also played some part in my realising that my jersey wasn't Madison too.) Here, rather than improving the quality of the sensory information which activated certain visual aspects of my **Diedre** representation, this way of testing attempts to activate other aspects of the same representation. For instance, if I walked around and saw her face, or if I heard her speaking and realised that her voice isn't anything like Diedre's, then this other sense-information would in some way inhibit **Diedre**'s activation, because these are not aspects of that representation.

To sum up: if we want a semantic theory to permit mental representations to have false contents, we need to allow for the fact that my senses mediate between my environment and my representations, and we also need to use the right examples to illustrate the explanation. We need to use cases where a representation misrepresents in the sense of representing something its content doesn't correctly apply to, because it was activated

inappropriately. And we need to acknowledge that this representation was activated inappropriately not by the wrong object, but because the outputs of the sense-organs carried incomplete or poor quality information. Being insensitive to these points is a major stumbling-block for many traditional approaches to the problem of misrepresentation.

1.9 Traditional approaches to the problem: General Tactics.

Nonetheless, there is a general tactic used in traditional approaches to the problem which merits examination and praise. The general tactic is this: It's clear that the Crude Causal Theory's thesis that a representation represents whatever can cause its activation won't do. Hence most theorists try to find a principled reason for saying that the representation represents the physical objects which cause its activation in certain "optimal" cases only, so that in other "non-optimal" cases it can misrepresent the object which caused its activation. (As I've just explained, many of these theories ignore the role of the sense organs.) This way of tackling the problem can be summarised as follows:

A) In certain "optimal" situations, representation R represents the thing(s) which activated it. I'll call these things "Xs".
B) In some situations we want to say that R *misrepresents* the thing which activated it. We can't justify calling this a case of misrepresentation *just* because this is a "non-optimal" situation, where the thing which activated R is a Y and not an X, because this would be *ad hoc.* and circular.
C) Because of their realisation of point B, theorists such as Millikan, Dretske and Fodor (the prime examples, whose theories I'll concentrate on) each try to give a theory of representation which explains–in a principled, non-*ad hoc.*, non-circular way—what the representation does (and does not) represent. They try to explain in such a way how R can correctly represent only Xs, and thus how it can misrepresent when activated by Ys.[9]

---

9 Note that I haven't used the word "content" here. Most theories claim that the job to do is to specify the representation's content in a principled, non circular way. But the way "content" is used in the literature is dangerously ambiguous between what's inside the representation, and what's at the end of the represents relation. A confusion between these two is endemic. I've made this mistake myself often., and have had to catch myself over and over again. An example is the discussion about the difference between disjunctive contents and descriptive contents mentioned earlier in this chapter.

Because of this ambiguity I'm going to stop talking about content, and instead talk about the sort of object a representation should represent, or the sort of object the representation correctly represents. I take this to refer unambiguously to the entity at the business end of the represents relation.

When I do get around to describing how we can *specify* the sort of object a representation should represent, I use a device which is neither part of the representation itself, and nor is at the other end of the representation relation. Rather it sits outside the representation, but is used in establishing *the way* the

The general tactic then, is to explain what a representation represents first. How each of the main proponents tackle the problem of misrepresentation differs in how they establish how representation R is first able to represent Xs and thus to misrepresent Ys.

Actually there are two versions of C). To establish how representation R is first able to represent Xs there are two approaches. Only the second uses the tactic I want to applaud. Bluntly, the difference between these tactics is this:

C1) Assume that representation R represents Xs. Use this to show how it can misrepresent Ys.

C2) Show how representation R comes to represent the things it represents (which just happen to be Xs). Because of this it can misrepresent Ys.

The C1 version is Fodor's. He tries to show how a representation can correctly represent one thing and misrepresent another using what he calls "asymmetrical dependence". He says that my **Diedre** representation's ability to misrepresent the stranger must be dependant on its ability to correctly represent Diedre; I couldn't misrepresent the stranger as Diedre unless I was able to correctly represent Diedre as Diedre. But this dependence is asymmetrical, it doesn't run the other way: my **Diedre** representation's ability to veridically represent Diedre doesn't depend on its ability to misrepresent the stranger. So R can misrepresent Ys because Y-caused activations of R are asymmetrically dependant on X-caused activations of R.

The aspect of this tactic I'm wary of is that it starts with the finished representation. Accounts like this invoke the spectre of circularity. You must be able to explain in a non-circular way how R can represent only Xs, and thus can misrepresent Ys. This is not an easy task.

Fodor's version of this story avoids the circularity by boot-strapping instead: he makes no attempt to explain how a representation can *come to* represent what it does. He avoids the responsibility for explaining how R comes to represent what it does, by hoping that he can help himself to the concept of an intact organism.[10] To illustrate: he says that "...misidentifying a cow as a horse wouldn't have led me to say 'horse' *except that there was independently a semantic relation between 'horse' tokenings and horses.*"[11] Fodor shrugs off the

---

represents relation points to what it does.

10 Fodor (1987) : pp106-110 and pp126-127.

11 Fodor (1987) : p107. (His italics, my bolding.)

12 Millikan (1984) and Dretske (1981).

responsibility of showing how this independently existing semantic relation is established. He starts out explaining how **horse** can misrepresent cows, by reference to its ability to correctly represent horses, without ever explaining how **horse** can come to correctly represent horses.

Fodor starts out *assuming* that **horse** obviously represents horses, and tries to explain how it can misrepresent cows. Millikan and Dretske don't start off assuming that **horse** represents horses. Their (C2) accounts avoid the charge of circularity by starting at the other end. Rather than beginning with a fully developed representation **R** and trying to explain how it can represent only Xs and not Ys, we begin with the question, "How does **R** *develop from scratch*, so that it comes to represent the objects it does (whatever those objects are)." (How the representation develops explains how Diedre's **kangaroo** representation can be activated by both kangaroos and wallabies.)

An explanation which starts with the raw material which develops to become the representation, doesn't incur any charges of being circular. The process is iterative rather than circular. Notice that by taking this tactic, these approaches explain how a representation comes to represent the object it represents, rather than asuming that, say, my **kangaroo** representation must have represent *kangaroos*. (Think about **representation #7934** again.)

One important lesson to learn here, is not to start an explanation of the sort of thing a representation represents with a look at the *finished representation*, and attempt a non-circular explanation of how that representation can represent what it does. Instead we describe how a representation *develops from scratch*, and in particular how it comes to represent the sort of thing it represents. By doing so we avoid any charges of being circular. Depending on how the representation has developed, the sort of thing the representation represents could be a vaguely specified class of things, or it could be quite specific, or it could be somewhere in between.

Theories which take this tactic differ in the ways they think a representation develops, and thus how the sort of thing a representation represents should be specified: Millikan argues that representations have Natural functions which develop through evolution, and Dretske (in 1981) argued that learning during the "learning period" is what specifies the sort of thing a representation represents.[12] Having accounted for what a representation correctly represents, these approaches then explain how because the representation correctly represents a certain sort of thing, it can misrepresent

12 Millikan (1984) and Dretske (1981).

when its activation is caused by something other than this sort of thing.

Dretske and Millikan take the following general tactic then: they firstly explain how a representation develops from scratch, in order to non-circularly explain how a representation comes to represent a certain sort of thing. With this established, they can determine when the representation veridically represents and when it misrepresents: it misrepresents when activated by objects other than this sort of object. (Or rather they *should* say: when activated by sensations caused by objects other than this sort of object.) Thus my **Madison** representation developed to represent a very specific class of things. It correctly represents a grey tabby cat with a big appetite, who lives at my house, who loves chocolate and cheese and who sheds hairs all over my favourite chair. So because grey jerseys are not the sort of thing this representation correctly represents, when **Madsion** is activated by my seeing, without my glasses on, my grey jersey where I left it on my favourite chair, it misrepresents the jersey

This *general* tactic is one I'll use too. I will however need to revise, append and replace some of the assumptions made in traditional accounts of representation (some of which, alas, Dretske and Millikan also take on board). In the next section I'll mention these assumptions, and briefly sketch the ways I'll revise them.

1.10 Some troublesome assumptions of traditional approaches to the problem of misrepresentation.

The following are a few assumptions which, it seems, most of the traditional attempts to solve this problem take on board. This thesis could be seen as an attempt to set out a new way of approaching the problem–a way which revises or rejects these assumptions:

(a) Physical objects activate representations.
(b) The sense organs' job is to convert the properties of objects into properties of representations of those objects.
(c) In explaining how we represent our environments, how those representations are used is relatively unimportant.
(d) Physical objects (as opposed to abstract ones) are the only kind of objects which can figure in an account of representation.

I'll deal with these assumptions in sequence.

I've already discussed the first of these. Assumption (a) refuses to take the sense organs into account. As I said earlier, because of the mediation of the

senses, a representation doesn't (correctly) represent every*thing* which causes its activation, because *things* don't activate representations anyway. It's only the outputs of the sense organs which do this.

Assumption (b), that the sense organ's job is to convert the properties of objects into properties of representations also needs to be revised. In the next chapter I'll argue that seeing the senses as transducers of *information* provides a refreshing perspective, which makes a lot more sense than one based on assumption (b). The idea here is that there is a lot of information already contained in the light waves, sound waves and so on that impinge upon our sense organs. The senses' job is not to convert properties of objects into properties of representations of those objects, but to convert information implemented as light waves, to information implemented as neurological impulses; the same information is transduced into a form more accessible to our brain processes.

We also need to reconsider the traditional perspective with regard to assumption (c). I'm going to show that the way a representation comes to represent what it does is intimately tied up with the way that representation is used in the production of behaviour. Our perceptions activate representations, and the activation of representations is used to produce actions appropriate to the situations and circumstances we represent ourselves as being in. A representation's job is not just to represent, but to coordinate action with perception. Traditional accounts need to take more notice of the relationships between perception and action which are embodied in our representations. Action and perception co-evolve, and by developing together the cognitive structures which undergird our representations are formed. Because of this co-evolutionary development of action and perception, what a representation represents is given by the way it is used to coordinate action with perception.

Assumption (d), that physical objects are the only kosher objects, and that abstract objects shouldn't figure in accounts of representation and misrepresentation can be rejected by looking at the developments made in the philosophy of language during the early part of this century. The work of Brentano, Twardowski, Meinong, and Frege showed that abstract objects *must* figure in our explanations of what a word means. As I'll show later on, the same goes for explanations of what a representation represents.

In the next chapter, I'll start building a position which rejects the above assumptions. I'll attempt the task of providing an account of how a representation is activated, what a representation is, what a representation

represents, and how a representation represents whatever it represents which accords with the revisions of the above assumptions. This account will use the general tactic I mentioned earlier; the tactic of specifying how a representation develops so that it comes to represent what it does, and then using this to specify when a representation correctly represents and when it misrepresents. To do this, I'll begin in the next chapter by taking a close look at the "nuts and bolts" of how the outputs of the sense organs activate representations, the way the sense organs act as transducers of environmental information, and the way a representation could be implemented in the human brain.

CHAPTER TWO

CONNECTIONISM AND THE FLOW OF INFORMATION

2.1 *The importance of implementation details*

Many of us have had the experience of walking into a room, and picking up a smell which we immediately recognise and which brings up memories associated with that smell. Recognising the smell somehow activates a certain representation, and this event brings to mind lots of other information you associate with the object or state of affairs you associate with this smell. The ability to *somehow* activate a representation on the basis of one piece of information which is associated with that representation is an important facet of human cognition. In this chapter I'll go through the details of how the sense organs are thought to work, and outline the methods by which their outputs could activate such representations. This will get us ready for the next chapter in which I discuss how representations like these might be implemented neurologically.

Now some claim –or at least imply– that how a representation is implemented in the brain is largely irrelevant; as long as we know what representations do (they represent) we can leave the details of how they're implemented up to the neurobiologists to figure out. They maintain that we should carry on with the game of figuring out how a representation represents and misrepresents. The nature of this mysterious representation relation which can hold between some piece of my brain (whichever piece that happens to be) and certain objects, like cups of coffee, kangaroos, and grandmothers "out there" in the world is difficult enough. What the piece of my brain which represents the coffee I'm about to drink would be like, and how it could function as a piece of my brain are neurobiological side-issues irrelevant to our investigations of the *real* problem.

I couldn't disagree more. By examining the details of exactly how a representation can be implemented neurologically we get a much clearer picture of exactly how a representation can represent. That is, if we have a reasonable picture of what such a piece of my brain would be like, and how it functions as a piece of a brain, we'd have an invaluable tool with which to

investigate the "real" question I just mentioned, that of how a representation represents and misrepresents, and what the nature of this representation relation is. The purpose of the next three chapters is to provide such a tool.

My aim is to explain how picking up on some piece of information can cause the activation of the representation this information is a part of. I'll do this by first illustrating how certain brain mechanisms operate, by way of a discussion of connectionist networks and how they work. An understanding of connectionist networks will help us understand how, for example, smelling coffee can activate my **coffee** representation. With a basic understanding of connectionist principles, and how these principles can explain how our perceptual systems can pick up information about our environments, I'll turn in Chapter Three to discussing connectionist representations themselves. Then in Chapter Four I'll what such representations would need to be like to explain how we use them. In Chapter Five I'll introduce the idea of where semantic relations like representation come from. Chapter Six is then an in-depth look at the representation relation itself; what it is, how it works and how it can hold between some part of the world and a representation like the ones I will have set in place. I'll then conclude, in Chapter Seven, by measuring the system I'm outlining against a set of criteria which have been suggested, rightly, as a yard-stick for any respectable theory of mental representations.

Before I get into discussing connectionism, and this new way of looking at representations and how they're implemented and activated, we need to get some terminology straightened out. In particular, we need to be very clear about what we are using the term "representation" to refer to. Many traditional theories hold that the sense organs output are themselves *representations* of environmental stimulus. For instance, everything "inside" my retina is a representation of what I see, the outputs of my olfactory sensors are representations of smells, the outputs of my tactile sensors are seen as representations of surfaces and textures, and so on. Against this, however, I don't want to claim that the outputs of my sense organs "represent" anything, much less that they "represent" the environmental information they transduce. This is mainly in the interests of avoiding confusion rather than any deep ontological dispute. That is, I want to avoid using the word "representation" to refer to the outputs of the sense-organs, because I want to be able to reserve "representation" to use to refer to a neurological entity which is used to stand for some *object* or *state of affairs*. I also want to avoid using "representation" to refer to the outputs of the sense-organs because "represents" is too suggestive of some sort of high-level cognitive process. The senses are merely low-level transducers; they

pick up on some aspect of a perceiver's environment and transduce this so that the information picked up is neurologically implemented (and thereby accessible to further brain processes.

I want to introduce another word for what is happening at the level of the sense organs. I'll use the word "encode" rather than "represent", to label how the sense organs' outputs relate to whatever it is they pick up on. I prefer "encode" as it's more in line with the sort of low-level transductive process we should expect sense organs to instantiate. Thus the outputs of my olfactory sense organs *encode* smells. From now on then, I'll reserve "represents" for the special case where a neurological item stands for an object or a state of affairs, rather than using it to refer to the outputs of the sense organs.

2.2 *What might a representation look like?*

When I first started thinking about representations and their parts, I thought of a representation as being like a long list of neurologically encoded sensations associated together. So my **coffee** representation for example, would be a list of all the images, sounds, smells, tastes, and tactile sensations[1] I associate with coffee. This "list" would include:

- The outputs of each of the different types of taste-buds I have, when I taste coffee with sugar, black coffee, instant coffee, strong espresso, and coffee with milk.
- The outputs of each of my different types of smell receptors, when I smell freshly brewed coffee, coffee beans, instant coffee powder, coffee-cake, coffee essence and so on.
- The tactile and heat-sensor sensations related to drinking cups of coffee– such as the feeling of holding a warm mug in my hands, and the burning feeling on the roof of my mouth when the coffee's too hot.
- The visual sensations I associate with coffee– images caused by coffee mugs, percolators and espresso machines, people sitting drinking in cafes, and people putting water in the kettle. Also those caused by written words and symbols which I associate with coffee–such as "Coffee", "Latté", "Café", "Decaffeinated", "Nescafé", and the "Robert Harris" logo.
- The auditory sensations I associate with coffee–such as those caused by kettles whistling, percolators gurgling, the tinkling sound made by spoons stirring in mugs, and slurping-drinking noises. Also those caused by the spoken word "Coffee"–spoken in all sorts of accents and with all sorts of inflections and intonations.

---

1 I'm using "sensations" here, and in the rest of this thesis, to refer to the outputs of the sense organs.

- My **coffee** representation would also be associated with certain other representations. These would include representations of facts like where to find it at Pak'n'Save, the fact that some people prefer tea to coffee, and the fact that there's never any left in the common room after 10:30am. It would also include representations like my **favourite-coffee-mug** representation, my **how-to-make-coffee** representation, and my **favourite-coffee-shop** representation.[2]

By having a list of neurologically encoded sensations like this *somehow* associated together, experiencing one of these sensations could activate the rest of the encoded sensations in the list. That is, one part of the representation (the smell sensation caused by coffee, for example) being encoded by the present outputs of my smell sensors, would *somehow* cause all the other encoded sensations which are parts of this representation to be labelled as important or relevant; to be *somehow* brought to my attention. Thus smelling coffee would start me thinking about coffee, recalling what it tastes like, planning where to go to drink some, and so on; it would activate my **coffee** representation.

Although we are now fairly certain that human memory does work like this, researchers call it *content addressable memory*, filling out the details of the above "somehow"s proved exceedingly difficult. The following is one way this could happen.

2.3 *The activation of a representation–Take One*

So how can my smelling coffee start me thinking about coffee? How can one particular encoded sensation, say an encoded smell, *somehow* cause the activation of the whole representation this encoded smell is a part of?

When I started thinking about this question, I first came up with the following tasks which must be done. For a start, I must have some sort of *recogniser* of encoded sensations. That is, something which can tell, for example, if the smell sensation I'm currently experiencing is a sensation which I've experienced before. Furthermore if this smell has been experienced before, this recogniser needs to be able to recognise which representation this encoded sensation is a part of. And of course, it must then cause the activation of that representation. So it seems that there are three jobs to be done between my smelling a particular smell and my **coffee** representation being activated:

---

2 My intuition is that all of these would themselves be composed of associations of encoded sensations and representations, which representations are also composed of associations of sensations and concepts, until everything grounds out in encoded sensations eventually.

(1) Check whether the currently experienced sensation is one which has been experienced before.
(2) Recognise that this sensation encodes the smell of coffee.
(3) Activate my coffee representation.

It seems as though there is a sequence of operations to be performed here. First (1), then (2), which then causes (3).

However, when we look closely at the job of activating a representation and at the neurological mechanisms responsible, it will turn out that this is not a *sequence* of *separate* jobs. Indeed, even without looking at neurological mechanisms, we can see that if I can do (2), then (1) doesn't need to be done as a separate task. If (2) is done, then (1) is done simultaneously. If I can recognise that this smell is the smell associated with my **coffee** representation, then I must have also recognised that I've experienced this smell before. Initially I had thought that we needed a *separate* (1) to explain how I can recognise a sensation, but not recall where I've experienced it before. For me, this happens a lot with peoples' faces: that familiar "I know I've met you somewhere before, but I can't remember who you are" feeling. But when we start looking at the sorts of neurological mechanisms responsible for (2), it will become clear(ish) that even this experience can be explained in terms of the apparatus which performs task (2) without any need for a separate apparatus to perform task (1).

Also, once we look at the way the "list" is actually implemented, it will become clear that (3) is not a separate job waiting to be done after (2) has been done. Activating a representation happens as an outcome of recognising that the encoded sensation is a part of that representation, but is not a *separate* job done by an "activator module" acting on the output of (2). When we have a good picture of how a representation is implemented, and how such a representation is activated, then it will become clear that recognising the encoded sensation as a part of a representation also simultaneously activates that representation.

What this means is that the implementation of content addressable memory hinges crucially on job (2); once we have that , the other jobs will take care of themselves. Activating a representation relies on recognising that the currently encoded sensation is the same encoded sensation as that which is a part of a certain representation. In theory, recognising the currently encoded sensation (2) and activating the relevant representation (3) could be done in two ways:

(A) I could search through all my representations, until I find one which has the same encoded sensation associated with it, and activate *that* representation. (Hopefully this would be my **coffee** representation.)

(B) I could have some way of knowing what each encoded sensation encodes the smell of, so that I would know that *this* encoded sensation encodes the smell of coffee; knowing this I would then look for my **coffee** representation and activate that representation.

Unfortunately neither of these are practical. Take method (A) for a start. An exhaustive search of all my representations is just not humanly possible. At the speed my brain works (even on a good day) it could take me a week to search through all my representations to find a representation which has *this* smell associated with it. The time factor is brought home by what is called the "100-step problem". Individual neuronal processes take somewhere around a few milliseconds each to occur, and many basic cognitive tasks take somewhere around a few hundred milliseconds. From this it's easy to argue that to perform basic cognitive tasks, such as recognising the smell of coffee for instance, the brain can't go through more than about one hundred serial steps. To carry out searches like the above one, however, would require a lot more than one hundred steps. So serial searches through all my representations just can't be carried out in the human brain. Searching like this through a large set of stored representations may be how a computer could do this sort of job, and a fast computer may be able to search a large database and find the relevant entry (using well-designed hash codes, or some other efficient searching method) in a "reasonable" time, but we biological animals must do this sort of thing in some other way.

Let's turn to method (B) then. This way of recognising the currently encoded sensation and activating the representation it is a part of is just as biologically unrealistic. Here I'm supposed to have some sort of look-up table (like an index of encoded sensations) which lists every encoded sensation together with the representation(s) it is a part of. So the smell of coffee would be encoded in a certain way, and this encoded sensation could be looked up in this index of sensations, where some way of specifying my **coffee** representation would be found under this encoded sensation's entry. There are two points against this idea. One is that such a look-up table of sensations would be *huge*. Even if we had a different look-up table for each sense-modality, the total space occupied by all the look-up tables would be close to same size as the total space required to store all my representations. The second point is that although this might be a faster method than searching through all my representations for a certain encoded sensation,

scanning through the look-up table of encoded sensations could still take time. It seems to me that this could take well over one hundred steps to complete.

Seeing the above as ways of doing job (2) is a consequence of implementing a representation as a "list" of encoded sensations. This in turn is a consequence of analogising the human brain to a computer, which *stores* memories and representations at a certain place and retrieves them from storage by somehow comparing the sensations I'm experiencing at the moment with something stored. This is also a consequence of the fact that we haven't really considered yet how sensations could be encoded biologically. The solution comes from casting off this "list" and "address" perspective, and looking at connectionist networks and the model they provide of the way sensations might be encoded and representations implemented, in the human brain. Such networks paint a simple picture of the way that job (2) can be done, which does not require lengthy look-up tables or time-consuming searches (this new way does *resemble* the second method I just discussed though). I'll explain connectionist-style encoding of sensations presently. But first there is another change in perspective which I need to explain. This is an epistemological change with regard to perception and what the job of our perceptual systems are; a change which comes from switching to a parallel processing perspective.

2.4 *Ecological psychology–The flow of information.*

Recall the traditional theories' assumption (c), which I described towards the end of the last chapter. This uses the assumption that the task performed by the sense organs is to turn properties of objects into properties of representations. When I introduced this assumption, I mentioned that a change in perspective is called for. Here is where I shall try outline the more sensible way of looking at perception which I claimed that we need to adopt. This change in perspective is provided by examining one of the insights of J. J. Gibson's *ecological psychology*.[3]

Ecological psychology carries with it a way of viewing perception which is radically different from the way traditional theories of semantics view perception. Here perception is not about turning properties of objects into properties of representations. Instead it's about the *pickup of information*. We need to cast off the idea that perception is the processing of images, sounds, etc. to somehow deduce, detect or extract properties of objects from these sensations. Gibson's theory situates each perceiver in a

3 Gibson(1979) , pp238-263.

"sea" of luminous, mechanical, and chemical energy, which is awash with information. This information is information which specifies the objects in the perceiver's environment.

(Gibson's theory of information pickup is intimately tied up with using the information picked up to guide action. For Gibson, perception is inseparable from action, from being in the world: "Perception is a psychosomatic act, not of the mind or of the body but of a living observer."[4] What information a perceiver picks up on, and how it is taken to be information, is intimately tied up with the perceiver's history of embodied action, of using perception to guide actions. I shall put off integrating action into this account of perception until Chapters Four and Five, however. Here I shall take on board Gibson's idea that the sea of energy in which we live contains a rich diet of information to be picked up on. In Chapters Four and Five I'll fill out the details about how this information *comes* to have significance to the perceiver, how it comes to be information *about* objects in the perceiver's environment. I'll do this by bringing actions into the picture. Very briefly, the outputs of my sense organs encode information which is informative because of my lifetime of experience in *using* that encoded information to guide my actions. I'll have to ask you to take it on faith for the moment that the information picked up already has significance to the (experienced) perceiver; I'll explain how the significance comes to be there in Chapters Four and Five.)

Gibson holds that information about the organism's environment, in particular information uniquely specifying the objects present and the actions or uses they *"afford"* to us, is contained in the "optic array"; the light input to the perceiver's retina. Perception is about *picking up* on this information, so that it can be used to control action.

There are two main ideas here: For a start the information picked up is not about object's properties, but only about objects. And furthermore, this information is not about an objectively specified perceiver-independent reality, but is –in an important sense– perceiver relative.

What information an organism picks up on depends on what the organism is looking for. That is, I only need to pick up on information which relates to my goals, needs and desires. The information picked up specifying the uses particular object affords is perceiver relative: a tree affords climbing to me but not to a giraffe, and a leaf affords food to a giraffe but not to me. The information picked up depends on the organism's training too. I can learn to pick up on information which specifies affordances I was originally insensitive to; I did this when I learned to walk

4 Gibson(1979), p240.

and needed to lean on things, when learned that paper affords making paper airplanes, when I learned that books afford reading, and when I learned that cornflakes afford making marmite and cornflake sandwiches. Also perceivers which have different perceptual equipment will be able to pick up different information: bees can pick up on information from the ultraviolet radiation reflected from flowers that I can't pick up on.

This does not amount to total relativism though: Gibson maintains that *all* this information is continually available in the environment. It isn't constructed or deduced, all this information is *there*, to be picked up by perceivers and used to produce actions. The relativism I've just been talking about is found only in which bits of the sea of information that we all swim in each of us are able, trained or predisposed to pick up on.

Gibson and his followers suggest two types of explanation for the relations between perception and action. One is in terms of the coupling between organism and environment, which explicitly refuses to discuss internal representations. Another is to postulate the existence of internal mechanisms for relating perception and action, stipulating that these mechanisms–whatever they are– must be described in the language of physics.[5] Advocates of each of these ways of relating perception and action avoid getting their hands dirty; they don't attempt any sort of detailed explanation of how internal processes which use the information picked up to guide action might be neurologically implemented. One of the aims of my thesis is to "muck in" and provide such an explanation, while still being sensitive to ecological psychology's central tenets.

2.5 *State spaces*

A plausible and biologically realistic explanation of how my sense organs pick up information and encode it in a neurologically useable form has been presented by Paul Churchland. This method employs what he calls "state spaces." Churchland shows how a representational scheme utilising state spaces "can account, in a biologically realistic fashion, for a number of important features of motor control, sensory discrimination, and sensorimotor coordination."[6]

---

5 These explanations both reject the ides of any sort of inference or deduction being necessary in perception; perception is not cognitively mediated, but direct. According to Gibson, information is picked up *directly* rather than having to be inferred from the sensations an organism receives from its environment. A criticism of this direct perception is mounted by Fodor & Pylyshyn (1981). This criticism is rebutted by Aaron Ben-Zeev (1988); he does this convincingly, I'd say.

6 Churchland (1986), p305. To show how state spaces can accomplish this, he uses examples from colour discrimination (colour blindness), tastes (discrimination of bitterness in rats, humans, and cats), smells (a dog's ability to distinguish every human by their smell), recognition of faces, and body configuration and locomotion. pp299-305.

The way the sense organs perform their jobs can be compared to the way a thermometer does its job. Both are basically transductive processes. A thermometer transduces environmental information about the temperature at present from one medium to another: that information is transduced from the mean kinetic energy of molecules in the air to the height of a column of mercury. The thermometer encodes the temperature at present as the height of a column of mercury. No intellectual process is happening in a thermometer, no meaning is given to anything (that happens when the thermometer is read by a person, and what is read is interpreted). The process of converting temperature information so that it's encoded by the height of a column of mercury could be described by rules–quite simple ones as it happens–but no rule-*following* is happening here. The rules for converting temperature into the height of a column of mercury are *embodied in the physical constitution of the thermometer.* [7]

The same sort of thing goes for our perceptual apparatus. Churchland's description of state-spaces provides a good example of how basically transductive processes which encode in neurological form the information picked up from the environment, can be embodied in the physical constitution of human perceptual systems' physiological structures. (Churchland doesn't actually talk about "information", but his description of the way the senses could work is very receptive to such an interpretation.) I'll start with a relatively simple two dimensional example, to illustrate the concept of state spaces, and then I'll show how this can be extended to multi-dimensional state spaces.

Example One: Good examples of two dimensional state spaces can be found in special areas on the outer surface of the human brain, which Churchland says are of particular interest because they "plainly constitute *topographic maps* of some aspect of the sensory or motor periphery, or some other area of the brain."[8] One of these topographic maps is the outer layer of the somatosensory cortex. This map implements a two dimensional state space which encodes the tension and the position of the muscles, and the pressure on the skin. This map appears to be organised so that some neurons are devoted to encoding the position of various areas of the epidermis, and some the pressure at those areas. For the sake of simplicity, I'm going to focus (for now) principally on the aspects of this map which encode our sense of touch and pressure. (I will, for now, ignore the other function, which encodes the position and tension of the muscles. Of course, this is gross over-simplification, because to be *information*, tactile sensations need to be related to the position and tension of the muscles. Otherwise it's

---

7 This point is from Aaron Ben-Zeev (1988)
8 Paul Churchland (1986) p281.

just uninformative input. Imagine holding a tennis ball in your hand and trying to identify it without having access to information about the position of your fingers and thumb, or how hard you were squeezing it.) I'll just deal with the encoding of tactile *sensations* for now, and I'll show later how these sensations encode tactile information by being combined with the aspects of this map which encode the tension and the position of muscles to form what Gibson calls a *perceptual system*..

The somatosensory cortex is a topographical map of the body's tactile surface. It's like a map of all the body's nerve endings, stretched like a rubber sheet into a deformed shape over the outer layer of a specific area of the brain. It's stretched according to a scaling of the map with regard to the sensitivity of the areas mapped: larger proportions of the map area are devoted to mapping the more sensitive and more often used areas of the body. For example, the part which maps the right hand, is much larger than the part which maps the outer right thigh, even though the thigh has more skin surface area. This is because the hands are more sensitive and more capable of intricate movement; there are more tactile sensors on the surface of the right hand to map (and more muscles to map also), than there are on the outer right thigh. On this somatosensory map, the neighbourhood relations which hold between areas of the body are preserved in the neighbourhood relations between areas of the map. For example, the part of the map responsible for the right index finger, is a neighbour of the part responsible for the right palm. This means that any closed area of the body's tactile surface is mapped by a corresponding closed area of the map.

When the body's tactile sensors are stimulated by pressure this stimulation causes a certain spiking frequency to be sent along the nerves, which causes neurons at corresponding areas of this topographic map on the surface of the brain to "fire." This "firing" is caused by each "spike" of the spiking frequency carried to the termination point of the nerve, the axon, causing the axon to release a synaptic transmitter. A higher spiking frequency means that more synaptic transmitter will be released. This synaptic transmitter released from the nerve axon excites the dendrite (receptor part) of the neuron cell on the topographical map by creating an electrical imbalance in the dendrite. A neuron cell can have just a few or many thousands of dendrites. If enough of a neuron cell's dendrites are excited in this way, then the neuron cell's output part, the axon, is triggered by the aggregate electrical imbalance, and the neuron is said to "fire". This firing causes a spike (or a spiking frequency) on the map neuron's axon,

which causes it to release synaptic transmitters which excite the dendrites of other neurons and so on.[9]

The degree to which the neurons on the somatosensory cortex are excited depends on the spiking frequency transmitted from the pressure sensors on the body's tactile surface, and this spiking frequency, in turn, is proportional to the amount of pressure these sensors are exposed to. So: the *physical location* of the excited neuron on the map encodes where in the body the sensation comes from. The amount this neuron is *excited* encodes the amount of pressure the sensors at this part of the body are experiencing.

The way a tactile sensation is encoded is probably related to the *spread* of pressure as well as the pressure itself. The pressure spread on the surface of the body causes a spread of stimulation over an area of the map. This spread of stimulation could be detected by the spread of dendrites connected to a neuron in the layer below the input layer of the map. These dendrites are connected to the axons (outputs) of the map neurons which take input from nerve fibres. Some of these second layer neurons could take input from a small localised area of the map, so that their firing encodes very localised pressure, like that caused by a poke with a pin. Some other second layer neurons could take input from map neurons spread further apart, so that their firing encodes less localised pressure. By the firing of different second layer neurons (or even third or lower layers) all sorts of sensations be encoded: pressures of different inensity, on different areas of the body, of different shapes and sizes. Thus patterns in the firing of nodes in this map can encode all the body's tactile sensations, from a pat on the head, to standing on a small sharp stone; different sensations are encoded as stimulation of different patterns of neurons of this topographical map.

According to Churchland, there are several such topographic maps in the central nervous system. The primary visual cortex maps the stimulation of the retina in the eyes. Likewise the auditory cortex is a two dimensional topographic map of frequency space. And there are many other areas of the surface of the brain whose structural organisation makes it apparent that they map in the same way, but exactly what they map scientists are still unsure about.[10] So topographical maps, as two dimensional "state spaces" are used to encode sensations from lots of different sense modalities.

Example Two: Churchland also describes how with a slightly different architecture, state spaces can be employed to encode the outputs of more complex senses, which require maps of higher than two dimensions. The sense of smell provides a good illustration of this different sort of state space. Our sense of smell must pick up on information carried in the chemicals in

9 See Patricia Churchland (1986)
10 Paul Churchland (1986) : p282-283.

the air (this information specifies the type of object present), and encode this information in a form which can be processed neurologically. This processing must facilitate the recognition of smells, so that we can tell what kind of object the chemicals detected in the air carry information about. For example, the presence of coffee molecules (or whatever they are) in the air around me convey the information that there is coffee present in my environment, this information needs to be encoded neurologically by my olfactory sense organs. Churchland's discussion of state spaces shows how the outputs of my olfactory sensory system could be encoded in a form which facilitates their recognition, and discrimination, so that the smell caused by coffee can be distinguished from other smells, and recognised as the smell of coffee.

Churchland points out that humans have six different types of olfactory receptors, each of the six encodes a different aspect of a smell (they're sensitive to the shapes of molecules). He suggests that at a conservative estimate, each of these types of receptor is capable of distinguishing between ten levels of stimulation.[11] All of the receptors of each type are connected up together to a parallel fibre along which a spiking frequency passes which is related to the amount of stimulation these receptors are experiencing. There are six such fibres then. The spiking frequency on each fibre encodes the level of stimulation of one type of receptor. So when I pick up a smell, the properties of my smell-receptors' outputs are encoded as a set of six spiking frequencies: one frequency on each of these fibres.

This allows for a six-dimensional state space in which to encode smells. Just as the somatosensory cortex is a "map" which extends in two dimensions, this can be thought of as a map which extends in six dimensions. Imagine this state space as being similar to a cubic volume, but it's a volume which extends in six directions instead of three; each direction being ten units long. Each point in this "volume"–which in the technical terminology is called a *hyperspace*–encodes a distinct smell sensation. It is perhaps easier to visualise each state space point as a set of six numbers–the six coordinates of the point, as there would be three coordinates in a cubic state space: height, length and breadth. Each coordinate in the set of six encodes the amount of stimulation of a set of one of the six different types of smell receptor. This state space gives us each the power to be able to encode 10^6, distinguishable smells– that's one million of them, no small feat.

So when I pick up a smell, the properties of my smell-receptors' outputs are encoded as a set of six levels of stimulation, each encoding how

11 More likely its a continuous range, rather than one with discrete levels. If this is so, then please forgive the oversimplification, but a discrete range, with ten distinct levels, is easier to describe on paper.

much one type of receptor is outputting (due to its stimulation). Differences between smells are implemented as differences in position in this six-dimensional state space, and similarities between smells are implemented as neighbourhood relations in the state space. Different smells will be encoded as points far apart in the state space, and similar smells will be encoded as points in the same neighbourhood. So for instance the smell of fresh basil could stimulate my smell receptors so that the co-ordinates (8,3,6,2,0,9) encode this smell, ammonia's sharp smell could be encoded as something like (1,9,1,0,0,1) a point deep in a corner far away from the smell of basil. Similar smells like the smell of fresh espresso coffee and the smell of instant decaffeinated coffee could be encoded as (5,4,2,5,3,6) and (5,4,2,5,9,3) respectively, which in the state space are close neighbours.

I should emphasise that I don't actually have a six-dimensional hyperspace in my brain. The idea of a high-dimensional space, is no more than a convenient way of *visualising* the relationships between smells as neighbourhood relations in the state space. However, writing these state space points as sets of six numbers does correspond to some degree with what actually happens in the brain. Recall that I said earlier that this smell state space is implemented by a set of six parallel fibres, each conveying the outputs of a set of smell receptors; each of these coordinates is implemented by a specific spiking frequency on one fibre of this set of six. So the "5" in the first position of the above encoding of the smell of fresh coffee, would be implemented by the first fibre carrying a spiking frequency just above mid-range; and the "4" in the second position, is implemented by the next fibre in the set carrying a spiking frequency just a bit lower than its mid-range.

A parallel situation is happening with the implementation of the smell state space as we found in the way the somatosensory cortex is implemented. Earlier I described how the location in the somatosensory map was linked physically with the pressure-sensors at the part of the body stimulated, and the frequency on the fibres connected to the map at that point encoded what was happening with the pressure sensors at that part of the body. With the implementation of the smell state space, each of the six fibres is fed through a physical link from a different set of smell receptors, and the spiking frequency this fibre is carrying encodes the level of stimulation of the corresponding receptors. Smells are encoded as patterns of frequencies on these fibres. But unlike the somatosensory map, the smell state space only has to encode one thing: the smell itself; the somatosensory map has to encode both where in the body the sensation is coming from, and the "type" of tactile sensation being picked up at that point.

So here we have are two similar but different ways these state spaces are implemented. State spaces in general encode the information picked up

by the senses using *frequency encoding*; the information is encoded as a pattern of frequencies over a set of fibres. For an n-dimensional state space (where n is bigger than 2) the information encoded is implemented by the "pattern" of frequencies across a set of n parallel fibres. This pattern of frequencies can be thought of as the coordinates of a point in an n-dimensional state space. Two dimensional state spaces, in addition to frequency encoding, also use *spatial encoding* of information. Rather than the frequencies on the parallel fibres encoding the *coordinates* of the point in the "map", the frequencies are on fibres actually *at* each point in the two dimensional map. The information is encoded as a frequency *spread*, or a spatial pattern of frequencies; it's encoded in the physical location on the topographical map of *all* the fibres which are carrying spiking frequencies, as well as by the frequency those fibres are carrying. For some perceptual systems then, the information will be encoded as a spread of frequencies over the *entire* two-dimensional map. For example, information picked up through retinal images will be encoded like this. Information picked up through the tactile sensory system, by contrast, will be encoded as a frequency spread over a localised area of the map, rather than over the entire map.

Thus in either case, when we look at the information encoded through state spaces at the implementation level, we're looking at information encoded as patterns of frequencies over sets of parallel fibres. These sets may be all the fibres which implement the state space, as happens in information picked up by my olfactory perceptual system, or it might be patterns of frequencies over small subsets of the entire set of parallel fibres, as information picked up through my tactile perceptual system would be.

2.6 *Perceptual systems as* active *gatherers of information.*

It could appear from the descriptions that I've given that a perceptual system is passive and static, where all that matters is an instantaneous "snapshot" of the activation of receptors and the transmission of nerve impulses through the relevant parallel fibres to neurons in certain areas of the brain. I yet haven't paid any attention to what Gibson calls the *adjustments* which need to be made by a perceptual system in order to pick up information, nor to the information picked up from the *flux* of ambient energy.[12] The static, passive appearance of the way I've been talking about encoded information is an unfortunate consequence of using state spaces as an introductory apparatus. The time has come to kick the ladder away, so to speak. We need to

12 Gibson (1979), pp 244-246.

concentrate more on the idea that the patterns of frequencies on fibres encode information, rather than sensations.

Gibson sees a perceptual system as an *active* system. The visual perceptual system is continuously making adjustments which serve the pickup of information: the eyes move in their sockets, the iris dilates and contracts to adjust light intensity, my two eyes work together to focus on the same spot, my head and body move as well; all this together makes up my visual perceptual system. All these elements work together in a continuous loop of input and output. This perceptual system doesn't passively receive stimuli, it actively obtains information. A normal act of visual attention is not the noticing of a single detail, but the active *scanning* of a whole feature of the ambient optical array. Basically this comes down to Gibson's comment that perception involves the continuous "coperceiving of the self."[13]

Because of this we should realise that in order to pick up information and encode that information neurologically, we need to incorporate information about the various adjustments made by the body to facilitate information pickup. For example, the position and tension of various muscles needs to accompany input to the somatosensory cortex for tactile information to be properly picked up. Recall that I asked you to imagine feeling a tennis ball without information about the position and tension of the muscles in your hand. To use tactile information to identify an object, or to pick up information which might aid in identifying it we need to incorporate this information. And as I mentioned earlier, the somatosensory cortex *does* receive information about the position and tension of muscles as well as sensations caused by pressure-detector stimulation. All of this is encoded as frequencies on fibres. Tactile information is encoded then, as patterns of frequencies caused by *all* of this: pressure detectors, muscle tension and muscle position. So when I speak of some piece of information being encoded as a pattern of frequencies over a set of fibres, it should be appreciated that perhaps not all these fibres will be at the same place in the cortex. Visual information, as well as being picked up through the activity of the retina, is picked up and encoded through the activity of the rest of the visual perceptual system. Thus patterns of frequencies which encode the position of the muscles which control eye focus and direction, those which control head and neck position, the body orientation and more will all be part of the way the information picked up is encoded. So when I talk about visual information being encoded as a patten of frequencies over a set of fibres in the brain, I mean a set of fibres which also includes the fibres connected to these muscles and so on.

---

13 Gibson(1979), p240.

Another factor we should consider is that a perceptual system picks up information from the *flux* of energy, through noticing how things change, and how they change or stay the same as the perceptual system changes. With the visual perceptual system for instance, information is picked up through noticing what parts of the ambient optic array are changing, and how they are changing. Information about movement is gained in this way. Information is also picked up by the perceiver's movement, by noticing how the optic array changes, and which bits don't change too, as the perceiver moves. Information about size, shape, relative position, and occlusion is picked up in this way. The descriptions I've given so far of the way information is encoded have made everything seem quite static. It's important to appreciate that a lot of information is encoded as *sequences* of patterns of frequencies over sets of parallel fibres. Another good example of the way information is encoded in changes in environmental energy is the information conveyed through sounds. Taking an instantaneous snapshot of the frequencies of vibration in the air will be particularly uninformative. Only when we attend to the sequences of vibrations in the air can we pick up on the information conveyed through sounds.

As Gibson says, information pickup requires perceptual systems, not senses.[14] And perceptual systems have enormous powers for encoding information as patterns of frequencies on certain fibres. They convert information conveyed in environmental media–in the ambient luminary, mechanical and chemical energy–into a neurological medium: certain brain fibres carrying specific frequencies. I'll now move on to discuss how information implemented as patterns of frequencies supports discrimination between, and recognition of the different bits of information picked up by each perceptual system.

2.7 *Connectionist pattern recogniser networks*

A connectionist picture of the brain illustrates how certain types of brain mechanism are able to recognise a particular pattern of frequencies outputted by my smell receptors, and realise that this pattern of frequencies encodes information which is a part of my **coffee** representation. And as I'll show in the next chapter, connectionist networks provide an illustration of a perfectly mechanical way in which the present outputs of my smell receptors being implemented as a certain pattern of frequencies can cause my **coffee** representation to be activated. But for now spend some time explaining what connectionist networks are. This will help explain how the

14 Gibson(1979), pp244-246.

neurological structures in my brain *could* do the job of recognising patterns of frequencies like those which encode information picked up by my perceptual systems.

The structure of a connectionist network is based on that of the interconnected networks of neurons in the brain. The theory is that the networks of neurons in the brain work in a similar to connectionist networks, and could do the sorts of jobs we've been able to get connectionist networks to do. A connectionist network is formed from arrays of small independent units, called "nodes." A node is a very simple processor whose only job is to take on a certain level of activation. A node takes input from other nodes, calculates a level of activation according to some function of its inputs, and then outputs this level of activation to other nodes (which is called "firing"). A node generally has input connections from all the nodes of the previous layer and its output is connected to all the nodes of the next layer. An example of such an arrangement is pictured in figure 2.1.

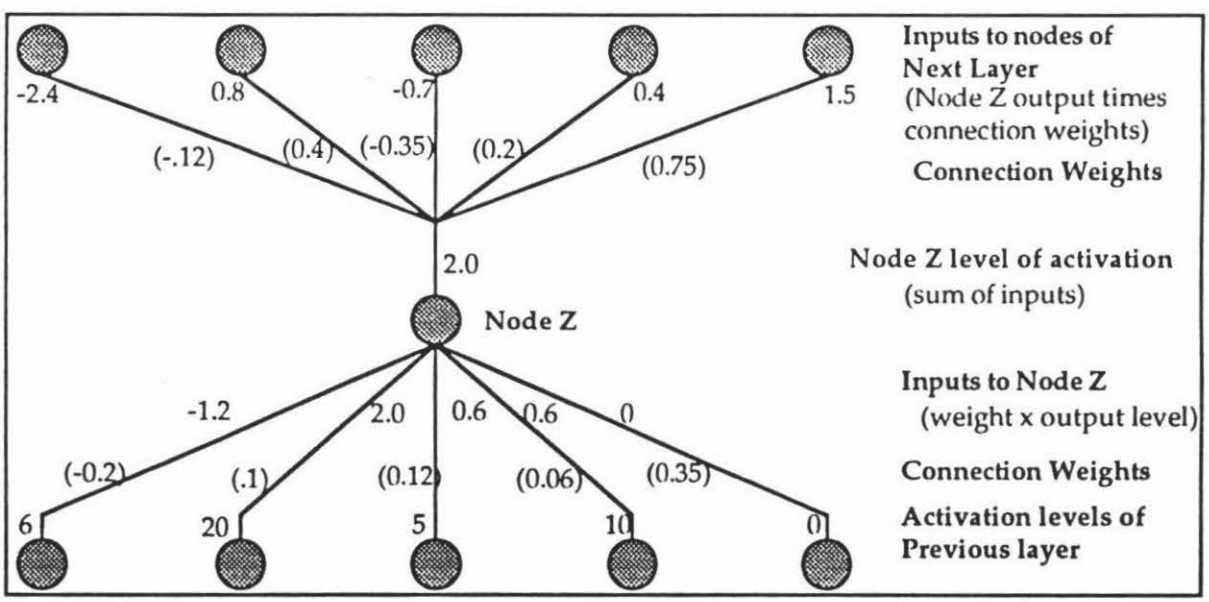


Figure 2.1 A node of a connectionist network and its connections.

Each node in the previous layer has its output modified by the "weight" of the connection between it and the node Z. This weight is simply a multiplier function. For example, the node on the bottom left has an activation level of 6 and the connection between that node and node Z has a weight of -0.2. Because of this, node Z receives from this connection the previous node's activation level (6) multiplied by the connection weight (-0.2). That is, the input it receives from this connection is -1.2. Node Z adds this input of -1.2 to the input it gets from all the other nodes which feed it, to get its total input. Node Z's activation level results from performing some function on it total input. Often this is a non-linear function, like a

threshold; if its total input is above a certain value (the threshold), then it will take on an activation level equal to this sum of inputs; but if its total input is below the threshold then it will produce zero as its output. For example, Node Z could have been set up with a threshold of zero, so that its activation level will be equal to its total input only if this total input is positive, otherwise it will be zero; thus the node will only "fire" if its total input is above zero.

As can be seen from the above diagram, the weight of a connection between nodes can be either negative or positive. A positive connection weight leading to node Z will cause a node to help "excite" node Z by increasing its total input, and a negative weight will make the total input to node Z lower than it would otherwise have been, and thus "inhibit" its firing.

A three layer network (as pictured in figure 2.2) is one of the most common configurations of these nodes, with layers of input, hidden, and output nodes. Each node of the layer of hidden nodes is wired up in the same way as node Z is in figure 2.1; for each node, its inputs come from the activation levels of all the nodes of the previous layer, and its activation level is input for all the nodes of the next layer. Again all the connections between nodes are weighted, but to avoid overcrowding the diagram I haven't drawn the weights in.

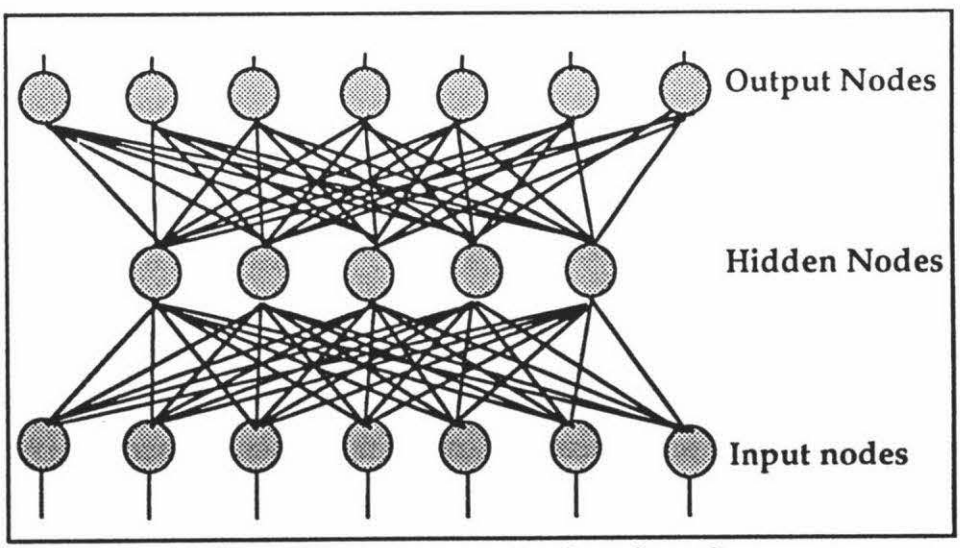


Figure 2.2 A common network configuration.

Such a network functions with all nodes working "locally", adding their inputs and producing the appropriate outputs. Each node is a processor working independently, of all the other nodes. There is no global control system (such as the CPU in a serial computer).

The important and interesting feature of these networks is that a network like this can come to have an arrangement of connection weights which will cause it to generate a particular "pattern" of activation levels across its output nodes only when presented with a certain pattern of activation levels across its input nodes. Indeed, a connectionist network like this can be trained to produce a particular output for each of many distinct input patterns. For example, the network pictured in figure 2.2 could be trained to produce the output activation pattern (2,1,0,1,2) only when given as input the pattern of activation levels (1,2,3,4,5), to produce (0,1,2,1,0) as output only when (5,4,3,2,1) is its input, and also to output the pattern of activation levels (1,1,1,1,1) only when its input is the pattern of activation levels (1,1,1,1,1). It could be trained to recognise and distinguish between these three input activation level patterns, demonstrating its recognition of each input pattern by the pattern of activation levels across its output nodes. Thus it can be said to "recognise" patterns in its input, and identify them by producing the characteristic output it has been trained to associate with each input.[15]

Such networks, usually starting with randomly assigned connection weights, are trained by being given examples of input patterns of activation, and being allowed to produce an output pattern of activation. The output is compared with the desired output, and then the weights of connections are altered in order to make the output closer to the desired output. This adjustment of connection strengths is done automatically, by a mathematical principle called the *delta rule*. This rule "back propagates" the errors –i.e. the difference between the desired output and the actual output at each node) from the outputs back to all the connections in the network. As it were, it apportions the blame for errors in the output patterns on the appropriate connections, and calculates the minute adjustments which need to be made to the weights of those connections, so that the error is a little less next time. Thus the nodes which fired too strongly, or which fired when they shouldn't have, have each of their input connection strengths reduced by a small amount so that their total input will be less the next time this input is presented. And those nodes which didn't fire when they should have, or which did fire, but whose activation level wasn't high enough, have their input connection strengths minutely increased so that they will have a higher input activation level next time. The amount the weights are adjusted is calculated by the delta rule. This will make them more likely to fire at the level which will make the pattern of activation on the output nodes closer to the desired pattern. This delta rule is nothing mysterious, it's

15 For a good discussion of the fact that recognising patterns is what connectionist networks do best, see Bechtel and Abrahamsen, Chapter4, pp106-146.

a perfectly mechanical series of computations. It does, however, require some fairly heavy-duty mathematics to explain fully. (Bechtel and Abrahamsen provide a fairly down-to-earth attempt at such an explanation on pp76-97.)

I'll try to quickly explain this learning rule by illustrating how one cycle of the application of this rule to a simple network might go. Pictured in figure 2.3 is an output node with two of its input connections. The rule for changing the weight of a connection is this:

The change in weight = the learning rate x (desired output – actual output) x the actual output of the previous node.

Or: $\Delta\text{weight} = \text{lrate}(d_u - a_u)a_i$

The "learning rate" is a constant set by the network's trainers. The larger the learning rate, the bigger the changes in the connection weight. A network generally needs a fairly low learning rate, because often networks are learning to recognise more than one pattern at a time, and we don't want the delta rule's application to cause large errors in the output for one pattern by making big changes to connection weights while learning a different pattern. It's best for a network to learn slowly. I've given the example in Figure 2.3 a learning rate of 0.1.

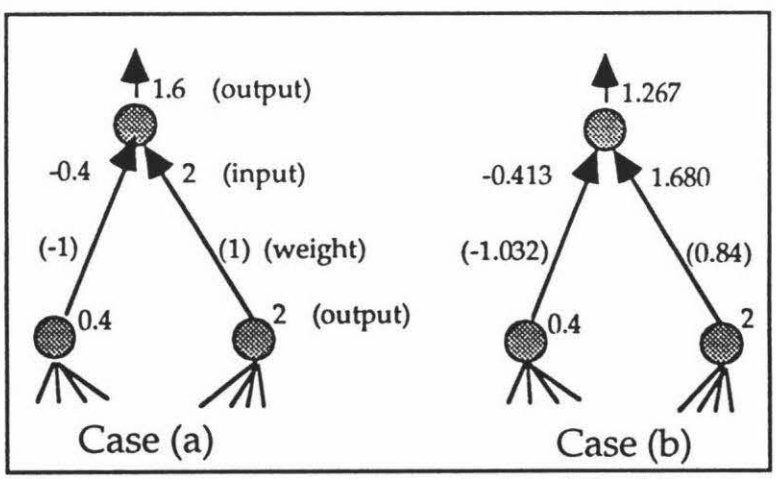


Figure 2.3 An example of the application of the delta rule

For case (a) in Figure 2.3 let's say that the desired output for this node is 0.8. The actual output of 1.6 is too high. By applying the delta rule to each of the weights of the connections pictured, the output will be lowered. For the left connection, the rule is applied by multiplying the learning rate (0.1) by the difference between the desired and actual outputs (0.8-1.6) and multiplying this by the activation level of the lower down node (0.4). Thus

$\Delta\text{weight} = 0.1 \times (0.8 - 1.6) \times 0.4$

$= -0.032$

Thus the weight on the left connection is decreased by 0.032, to 0.978. For the right connection the same formula is applied:

$$\Delta\text{weight} = 0.1 \times (0.8 - 1.6) \times 2$$
$$= -0.16$$

So the right connection weight is also decreased, this time by 0.16 so that it becomes 0.84.

(In reality the delta rule would now be applied to the weights of the connections which carry input to the two lower nodes pictured, which would lower the input activation levels of these two nodes next time. But in order to show the effect which the delta rule has on the connection weights between these nodes and the upper node, I'm going to assume that there is no change to these lower connections, so that the output activation levels of the lower nodes are unchanged.)

Changing the connection weights by the amount calculated gives case (b). The actual output activation level of 1.267 is now closer to the desired output of 0.8. By repeated applications of this rule, the network's output progressively becomes closer and closer to the desired output.

An often-cited example of a successful connectionist network is one which was trained to examine radar echoes from undersea objects, and to distinguish whether the echo came from a mine or a rock.[16] Its inputs were the relative strengths of the different frequencies (the *frequency response*) of a reflected radar signal. This network (see figure 2.4) had two output nodes; one was to indicate (by taking an activation level of one with the other being zero) the network's recognition of a mine-echo, and the other to indicate (in the same way) the network's recognition of a rock-echo. Thus the expected output activation pattern was either (1,0) or (0,1).

16 Paul Churchland (1986) : pp 202-204.

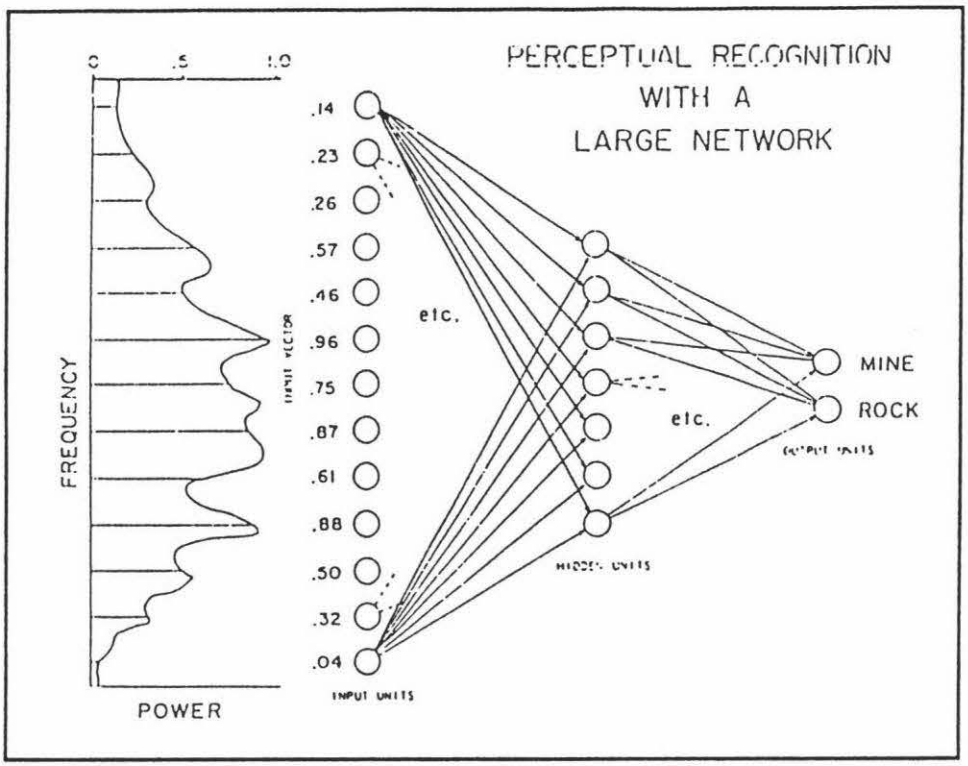


Figure 2.4 The rock/mine detector[17]

This network was shown examples of the frequency responses of echoes caused by mines and by rocks; the input nodes' activation levels corresponded to the amount of reflected signal at graduated frequency intervals. These activation levels passed through the network's connections (being multiplied by their weights) and caused activation levels on a hidden layer which then, in turn, on the two output nodes. Whenever it gave an incorrect response (which was very often, in the early stages of its training), the correct answer ((1,0) or (0,1)) was applied to the output by the network's trainers, and the errors were "back-propagated" using the delta rule through the network to adjust the appropriate connection strengths. As the network was trained, its errors slowly became less and less, both in size and in frequency. After extensive training, being shown the frequency responses of many thousands of mine-echoes and rock-echoes and having its errors corrected, this network eventually learned to tell the difference between mines and rocks from the frequency content of their radar echoes.

At this stage the network could also generalise its learning to echoes which were not part of the set of examples it was trained with. That is, it could correctly identify echoes it had never seen before as either being caused by a rock or by a mine.

There is now an extensive literature about how connectionist networks can be trained to recognise patterns. By using the delta rule to tailor the connection strengths, a network can learn to distinguish between

17 This diagram is stolen from Paul Churchland (1986), p159.

many distinct patterns of input, and generate a distinct output pattern for each distinct input pattern.

2.8 The flow of information.

It seems that jobs somewhat like this are doable, and done, by mechanisms in the human brain as well. They might not be done in *exactly* this way, and they almost definitely aren't trained in the same way (they're not trained against large sets of unequivocal examples and non-examples, and they may back-propagate errors, but will probably not use the delta rule when doing so). But it seems that there are brain mechanisms which do this job of learning to recognise patterns too. Some parts of the massively interconnected networks of neurons in the brain seem to implement pattern recognisers very similar to connectionist networks. There are networks of neurons which are capable of taking a certain pattern of activation as input, and producing a characteristic pattern of activation as output for each (familiar) input pattern–effectively "recognising" that pattern,

A good place to look for pattern recognisers like these is at the outputs of the sense organs. It is postulated that for each perceptual system we have a pattern recogniser (or several pattern recognisers) dedicated to assessing the pattern of frequencies on the set of fibres which implement that state space. Each recogniser network produces a characteristic output pattern for every "familiar" pattern of frequencies it takes as input. Viewing our sensory systems as employing these connectionist pattern recogniser networks, we can see how *information* available in the environment is picked up by a perceptual system and carried through into the representations themselves. This contrasts sharply with traditional theories' treating the senses as a boundary, where properties of objects in my environment are *converted* into structured information contained in the formal properties of representations. To illustrate the way information can permeate through from the environment to representations, I'll discuss in detail three examples.

Example One: The pattern recogniser which processes the patterns of frequencies on the fibres which implement my smell state space, would produce a characteristic output for every distinct smell I'm able to recognise.

(In order to simplify the explanation of how this can happen, I'm going to assume (for the moment) that these pattern recogniser networks signal their recognitions of input patterns by the firing of a particular output node (much as the rock/mine detector did), rather than a pattern of activation levels of all the network's output nodes. So for example, I'll say that the network which recognises the outputs of my smell sensors, encodes

its recognition of an encoded sensation by activating a particular output node.)

The pattern of frequencies (5,4,2,5,3,6), which encodes the information that there are coffee molecules in the air around me, would cause the pattern recogniser to produce a distinct output pattern of activation which it only produces when it receives this particular pattern as input. This network could recognise the smell of coffee, and distinguish it from the smells of cowdung, mown grass and fresh basil. Each of these other smells would be implemented as further patterns of frequencies which when processed by the recogniser network would each cause a characteristic pattern of activation over the network's output nodes.[18] In this way the network could produce a different output pattern to indicate its recognition of each smell familiar to me.

What this means is that I could have brain mechanisms capable of doing job (2) –recognise that this sensation encodes the smell of coffee (see section 2.3)– by having networks of neurons similar in structure to a connectionist pattern recogniser as parts of my brain. These mechanisms would be capable of recognising each piece of encoded information I'm familiar with, and generating some characteristic output to indicate the recognition of this information. In fact, notice that by being processed by such a recogniser network, jobs (1) and (2), the recognition of an encoded smell as a familiar one together with the recognition of what it is the smell of, happen simultaneously. They are not separate tasks. If the smell is one which has been experienced before, then "realising" that it's a familiar one and "recognising" what it is the smell of are the same event: namely the activation of a certain output node. The next step then, is to explain how the output of such a network manages to activate the appropriate representation.

Example Two: Now consider one of the networks which recognises certain features of the sounds we hear. There are perhaps four major networks which process and recognise sounds, each dedicated to picking up different aspects of sounds, such as speech sounds, musical melodies, voice characteristics, and known sounds such as telephones, the tone qualities of musical instruments, cars, spoons stirring coffee in mugs and so on.[19] The network which recognises speech sounds will do fine for this purpose.

The inputs to this network are the raw unprocessed properties of the auditory sense organs, encoded as spiking frequencies on fibres in the auditory cortex. Certain patterns in this network's outputs could be

---

18 I'm using "nodes" instead of "neurons" to indicate that this system is a *model* of what could occur in my brain. The "nodes" I mention could be neurons, or nodes in an artificial system. I want to leave this ambiguous.

19 This division is from Martindale (1991). Most of this discussion of the processing of speech sounds is based on apparatus Martindale sketches on pp 53-55.

produced whenever it recognises a feature of human language in the encoded sounds it takes as input. It's been suggested that there are about ten distinctive and perceivable features of a spoken sound: such as being voiced or unvoiced, being a consonant sound or not, being an explosive sound (like that produced by pronouncing the sounds made by t, p, b, k) and so on. So this network would have outputs whose activation encodes each of these features; outputs which fire whenever that particular feature is recognised. The outputs produced by this network would encode the features that are present in the sound encoded. For example to encode the *p* sound in "dip", the recognition of this sound would cause the activation of the output nodes which encode a stop, an unvoiced sound, and whatever other features are necessary to distinguish this sound from the other sounds of human language.

This network's output is fed to a deeper network layer, which recognises patterns in the outputs of the above feature recogniser network. This deeper pattern recogniser network would take as input the encoded features of sounds, and have outputs which encode the phonemes each combination of features makes up. Phonemes are small units of vocal sounds. The *ee* sound in "deep", is a phoneme, and is a different phoneme from the *i* sound in "dip". There are 30 to 50 different phonemes of human languages, so this network would need to have between 30 and 50 different output patterns–one pattern to encode each different phoneme.

This second network's output activation patterns, which encode phonemes, could be the inputs for a third network which recognises patterns in sequences of phonemes, and which has output nodes whose patterns of activation encode syllables. This network would need to have about 10 000 distinct output patterns to encode all the syllables I use. And at an even deeper level could be a network which takes as input the outputs of the syllable recogniser network, and which recognises patterns in sequences of encoded syllables, and produces output patterns which encode the words which this sequence of syllables make up. It would produce as its output a distinct pattern to encode each word known to me.

Through such a series of networks the information encoded in environmental sound waves is processed to extract, or recognise, relevant features, and to encode them in a form which can be used to activate representations. My hearing someone say the word "coffee" would cause this series of networks to transform this piece of environmental information from being implemented as pressure waves in the air into a piece of information encoding a sequence of sound features, phonemes, syllables, and finally words. In this way the node which encodes the word "coffee"

comes to be activated, which causes the activation of my **coffee** representation.

Example Three: A similar example of this sort of process is found in our visual pattern recognition systems. Patterns of light intensities of specific frequency ranges are encoded as activations of nodes in the visual cortex. In the human visual system we have several different layers of pattern recognisers, each layer recognising more complex patterns, and thus extracting more complex information from the optic array.

The first layer takes as input the raw unprocessed properties of the visual perceptual system, encoded as activated neurons in the visual cortex. (I'm ignoring all aspects of the visual perceptual system except the encoded light itself, for the moment. I'll come back to it soon.) The visual cortex is a two dimensional topographical map of the rods and cones in the eyes.[20] Activated nodes here correspond to whether or not light of a certain frequency range is being picked up by a certain rod or cone in the retina of a particular eye. One layer of cortical neurons down are nodes which extract information from these patterns of light intensities. These nodes look for edges and lines in the patterns of activated neurons.[21] The activation of a certain output node encodes the detection of a line of a certain orientation at a certain place in the visual field.

This output is fed to another deeper layer, which recognises patterns in the outputs of the above line detectors. This pattern recogniser would take as input the encoded edges and lines, recognise combinations of these features, and have outputs which encode certain basic shapes, such as angles, circles, curves and rectangles of certain sizes and orientations. The output activation patterns of these first two networks are the inputs for a third network which recognises more complex patterns in the shapes and lines encoded. And at an even deeper level could be a network which takes as input the outputs of these first three, and processes these to extract information about shapes, lines, curves and so on.[22] And as we get deeper, we come to nodes whose activation encodes visual information about more and more complex shapes, like hands, pens, chairs, grandmothers and ice-creams. [23]

Thus I have a *grandmother* node whose activation encodes the information that my grandmother is present; information that was encoded in what Gibson called the "optic array," and was picked up by my visual

---

20 The visual cortex is actually a set of four two dimensional topographical maps of the rods and cones in the eyes. The four maps are compared to each other to produce colours. See Land (1977) for more details. For our purposes, it suffices to consider these as just one big map.

21 Poggio (1984) explains how the neurons here can detect edges, by physically he what he calls a "Mexican Hat" function.

22 For a more detailed discussion of these layers, and how they work see Ornstein and Thompson (1985) pp43-58 and pp103-129; and also Martindale (1991) : pp39-44.

23 Ornstein and Thompson (1985) : p57.

perceptual system. Of course, contained in the optic array is more than just the information that my grandmother is *present*. The optic array contains the information that my grandmother is sitting on a chair in front of the TV about three paces away from me. To encode and pick up this information I need to include information from other parts of my visual perceptual system, information about where my eyes are focussed, which direction they're pointing in, which direction my head is pointing, how my body is moving at the moment, and so on. I also need to include other visual information picked up by pattern recogniser networks; information about the chair, sitting, distances, and the TV.

With connectionist perceptual systems such as these three, the information already present in the environment is converted from one carrier to another, from light waves, sound waves, and other environmental information carriers to the activation of nodes in networks; the information itself percolates through from one medium to the next.

CHAPTER THREE

CONNECTIONIST REPRESENTATIONS

Now that we have a picture of how the senses might work, what their job is, and how connectionist pattern recognisers work, we can move on to discussing connectionist representations and how these pattern recognisers can activate such representations. A connectionist perspective revolutionises the traditional picture of what a representation is, how the parts of a representation are organised, and how it can represent. In this chapter I'll illustrate connectionist representations, and show that this connectionist version of what a representation is has some important advantages over the traditional picture. We'll end up with an account of what a representation is and how it's activated which is drastically different from the way in which many traditional theories view representations and the way they function. It gives us the equipment to make distinctions, explain cognitive capacities, understand subtleties, and account for phenomena the traditional theories aren't equipped to deal with. Armed with such equipment, the problem of representation and misrepresentation becomes much easier to deal with.

A further advantage of connectionist representations is that many characteristics which emerge as side-effects of implementing representations in this way are characteristics which are exhibited by the human mind. This indicates that connectionist representation and misrepresentation is biologically realistic.

Disclaimer: I should make it clear that I'm not claiming that this chapter provides a description of how our mental representations actually *are* implemented; they might turn out to be implemented in a different way. What I mean to do here is give a description of a way that representations as I depict them *could* be implemented in the human brain. I intend this description to be both as simplistic and as neurobiologically realistic as possible. These may turn out to be slightly conflicting goals. The necessity of simplification may make it sound like I'm doing "folk neurology". But by this perhaps slightly "folky" explanation, I hope to accomplish two things: In this Chapter I want to show that a representational and perceptual system like the one I'm outlining fits quite well with and explains a lot of our cognitive and perceptual capabilities. And by the end of this thesis I aim to have shown that it also gives us the equipment to provide a satisfactory

account of mental representations and a solution to the problem of misrepresentation.

3.1 Connectionist representations

Activating a representation is a bit like playing one of those television competitions where you're shown a picture of some famous person's eyes, and you can win a prize by figuring out who the eyes belong to. You need to recognise *where* you've seen those eyes before, plus distinguish them from all the other sets of eyes you've seen, plus picture the rest of the face those eyes normally belong with. Activating a representation is very similar: you experience a *part* of some entity and you have to recognise this part, plus distinguish it from all the other parts of this type you've encountered, plus somehow recall the *whole* which this is a part of. Take recognising a person on the street for instance. You see them, and picking up this visual information *somehow* activates your representation of the person, so you can recall who they are, what their relationship to you is, (hopefully) what their name is, and so on.

In section 2.3 I described two ways in which the smell state space point (5,4,2,5,3,6) being encoded by my smell sensors' outputs could be used to activate the representation this encoded smell information is a part of; both of these ways I concluded then were impractical. A search of all the representations (each a "list" of encoded sensations) in order to find those which have this encoded smell as a constituent would take far too long. And having some sort of index of encoded sensations is almost as impractical. This list would be very long, and would probably still take a good while to search through. These impractical methods looked like the only alternatives because I was stuck in a mind-set where I thought of a representation as a "list" of all the encoded sensations associated around some central theme, as might be stored in a serial computer.

Breaking out of this perspective, and taking connectionist networks as a serious and useful approach to understanding the brain, gave me a refreshingly fitting picture of what a representation is, and what its proper parts are. This revelation occurred when I came across the concept of a *distributed representation*.[1] I'm going to lead into this slowly though, via the concept of *local representations*. (The difference between local and distributed representations is caught up with the question of whether the output of a pattern recogniser is the activation of a single node (local), or a

1 The name "distributed representation" comes from Hinton, McClelland et al. (1986).

pattern of different activation levels distributed over a set of nodes (distributed).)

Local (and distributed) representations revolutionised my whole idea of what it is to be a representation, and of what it is to be a "part" of a representation. Being a part of a representation is tied up with the output nodes of pattern recogniser networks. The smell recogniser network signals its recognition of the smell encoded by the state space point (5,4,2,5,3,6) by activating a certain output node on receiving that pattern as input. This output node's activation could be said to be another way of encoding that piece of information. The information is first transduced, from being encoded as the shapes of molecules in the air around me, to being encoded as a pattern of frequencies on a set of parallel fibres. Then it is further transduced from being encoded as a pattern of frequencies on a set of parallel fibres, to being encoded as the activation of a certain node in the output layer of a recognition network. It is the same piece of information each time, but it's encoded in a different form. Using this idea we can keep the notion that a representation is in some sense "made up" of "parts." But its not a "list" of sensations somehow encoded neurologically. Rather it's an association of encoded pieces of information, which information is encoded as the activation of certain nodes.

In short, the representation itself is just a large set of activated nodes. When activated, these nodes encode all the information which constituted the "list" I mentioned in section 2.2. But now we have encoded bits of information rather than encoded sensations. Each sensation I included before in the "list" is now properly to be seen as a piece of information encoded by the activation of a certain node. And rather than these encoded sensations being "somehow" associated together, these nodes can be *physically connected* together, so that the activation of one node can send an excitation signal out which results in the activation of all the other nodes which are constituents of that representation.

The best way to illustrate the way such a representation is organised is to return to the example of my **coffee** representation, and show how all the encoded sensations fit together. This representation is a physically interconnected set of all the nodes whose activation encodes the different bits of information I've learnt through experience to associate around the central theme of coffee. So this representation would have activated nodes as its constituent parts, nodes whose activation amounts to recognising all the information which I earlier included in the list of encoded sensations. My **coffee** representation as a whole is just this *interconnected network of activated nodes*. It might be pictured as something like figure 3.1.

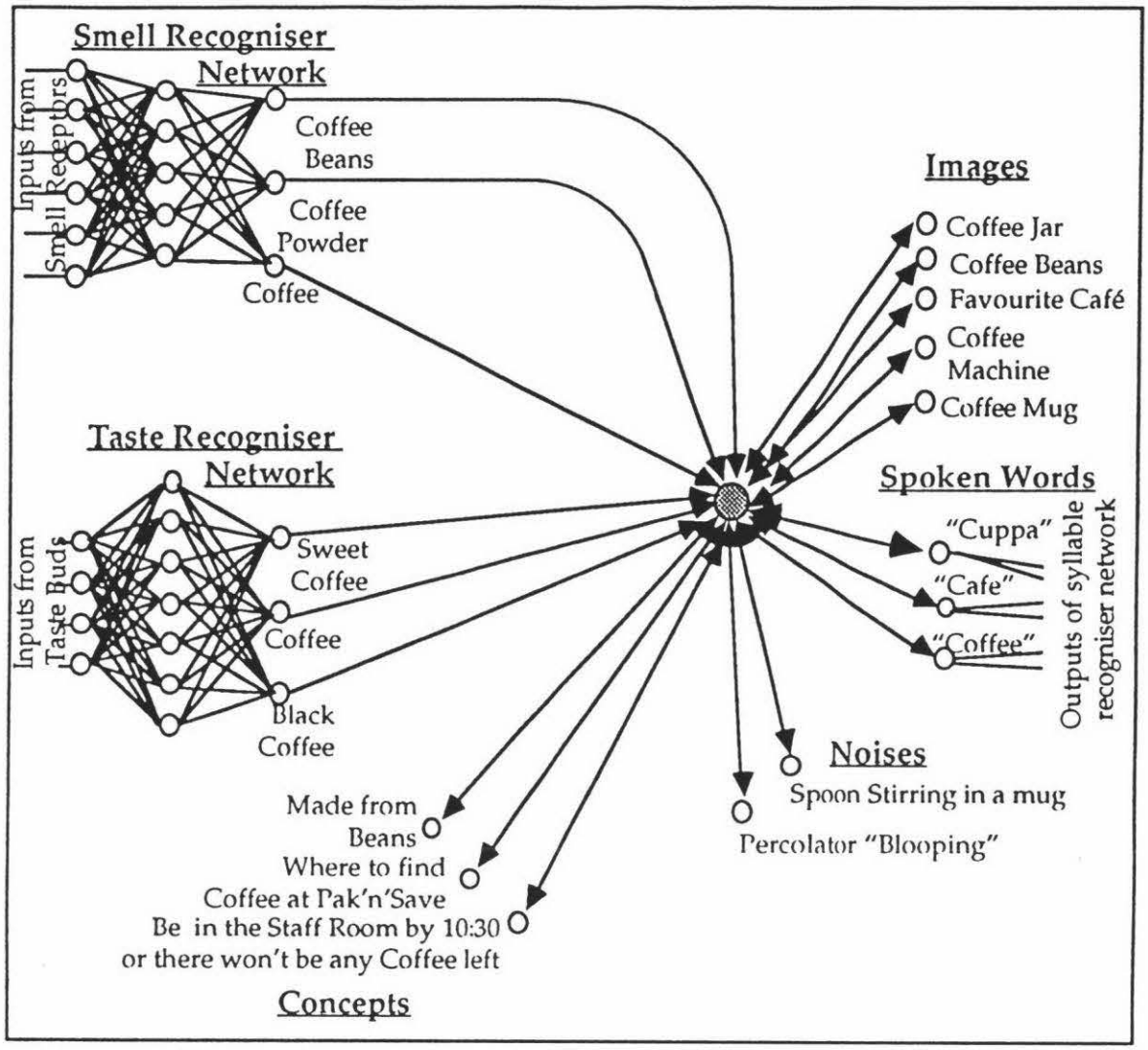


Figure 3.1 A connectionist depiction of my coffee representation.

The arrows in this diagram indicate inter-activating physical connections between the nodes at each end of the arrow. If a node at one end of a connection is activated, then this sends an excitation signal to the node which is at the other end of the connection, which activates this second node. And then this node sends excitation signals to all the nodes which are connected to it, and so on; the activation spreads so that this interconnected set of nodes are all activated.

It's important to realise that a node doesn't necessarily have just two states: fully activated or not activated at all. There is a spectrum of possible levels of activation. A node may be only weakly active, very strongly active, or somewhere in between. This is important because the strength of one node's activation can affect the activation of other connected nodes. A node's being activated doesn't guarantee that other nodes connected to it will be activated. Whether or not one node being activated will cause the activation of a second node depends on the activation level of the first node and the weight of the connection between them. For example, the first node may be only activated weakly, and thus may not send an excitation signal

strong enough to activate the other node. But if this second node is also connected to a third weakly activated node, then the combined weak excitation signals from nodes one and three may be enough to activate node two.

Note also, that a node's activation can also send out *inhibitory* signals which suppress the activation of certain other nodes. To avoid overcomplicating the above diagram I haven't attempted to draw these in. These inhibitory signals are implemented through having connections between nodes which have negative connection weights.[2] These can prevent certain nodes from being activated, which could happen when mutually incompatible pieces of information are encoded as constituents of the same representation. For example, the activation of the node which encodes the taste of sweet coffee is mutually incompatible with the activation of the node which encodes the taste of "normal" coffee: milky, strong and not sweet, the way I like it. If my sipping a cup of coffee activates one of these nodes, I can be sure that the other will not be activated. But if either one of these nodes were activated, my coffee representation would be activated (if it hadn't already been activated) without activating the other node. Thus the activation of one of these sends inhibitory signals to the nodes which encode other tastes I know this smell is not compatible with. Inhibitory signals like this occur all over my brain, between representations as well as between constituents of representations. This ensures that the spreading activation only spreads so far. Without such inhibition, activation would spread throughout the multiply interconnected network of neurons which make up my brain, until all my brain's nodes were activated.

My **coffee** representation isn't stored in any particular place in my brain, rather it's implemented as the activation of a set of nodes which are a small subset of the billions of nodes in my brain. These activated nodes are scattered all over my brain, but are physically connected together. So if we consider the overall network of nodes which makes up my brain, when my **coffee** representation is activated some nodes are activated but many more are not activated. If anything can properly be called my **coffee** representation, it is this resultant PATTERN OF ACTIVATION over all the nodes in my entire brain.

I call this representation a "PATTERN OF ACTIVATION" (written with *capital letters*) to distinguish it from the sort of pattern of activation (written with *lower-case letters*) which might appear across a small set of nodes, like the nodes which receive the pattern of frequencies which encode a smell.

---

2 This sort of inhibitory connection also appears in the connections between nodes in the human brain. See Patricia S. Churchland (1986) : especially pp51-55.

Each encoded smell could be said to be encoded as a pattern of activation across these six nodes. Thus bits of information which are constitutents of a representation can sometimes be a pattern of activation over a set of nodes rather than the activation of one node. (Implementing representations' constituents as patterns of activation like this has several advantages over implementing each constituent as the activation of a single node; I'll discuss these advantages soon. I'll stick with the idea that each encoded piece of information is encoded as the activation of a single node for a while longer.) Thus a PATTERN OF ACTIVATION is a whole representation, which can have patterns of activation as constituents.

By implementing a representation as a PATTERN OF ACTIVATION over a set of interconnected nodes, the process of activating the whole representation because one of its constituents is activated becomes a purely mechanical process. The activation of any one of these nodes causes the activation to spread through the connections between the nodes so that all these interconnected nodes become activated. By having all these nodes activated, I become aware of all the other bits of information I associate with coffee; I'm aware of each bit of information encoded by the activation of one of these nodes.

Note that the representation is not the *network* of nodes itself, but the PATTERN OF ACTIVATION over the network of nodes. Just as a voice signal on a telephone line doesn't exist in a certain place as a physical entity but merely as a series of electrical impulses on a telephone line, in a similar way a representation doesn't exist as a physical localised object, but rather as a temporarily activated PATTERN OF ACTIVATION over a set of nodes. A representation doesn't exist all the time, but merely has the potential to exist. It only exists when the relevant nodes are activated. Likewise, each part of a representation is the *activation* of a certain node, and not the *node* itself. So the node which is activated whenever the smell recogniser network recognises the pattern (5,4,2,5,3,6), does not itself encode the smell of coffee, and is not itself a part of my **coffee** representation. Rather it's the *activation* of this node which encodes that sensation and is a constituent of my **coffee** representation.

This way of construing a representation gives a physical, mechanical picture of how my recognising the smell of coffee can activate my entire **coffee** representation. The present state of my sense organs' outputs are encoded as patterns of spiking frequencies on a certain set of parallel fibres. This pattern of spiking frequencies causes a pattern of activation on the input nodes of the pattern recogniser network. Because of the weights of the various connections between the nodes of this pattern recogniser network, this input pattern causes the network to activate one of its output nodes.

This output node's activation sends an activation signal to all the nodes connected to it, which causes the activation of all the nodes connected to it and so on. The whole set of activated nodes which results constitutes my **coffee** representation. By this means my **coffee** representation can be activated by the activation of any of the nodes whose activations are constituents of this representation. The activation of the nodes which encode the information contained in hearing the word "cuppa", smelling coffee or seeing the coffee machine are all constituents of my **coffee** representation, and so their activation will initiate the activation of this representation.

Now we have seen how the "list" of encoded sensations is actually implemented, it becomes clear that job (3) (which I talked about in section 2.3) – the job of activating my **coffee** representation– is not exactly a separate job to be done after job (2) is done –the job of recognising that the smell information picked up encodes the smell of coffee. Nor is job (3) separate and precedent. I certainly don't recognise the smell as the smell of coffee, and *then* initiate a process which finds my **coffee** representation and activates it, as I imagined when I earlier set out (1), (2) and (3) as a sequence of tasks. Activating a representation is accomplished by activating a node whose activation is one of the representation's constituents. And activating this node is exactly what job (2) consists of. So recognising that the encoded information is encoded as a constituent of my **coffee** representation *is* activating the representation. (Jobs 2) and (3) are not sequential processes performed by separate components: a recogniser module and an activator module. Thinking of these jobs in this way is a symptom of the "computational" metaphor of the brain. From a connectionist perspective, recognising a state space point and activating a representation is an *event*, not a sequence of processes. Indeed, both of these are actually one and the same event: the activation of a certain node.

Likewise job (1), for which I thought needed a separate mechanism, can also be explained in terms of the outputs of a recogniser network and how it activates a representation. I thought I needed job (1) to explain experiences like seeing a person's face, and knowing that you've met them before but not knowing where you met them, or who that person is. It could be that the connection strengths in the face recogniser network[3] are such that the image caused by this person's face does cause the activation of a certain output node, but the activation is only weak because the face is not a very familiar one. The weak activation of this output node may not be sufficient

3 There is evidence that we have a face recogniser network, in the fact that this specific ability can be lost due to damage to a certain area of the brain.

to activate any other nodes. So the face recogniser's output node being activated indicates the recognition of the image as a familiar one, but the output isn't strong enough to activate the representation by activating any other nodes. This representation may be activated, however, as the result of more information which helps remind you of the situation you met them in (such as someone reminding you of her name, or of the person who introduced you). This extra information may activate other constituents of the representation, whose activation adds to the weak activation from the face recogniser network's output node, activating the representation.

3.2 How contextual activity can affect the activation of representations.

Remember that earlier I said that some constituents of a representation could themselves be representations. I've down-played this aspect so far, to keep the initial diagram simple. Figure 3.1 showed connections from my **coffee** representation to other representations. I'll expand the diagram now, to make it a little more realistic. This expanded diagram shows how representations can be constituents of other representations, and helps illustrate how inhibitory connections and the weights of connections can help limit the spread of activation. It also illustrates how contextual activity in one part of a large network can affect the PATTERN OF ACTIVATION which results from a pattern recogniser recognising some piece of perceptual information.

In figure 3.2, node (d) isn't connected simply to nodes at the outputs of pattern recognisers. Some of the nodes it's connected to are themselves connected up in the same way as itself, as the "hub" of a network of nodes. Each of these "hub" nodes could be said to be at the centre of representations which are related to the central one. The node at the centre of Figure 3.1 wasn't connected simply to the output nodes of pattern recognisers either. It was also connected to representations. The difference between these was down-played at the time, by illustrating these other representations as single nodes. They would in fact be smaller networks like those in Figure 3.2.

Note that some of the lines in figure 3.2 are drawn dotted. These indicate inhibitory connections which work to prevent nodes from being activated. If a node at one end of such a connection is activated, then this will send a signal to the node at the other end which will lower its activation level. This is necessary because some of the nodes whose activation encode constituents of a representation are also connected to nodes whose activations are not constituents of that representation. These other nodes

could be activated inappropriately, unless inhibitory signals are sent to them to counteract the activation signals they are sent.

Note also the connections drawn with arrows on the ends. These indicate connections to nodes which are out of the picture. Nodes (a) and (b), for instance, can be activated by activity outside the system pictured, as well as by the activation of nodes they are pictured as being connected to. And node (h) can receive inhibition signals from outside the diagram. This outside activity can profoundly affect the PATTERN OF ACTIVATION which results.

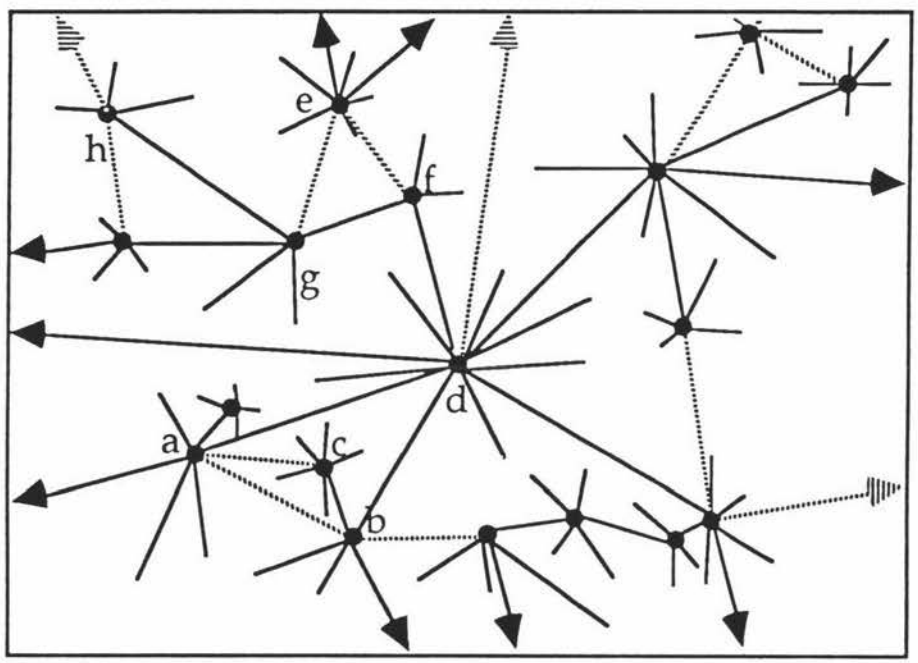


Figure 3.2 A more realistic network of representations[4]

Here are a couple of examples which illustrate the way the activation could spread and be inhibited in the diagram in figure 3.2. Suppose node (d) is activated by an excitation signal received from outside the diagram. This would send excitation signals to nodes (a) and (b), among others. But there is an inhibitory connection between (a) and (b), which means that if one is highly activated then this will send inhibition signals to the other, so that it is less likely to be activated. At the moment it's difficult to say which would be activated; this would depend on the strength of the activation each receives from (d). But if (b) also receives a strong activation signal from outside the diagram, then this could increase its activation level. Thus (c) would receive a strong activation signal from (b), and (a) would receive a

4 Here I've incorporated an idea which is perhaps not that realistic. In a real brain the connections between neurons are one-way. The connections come from dendrites, and go out through the neuron's axon. However, the connections in figure 3.2 are illustrated as mutually activating, two way connections. In a real brain two nodes may be able to be connected so that they are mutually activating by having two separate connections: from the first to the second, and another from the second to the first. I'm assuming, for the sake of the illustration that all the connections pictured are two way ones.

very strong inhibition signal from (b) and (c). If this inhibition which (a) receives from (b) and (c) outweighs the activation it receives from node (d), we would end up with (a) unactivated and (b) and (c) activated. But with different activity outside the diagram, the activation of node (d) might have resulted in a different final pattern of activation: (a) might have been activated and (b) not.

Similarly, if both node (d) and node (e) receive activation signals from outside the diagram, then node (f) could be activated or not, depending on the relative strengths of the inhibition which it would receive from (e) and the activation it would receive from node (d). The overall pattern of activated nodes which eventuates can be affected by the way inhibitory and excitatory connections interact in a network and also by subtle differences in activity further away.

When we look at real life, this should be expected. Factors in the context of a sensation may make some constituents of a representation more salient than others. The context of a representation's activation affects which of its constituents are brought to my attention. While I'm sitting in the dining-room at home, if my flatmate asks me if I'd like a cup of coffee, I don't usually start thinking about where to find coffee on the shelves at Pak'n'Save. This happens because certain nodes are activated and certain others are not, because of excitation and inhibition from the parts of the overall network which encode where I am and what I'm doing. If my flatmate asks me the same question while we're in the supermarket, this node might very well be activated.

When one node's activation is a constituent of two or more (mutually exclusive) representations, contextual factors can influence which representation is activated and which is not. The activation of this common constituent results in two representations competing to be activated, and the context may sort out which pattern of activation eventually results. For example, the sensation caused by hearing the spoken word "cuppa", when recognised, causes the activation of a particular node. The activation of this node is a constituent both of my **coffee** representation and also my **tea** representation. Since it is a constituent of both of these representations, recognising this sensation could cause either of these representations to be activated. And since I can't have (and usually don't want) both a cup of coffee and a cup of tea, my **coffee** and **tea** representations would probably have inhibitory connections between them, so that if one is activated, then the other's activation is inhibited. Context can tip the balance, to help activate one of these and not the other, and this can happen in two ways: by patterns of activation pre-existing before the activation of the common

constituent, or by sensations encoded concurrently with the common constituent.

If certain nodes are already partially activated before being activated, this pre-existing pattern of activation may add to (or inhibit) the activation of certain nodes. For example, suppose I haven't had any coffee all day, and my system is suffering caffeine-withdrawal. This fact could be encoded by the activation of a node somehow connected with my **coffee** representation, but this node may not be activated strongly enough to activate my **coffee** representation by itself. But suppose now my flatmate asks if I want a cuppa, which causes a node connected to both my **tea** and **coffee** representations to be activated. The first activated node, which encodes the information that I need caffeine, tips the balance and my **coffee** representation becomes activated more strongly than my **tea** representation (which also helps inhibit my **tea** representation's activation). In this way nodes already activated in one part of a network can affect what representation an activated node activates.

Other sensations experienced at the same time may also cause the activation of one representation rather than another. Say for example, I recognise the image of the coffee plunger in my flatmate's hand at the same time as this node whose activation encodes the spoken word "cuppa" is activated. This image sensation also activates a constituent of my **coffee** representation. The activation of this other constituent of **coffee** can tip the balance between the activation of my **coffee** and **tea** representations.

The idea that some nodes connected together need not be activated by the activation of one of their number further revolutionises our idea of what it is to be a representation. It can be seen that the activation of a particular node can cause many different patterns of activation of nodes, both because one node can have connections to *many* other nodes, as in figure 3.2, and thus could possible activate many representations, and also because the overall pattern of activation which results in any given situation can be affected by contextual activity in other areas of the overall network.

This gives us a much more flexible picture of a representation. We are accustomed to thinking of a representation as a static "list" of encoded sensations which are *all* activated when the representation is activated. With network representations a representation can be implemented as a PATTERN OF ACTIVATION over a certain set of nodes, but each PATTERN OF ACTIVATION can be subtly different, by having a different sub-set of those nodes activated. This gives potential for the same set of nodes to implement a very large number of distinct representations. The variations in the overall pattern of activated nodes which can result are virtually limitless: some nodes can be activated and some not activated. So as well as no longer

being a physically localisable chunk of matter, a representation is no longer a fixed, static, entity but a flexible, dynamic one.

For example, what I would have called my **spaniel** representation could be implemented as a PATTERN OF ACTIVATION over a set of interconnected nodes, but for each instance of its activation only a subset of those nodes will be activated, and some to a higher degree than others. Thus this set of nodes would be activated in different configurations to represent: a black spaniel, a white spaniel, a black and white spaniel, a noisy spaniel, a mute spaniel, a blind spaniel, a spaniel which is very old, a spaniel puppy, a concrete statue of a spaniel, a toy spaniel, and a vast number of other variations, combinations and permutations on this theme. (We could perhaps say that there is a "family resemblance" between all these Patterns of Activation.) Thus using network representations offers a system in which a relatively small set of hardware units can implement multitudinous distinct representations. This is a staggering improvement, both in representational power and in realism, over the traditional "list" of encoded sensations which I started with!

3.3 *Advantages of connectionist representations*

Another impressive characteristic of a connectionist-based representational system is the fact that many of the weaknesses and strengths of human cognitive faculties fall out as side-effects of network-style representations. The fact that these systems share similar proclivities with the human mind is very good psychological evidence that we're on the right biological track here. In this section I'll illustrate some of these faculties which networks and humans share. Figure 3.3 is a diagram which illustrates how representations of five different people might be implemented in a network-based system.[5]

5 The idea for this diagram is taken from Martindale (1991) p16. We can trace the ancestry of Martindale's diagram to the "Jets and Sharks" example, in McClelland, Rumelhart et al. (1986) : pp27-31.

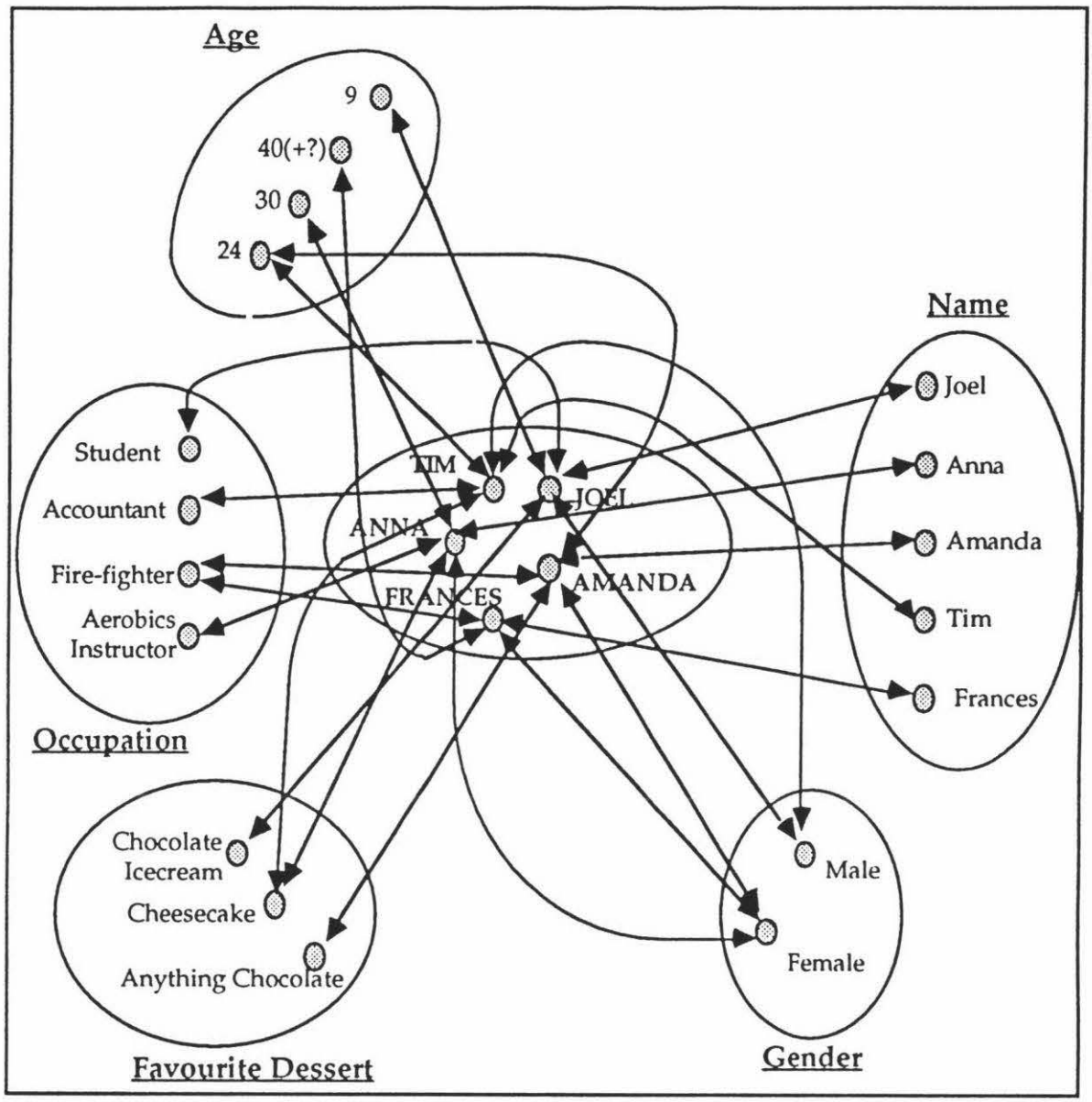


Figure 3.3 The interaction between connectionist representations

The arrows in this diagram again indicate mutually activating connections between nodes. As well as excitatory connections this system also has inhibitory connections between the members of each group; but drawing dotted lines from each member of a group to all the others would have been messy and complicated, so I've indicated inhibitory interconnections by drawing an oval around all the mutually inhibitory nodes. If one node within an oval is activated, then this sends inhibitory signals to all the other nodes in that oval which reduces their activation levels. Each of the nodes in this diagram can be activated either indirectly (through the other nodes pictured) or directly from outside connections to perceptual systems. Through these outside connections, the network could be questioned about Amanda by activating the node whose activation encodes that name. If one of the nodes is directly activated in this way, this

sends out a pattern of activation and inhibition which eventually settles down to a stable pattern with some nodes activated and some not.

By following the arrows, and seeing how the nodes are connected, it's possible to see all the constituents of, say the **Anna** representation. If the "Anna" node is directly activated, then the nodes for her name, gender, age, occupation, and favourite dessert would end up being activated (indirectly); all others would be unactivated.

Such a system of organising representations can be shown to have very human-like qualities.[6] As I've been explaining, this system displays what is called *content-addressable memory*. Each representation is activated by the activation of any one of its constituents. We can find out about Anna, by activating the node whose activation encodes the name "Anna", or by activating the node whose activation encodes the occupation "aerobics instructor", or by activating any of the other constituents of the **Anna** representation. The rest of the nodes which are constituents of this representation are automatically activated whenever one is activated.

A system which employs network representations like this also is capable of making *generalisations*, an ability humans definitely have. We make generalisations by assuming that if two things are similar in some respects, then it's likely that they will be similar in other respects too. For example, note the similarity between the **Frances** and **Amanda** representations in figure 3.3. Each of these representations has the activation of the nodes for "female", and for "firefighter" as constituents. Martindale says that because of this similarity, this network would activate the node for "Anything Chocolate" when the **Frances** representation is activated, even though there is no connection here. This is because activating the **Frances** representation would activate the "female" and "firefighter" nodes which are also constituents of the network's **Amanda** representation. This would cause the *partial* activation of the **Amanda** representation, which cause the "Anything Chocolate" node to be partially activated. The partial activation of this node would inhibit the activation of all the other nodes in this group. Thus the network would "assume" that Frances' favourite dessert is "Anything Chocolate".[7]

This sort of network is good at another (related) sort of generalisation: extracting the characteristics which two representations have in common. This can be done through the activation of a node which is too general to activate any one representation. For instance the node whose activation encodes the occupation "fire fighter" is a constituent of two different

---

6 These ideas are also from Martindale (1991) p16.

7 Martindale (1991) : p16.

representations. So asking the network about firefighters by directly activating that node, is activating a constituent common to both of these representations. This would spread a pattern of activation and inhibition throughout the network, which would result in the nodes common to both representations being activated strongly. In groups where each representation has different constituents, the nodes activated would also inhibit each other, leaving them only partially activated. That is, the nodes for "female" and "firefighter" would be activated, while all the other constituents of the **Amanda** and **Frances** representations would only be partially activated (except that "anything chocolate" would be activated strongly, none of its competitors would be activated).

This network is also able to use *default assignments* based on its generalisations. If this network was subsequently told about Grant, who is also a firefighter, but the network was not told about Grant's gender, the network would "assume" that Grant would be female, (just as it assumed that Frances' favourite dessert would be Anything Chocolate). Because all the firefighters already represented by the network are female, the "female" node would be partially activated, and the network has no information which would cause the inhibition of this node. This sort of prejudiced expectation is also typical of human beings. The advantageous aspect of this feature is that this prejudiced expectation is quick but also correctable. It can be fixed by representing information about a wide range of subjects, so that the network's representations reflect the diversity found in the true population rather than some quirky sub-sector of it.

The performance of such a network can also be seen to exhibit *graceful degradation* in its performance under damage, and in its performance in coping with noisy or erroneous information. If we ask the network of figure 3.3 about a 30-year-old *male* aerobics instructor (by directly activating these three nodes), the "Anna" node, will end up most strongly activated even though some of the information is misleading. Networks like these aren't paralysed by incomplete, noisy or erroneous information, but as the information gets worse, so does the network's performance. Humans also exhibit this feature. When faced with misleading or incomplete information we can often arrive at an answer which best fits the information provided. Most people would arrive at the same response to the following description, even though part of it is misleading: it is an actor, it is very intelligent, it is a past president of the USA[8].

As well as being a feature of network representations, graceful degradation also falls out as a feature of the pattern recogniser networks I

8 This example is from Hinton, McClelland et al. (1986) : Vol 1, p79.

discussed earlier. This happens in two ways. Firstly, if a section of a recogniser network is damaged, the overall network will still continue to function, although less than optimally. And as more of the network is damaged, its performance degrades proportionally. This degradation of performance in recogniser networks has also been found to occur in humans. Very localised brain lesions have been found to cause small errors in humans' perceptions. Martindale reports on a person who had very localised damage to the part of his brain which is responsible for recognising spoken words. This person could perform well in recognising most words, but was unable to distinguish a certain range of similar ones. For example, he reported hearing "pat" when "bat" was spoken.[9] The way the performance of this part of his brain degraded due to the brain lesion is very similar to the way a (fully trained up) recogniser network would perform if a small section of it was damaged.

The second way in which pattern recogniser networks' performance degrades gracefully is in their performance in working with noisy or incomplete information. Pattern recogniser networks will attempt to match such information to the known input which best matches the present input. Depending on its architecture, it will either produce the output pattern or activate the output node for a known input pattern which most closely resembles the present noisy or incomplete input pattern.

We had an example of this in Chapter One. Recall the example of my seeing (without my glasses on) my grey jersey on the chair that my cat Madison likes to sleep on. This sensory information was processed by a visual pattern recogniser network which produced the output which most closely matched the stimulus it was getting: it activated a node which encodes the image of Madison sitting on the chair. Since I often see Madison asleep on that chair it's a node which is activated often. The activation of this node caused the activation of my **Madison** representation. In this way my perceptual systems' picking up noisy or incomplete information can cause the activation of representations which wouldn't be activated if better quality or more complete perceptual information were available.

The features of networks performance I've just been talking about –that they make generalisations, exhibit content-addressed memory, use default assignments and have performances which degrade gracefully– are sometimes called "emergent" properties of networks. No-one had to build these features in; they just fall out as side-effects of this way of implementing representations. And because all of these are features of human cognition and memory, we have good reason to expect that human cognitive faculties,

---

9 Martindale (1991): p54.

and in particular our representational powers, can be explained in terms of networks organised along more or less these lines too. Network representations provide realistic models of human representations .

3.4 *Distributed representations.*

All the network representations we've looked at so far have encoded sensations by activating a single node. As I mentioned earlier, this one-to-one correspondence between entities to encode and units whose activation does the encoding is very much an oversimplification. *Local* representations like these have been useful as an introduction to connectionist representations because they are easier to understand because the structure of the relations between pieces of information represented is directly mirrored in the structure of the network. But Rumelhart and McClelland have shown that *distributed* representations, which encode each piece of information as a pattern of activation over *several* nodes rather than as the activation of a *single* node, have important advantages over local representations. They show that local representations require an excessive number of hardware units, have difficulty in accommodating new concepts, and although local representations demonstrate some important features of human cognition, there are nonetheless certain other features of human cognition which they lack, and which distributed representations provide. The following discussion explains some of the advantages of using distributed representations.

The extra efficiency of distributed representations is one of their most attractive features. Encoding a constituent of a representation as a pattern of activations over a number of nodes, rather than dedicating a separate node to each constituent, can drastically reduce the number of nodes required. A comparison of the different way numbers could be encoded in systems which use local representations and in systems which use distributed representations will demonstrate the dramatic efficiency of the latter. The "local" way to encode all the numbers from 0 to 255 would be to have a separate node to encode each number: encoding the number 99 would be accomplished by the activation of one particular node (that is, node 99), and the number 3 would be encoded by the activation of a different node. Thus we would need 255 separate nodes to encode all these numbers. A more efficient way would be to employ the same technique used in digital computers: binary number encoding. Such a system would use eight nodes which encode by their activation the numbers 128, 64, 32, 16, 8, 4, 2 and 1; by activating a certain subset of these nodes, we can encode every number from

0 to 255 as the sum of a subset of these numbers. Each number is encoded as a pattern of activation across all eight nodes. So the number 255 is encoded by 11111111, and 0 is encoded by 00000000. To encode the number 99 , which is the total of 64+32+2+1, we activate the nodes which stand for these numbers, and deactivate the others. Thus 99 is encoded as the pattern 01100011 (the addend 1 is encoded by the right-most node, and 128 by the left-most one). So in a distributed system, rather than needing 256 nodes as we would in a system which encodes numbers locally (one node for each number), we only need eight nodes to encode 256 different numbers; an impressive gain in efficiency.

This gain in efficiency uses nodes with only two levels of activation. With more levels of activation we can increase efficiency even further. In the last chapter I showed how using only six fibres, each fibre capable of carrying ten different spiking frequencies, we can encode one million different smells; each as a (distributed) pattern of frequencies over this set of fibres. (In a similar way, using six nodes with ten levels of activation each we can encode all the numbers from 000000 to 999 999: one million different numbers implemented as different patterns over six nodes.)

Establishing a link between a pair of patterns of activation on different sets of nodes, so that the presence of one pattern of activation causes the activation of the other pattern, is easier that it might seem. We've already discussed one apparatus which can do this: pattern recogniser networks. These networks, recall, take a pattern of activation on their input units, and through having the appropriate weights on the connections between their nodes generate a pattern of activation on their output nodes. This output pattern of activation is produced only when that particular input pattern is produced. Different input patterns cause different output patterns. One especially useful feature of these networks is that the input and output patterns can be over different sized groups of nodes; for instance, a pattern of activation over ten nodes can be associated with a pattern of activation over six nodes, and vice versa.

Using these networks, which McClelland, Rumelhart and Hinton[10] call *pattern associators*, the appearance of a certain pattern on one set of nodes can cause the activation of another pattern on another set of nodes. Because of this factor, all the networks which use local representations that I've discussed so far can be implemented using distributed representations; the jobs done by the networks illustrated in figures 3.1, 3.2, and 3.3 can all be accomplished by networks which use distributed representations. Using pattern associators to associate a pattern of activation over one group of

10 Hinton, McClelland et al. (1986) p33.

nodes with a pattern of activation over each of several other groups of nodes, we can implement all the same networks using patterns of activation to encode each piece of information.

Figure 3.4 is an illustration of a small section of the way such a distributed representational system could be implemented. There is a set of nodes over which we can implement patterns of activation to encode smell information, another to encode taste information and another to encode images. By having connections with the appropriate weights, each pattern associator relates a pattern of activation over one set of nodes with a corresponding pattern of activation over another set of nodes. So if the six nodes whose patterns of activation encode smell information were activated in the pattern (5,4,2,5,3,6), then the pattern associator could cause a corresponding pattern over the central set of units (which might be said to encode the information that coffee is present). The presence of this particular pattern of activation could cause, via the connection weights in the respective pattern associator networks, the patterns of activation which encode information about the taste of coffee and the image of a cup of coffee to become activated.

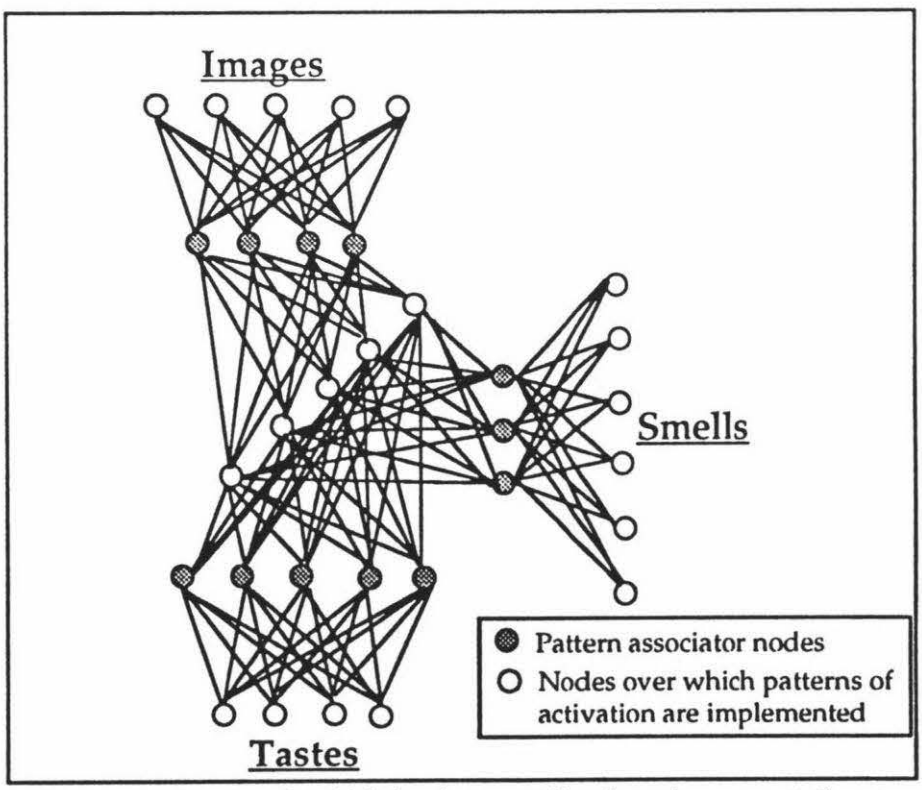


Figure 3.4 A network which implements distributed representations.

The remarkable thing is that if each node is capable of ten levels of activation, the network pictured in Figure 3.4 could be used to implement thousands of images, smells and tastes, and associate the corresponding ones

together. Thus representations of **coffee, garlic, steak, basil,** and hundreds, perhaps thousands, of other associations of smells, tastes and images could all be implemented on the above sets of nodes. (This might require a few extra pattern associator nodes though.) Through such a system, seeing a barbecued steak would cause the patterns of activation which encode the smell and taste of a barbecued steak to be activated, and smelling basil would cause the patterns which encode the taste and image of basil to be activated.

The only disadvantage in using distributed representations is that they are less easy to visualise, because the relationships between information represented isn't visibly reflected in the structure of the network. With a system which uses distributed representations, we can't assign "labels" to patterns of activation to illustrate on a diagram what it is that the pattern of activation encodes. Also the relationships between different patterns of activation are hidden in the weights of the connections between nodes, they're not illustrated by easy-to-follow arrows as we had with diagrams of local representational systems. Compare figure 3.4 with figure 3.5, which illustrates how the same associations between tastes, smells and images could be implemented in a system which uses local representations. Here we can label all the encoded sensations, and all the relationships between encoded sensations can be clearly seen. By following arrows, it can be seen that the activation of the node whose activation encodes the taste of basil will cause the activation of the nodes whose activations encode the image and smell of basil too.

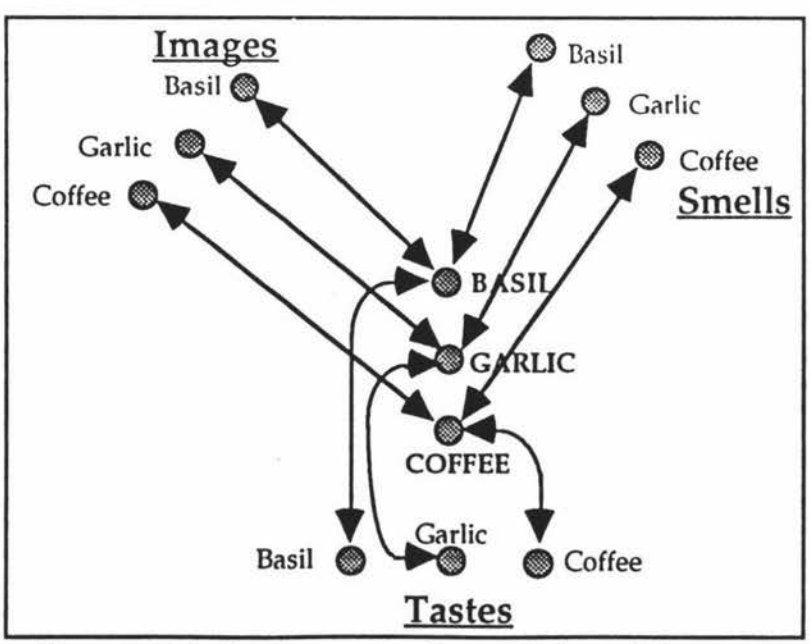


Figure 3.5 The local representation equivalent of Figure 3.4.

Of course, to be able to encode the thousands of representations which the distributed network of figure 3.4 could implement, this network would

need thousands of nodes in each section. But still, in figure 3.4 the structure of the information represented is easily seen because it is mirrored directly in the structure of the network itself. With figure 3.3, however, the structure of the network gives little clue to the structure of the information represented. But this disadvantage is merely inconvenient from the point of view of *visualising* how the network does its job. As long as it has the appropriate weights on all its connections, the distributed representation network of figure 3.4 can still *exhibit* all the associations between encoded information which a network which uses local representations like the one in figure 3.5 could exhibit, and many, many more. The network of figure 3.4 can still do the job, and that is what matters most.

A system which employs distributed representations has another important advantage over one which employs local representations: it is remarkably easier to introduce new entities to represent. The ability to learn new concepts to represent is essential to any representational system which is intended to be useful. It is even more essential that a system be able to represent new entities which weren't even foreseen at the time the system was designed and built. With local representations this isn't that easy. The network in figure 3.5 encodes each smell, taste, image and so on by the activation of a separate single node, so to represent a new entity this system needs to acquire a new node in each of the relevant groups, and to wire up these newly acquired nodes in the right way. In order to have the ability to incorporate new representations, either the system's designers need to anticipate every representation the network will need to represent, or the system will have to have the capacity to grow new nodes on demand, and to grow connections between specific nodes, also on demand. Each of these is difficult to implement.

But with the distributed system pictured in figure 3.4, representing new entities is much easier. No new nodes are required, and no new connections between nodes need to be forged. To implement a new representation, the network needs to learn a new pattern of activation over one set of nodes and it needs to learn to associate this pattern with a new pattern on the other sets of nodes. This requires no new nodes, all that needs to be done is to adjust the weights on the connections in the pattern associator networks so that a new stable input/output pattern association is learned by each pattern associator. If this adjustment of connection weights is done slowly (by having a low learning rate so that only very small alterations are made in the connection strengths), then these new pattern associations can be learned without disturbing the associations between patterns of activation which the system has already learned.

Another welcome advantage of distributed representations is that a system which employs distributed representations can make even more effective generalisations than a system which uses local representations. With local representations, the network notices the similarity between representations only if the representations each share the activation of a particular node as a constituent. In figure 3.3, the similarity between the network's **Amanda** and **Frances** representations was that they each shared the activation of the "firefighter" and "female" nodes; that is, there were constituents which were common to each representation. But this sort of network is entirely insensitive to the ways in which representations could have constituents which are very similar, but which are not identical. For instance, this network wouldn't be able to notice the similarity between the taste sensations associated with the two desserts "chocolate icecream" and "anything chocolate".

Distributed representations can take advantage of *similarities* between patterns implemented over the same set of nodes. They aren't limited in their generalisations to representations which have *identical* constituents. For example, in the network of figure 3.4 the pattern of activation which encodes the outputs of the taste sensors when tasting chocolate icecream will be very similar (though not identical) to the pattern over these same nodes which encodes the sensation caused by tasting chocolate cheesecake. These patterns will be alike in some respects and different in others. Let's say for the sake of this example that the first two nodes are excited to the same level of activation in each pattern, and the third and fourth are different; thus the generalised characteristic of these two patterns could be expressed by (3,8,?,?). The network can be sensitive to the similarity between these two patterns, and relate them to patterns of activation on the central set of nodes which will be similar in some respects also. This similarity in the patterns on the central set of nodes would probably be reinforced by the similarity in the patterns of activation which encode the sensations caused by smelling chocolate icecream and chocolate cheesecake.

The generalisation that things which cause similar tastes often also smell alike falls out as a side effect of this way of implementing representations, precisely because they are sensitive to similarities in different patterns of activation over the same set of nodes. Suppose the system of figure 3.4 is exposed to a new taste sensation, again encoded by the pattern of activation (3,8,?,?) caused by tasting chocolate cake; but no smell sensation accompanies this taste. This pattern of activation over the taste nodes would cause a pattern on the central set of nodes similar to those patterns caused by the other two encoded taste sensations. And because these central patterns are similar, this could then cause the smell nodes to be

activated in a pattern similar in some respects to those patterns of activation which encode the smell sensations caused by chocolate icecream and chocolate cheesecake. So the network's connections could implement the generalisation that because the pattern which encodes this new taste sensation is similar to the other two (in the same way they're similar to each other), then it will probably also cause a smell sensation which is encoded by a pattern which incorporates the common characteristics of these other two patterns. The network "assumes" that because chocolate cake tastes like chocolate icecream and chocolate cheesecake, then it will smell like chocolate icecream and chocolate cheesecake too.

But distributed representational systems like these don't necessarily attune to similarities; similar patterns over one set of nodes can be correlated with quite dissimilar patterns in another set of nodes.[11] The patterns of activation which encode the images associated with chocolate cheesecake and chocolate icecream will be quite dissimilar.[12] So the pattern associator which associates patterns of activation on the central set of nodes with patterns of activation on the set of nodes which encodes images will need to be capable of *ignoring* the similarity between the patterns of activation on the central nodes which encode **chocolate cheesecake** and **chocolate icecream**, and be able to concentrate instead on the aspects of the two patterns which are different. And this too is quite possible. Pattern associators *can* be trained to ignore similarities, and associate similar patterns of activation on one set of nodes with dissimilar patterns of activation on other sets of nodes.

So apart from being more difficult to see *how* a system which uses distributed representations does its job, in that it's less easy to see the structure of the information represented from the structure of the network itself, such a system has several important advantages over one which uses local representations. It is easier to learn new entities to represent, it is more efficient in the use of nodes, and it can make better generalisations about similar entities. And as well as having these advantages over local representations, a system which employs distributed representations still has all the advantages which made systems which employ local representations look so wonderful earlier. Such systems still exhibit content-addressable memory, and their performance still degrades gracefully if either the

---

11 This is so only if the network uses a fairly sophisticated learning rule. Hinton, McClelland et al. (1986) say that the Hebbian learning rule can only learn associations between an entire ensemble of patterns of all the patterns are uncorrelated. To create a system which can handle correlated and uncorrelated patterns a more sophisticated learning rule needs to be used, such as the delta rule I discussed in section 2.7

12 Illustrating the images as being encoded by a small set of nodes like this is a little misleading. In fact, the patterns of activation which encode different images are quite likely patterns of activation over different sets of nodes. In reality, my image of a garlic plant may be encoded as a pattern of activation over one set of nodes, where my image of an icecream could well be implemented over an entirely different set of nodes.

network is damaged or the information it is given is misleading, noisy or erroneous. They are also capable of making default assignments based on their generalisations.

3.5 Some minor difficulties with distributed representations

Implementing huge interconnected networks of representations within representations within representations, like that illustrated in figure 3.2, requires an important modification to the way things are done, as I've explained them so far. For a start one pattern of activation can't be a constituent of two representations if the two representations use the same central set of nodes. This is because a pattern associator can't associate one pattern of activation on a set of nodes with two different patterns on another set of nodes. The pattern associator can only perform a one-to-one mapping of patterns. So if a sensation is a constituent of two mutually exclusive representations, then these two mutually exclusive representations can't be implemented using the same central set of nodes. For example, using the distributed network of figure 3.4, the same encoded image (like that caused by a steaming mug of brown liquid) can't be a constituent of both my **coffee** and **tea** representations, if each of these representations involves a different pattern on the central set of nodes. Implementing these representations with the same encoded image as a constituent of each would require a different central set of nodes for each different pattern.

As well, a pattern of activation can't be activated to a lesser or stronger degree in the same way that a node can be strongly or weakly activated. This is an important difficulty when we consider how the interaction of excitatory and inhibitory signals can affect a node's activation level. Doing the same sort of thing with patterns of activation will be a lot more difficult. The pattern of activation (5,4,2,5,3,6) can't be weakly or strongly activated; if it's weaker or stronger then it's simply a *different* pattern.

These difficulties are formidable, but not devastating however. Perhaps we don't need to worry overly much about how the interplay of activation and inhibition could be implemented in a system which uses distributed representations, because there is no particular reason why a connectionist representation has to *exclusively* use distributed representations. A system could well use distributed representations to implement some representations, but use local representations to implement others.[13] Alternatively perhaps we could have a system which

13 There is some evidence to suggest that neurologically this is just what does happen: some information is encoded locally. There are certain nodes in what is called the TE area of the brain whose activation is

uses the activation of single nodes to implement the interactions between representations, and patterns of activation to encode the constituents of representations. We only *need* distributed representations where there is a need for efficiency, or where the similarity relations between bits of information encoded needs to be reflected in similarity relations between the patterns of activation which encode that information. As a matter of neurobiology, maybe we only use distributed representations where we need to use them.

I do not know enough detail about the interplay of patterns of activation and their implementation to do anything more than speculate here, however, so I'll stop now while I'm hopefully still making some sort of sense.

In this chapter I've illustrated the advantages of making the shift from the picture of human representations offered by the computational picture of the brain to that offered by the connectionist way of viewing the brain. This shift in perspective transforms our account of what representations are, what a part of a representation is, how representations are organised and how they're activated by sensations into a much more realistic picture. My **spaniel** representation, for instance, is no longer a static "list" of encoded sensations stored in a particular place, which is either activated in its entirety or not at all. Our picture of this representation changes dramatically when we view representations as implemented by networks of nodes, encoding information as patterns of activation over sets of nodes. My **spaniel** representation becomes a flexible entity capable of implementing representations of a vast array of similar but different entities, by varying which parts of this overall set of activated nodes are activated on each occasion and which are not, it can represent all the variations on the theme of spaniels I can conceive of. And the characteristics which emerge as side-effects of this system of organising representations, seem to be exhibited by human representational systems. As Hinton, McClelland and Rumelhart remark:

> **"the best psychological evidence for distributed representations is the degree to which their strengths and weaknesses match those if the human mind."[14]**

---

caused by picking up information about certain shapes. The activation of this single node encodes this information. See Ornstein and Thompson (1985).

14 Hinton, McClelland et al. (1986) : p78.

Because of all this we have good reason to believe that the representations implemented in the human brain are organised along very similar lines to those discussed in this chapter.

This makes it possible to demystify the representation relation. In the rest of this thesis, I hope to show how connectionist representations explain how a bit of my brain can come to stand in the representation relationship with some particular part of my environment. In the next chapter I begin this task by discussing the constituents of connectionist representations like these, and the sort of information they must encode if we're to be able to explain human representation and misrepresentation.

CHAPTER FOUR

WHAT ARE THE CONSTITUENTS OF A REPRESENTATION?

4.1 So where is all this going?

I'll start this chapter with a brief look at where I've got to so far and then have a not-quite-so-brief look at where I'm trying to get to, and how I plan on getting there. Over the last two chapters I've given the foundations of a connectionist picture of what a mental representation might be like; a picture which has several advantages over that offered by the traditional views of representation. The picture we have so far is that a representation is a PATTERN OF ACTIVATION over a large network of nodes, made up of a physically interconnected set of smaller patterns of activation each of which encode information about some aspect of the environment. In Chapter Five I'll introduce the semantic relations which must hold between a representation and some piece of the world, and show where these semantic relations come from. Following that, in Chapter Six, I'll approach the problem of representation and misrepresentation from the perspective I'm developing, explaining how a representation can represent some things, and can misrepresent other things. In this Chapter I'll make some refinements to my connectionist story so that an explanation of representation and misrepresentation will be possible.

At the end of Chapter One I listed four assumptions which lie unquestioned by many traditional approaches to the problem of representation and misrepresentation, and claimed that these all need to be re-thought if we're going to provide an account of representation which allows for misrepresentation. These assumptions were:

(a) Representations are activated by physical objects.
(b) The sense organs' job is to convert the properties of objects into properties of representations of those objects.
(c) In explaining how we represent our environments, how those representations are used is relatively unimportant.

(d) Physical objects (as opposed to abstract ones) are the only kind of objects which can figure in an account of representation.

My rejection of (a) has already been explained, and to some extent justified. We're not going to be able to explain misrepresentation if we continue to ignore the role of the perceptual systems' outputs as the *actual* activators of representations. I've also explained the two counts on which I disagree with (b): we don't have senses and the job which needs to be done is not that of picking up on properties of objects. Instead of separate sense organs, we have integral *perceptual systems*. And what perceptual systems do is not pick up on properties of objects, they pick up *information*. What I mean is this: the job which needs to be done is not taking the properties of objects and turning these into the properties of representations of those objects. What perceptual systems do is take information about the properties of objects, which is already encoded in environmental energy, and encode this same information in a neurologically useable form. However, this rejection of (b) by switching to talking about "information" hasn't yet been fully justified. In this chapter and the next I'll make good on my promise to justify the claim that our perceptual systems can pick up information encoded in chemical, mechanical and luminous energy. I'll do this while explaining my rejection of assumption (c). Indeed one of the major points of this chapter is to put forward the idea that the way representations are used in the production of behaviour is crucially important to an explanation of what they represent. The way a representation is used gives the sort of thing the representation should represent, and the way environmental energy is used by being picked up and encoded as constituents of representations gives this encoded energy its informativeness. While discussing the way information is encoded as constituents of a representation, I'll also begin to address assumption (d), that abstract objects are not kosher, and should not be included in an explanation of perception and representation. I'm going to argue that abstract objects *need* to be accepted as a part of our ontology if we're going to be able to explain misrepresentation.

Doing all this follows the general tactic I applauded in Chapter One; that of explaining, in a principled, non-*ad hoc*. non-circular way, how a representation comes to correctly represent some things and not other things. And as I argued in Chapter One, if we can specify what the representation does represent, we can identify cases of misrepresentation as cases where this representation was activated by the perceptual systems picking up incomplete or poor quality information, and where this representation would not be activated if the information picked up was of better quality or was more complete. The main job of this chapter is to follow this tactic through.

I won't be providing much in the way of deductively valid arguments showing that things *must* be this way. Instead I'll demonstrate the worth of these positions by showing that these are *sensible* positions to hold, and showing that they give us the equipment to paint a coherent picture of representations and their use in representing and misrepresenting objects. We all know of many pieces of philosophical apparatus whose existence or necessity is not deductively proven, but these are still accepted as being pretty darn useful ways of looking at their respective problems. Here, I propose, is another.

To explain how a PATTERN OF ACTIVATION over a large interconnected network of nodes can do some representing, requires several pieces of a large jigsaw of interlocking ideas to all be in place. Many of these issues have been overlooked by traditional theories of representation; but they're issues which we *need* to be sensitive to. Some of them are simply essential facets of what it is to be human. Some of them are important when we look at what we have representations for. But the main reason for being sensitive to each of these issues is that when they are all fitted together they in fact form a picture of how a PATTERN OF ACTIVATION can represent (and also misrepresent) something. Until this jigsaw if ideas is assembled, it may not be clear exactly why a particular idea is important, or exactly where it fits into the big picture. Nevertheless the importance and placement of each of the pieces will be apparent when all the pieces have been laid out; then we will be able to tell a coherent story about what a representation is, and what it represents.

I'll identify each of the ideas we need to incorporate first. Then I'll go through and explain them all slowly (though not always separately) over the next two chapters. The pieces of the puzzle are these:

(1) We *use* our representations to do more than just represent things with them.

(2) The use of our representations has little to do with getting from perceptions to knowledge of which objects are "out there". (This is done in traditional stories by employing some knowledge of objects and their behaviour, and accomplished through a reasoning processes like abduction and induction.[1])

(3) The main use of our representations is to successfully interact with objects in our environments.

---

1 See Peirce (1955) Abduction is the adoption of an explanatory hypothesis, and induction is rating the hypothesis with a certain amount of confidence when it facilitates the prediction, and subsequent observation, of novel phenomena (under the assumption that this event is representative of a whole class of similar events).

(4) The representations I have now weren't magically "given" to me; they developed slowly over my lifetime from the representations I had in infancy.

(5) I don't first develop a representation and *then* subsequently use it, rather my representations are developed *by being used* .

(6) My representations are used to help me employ objects to them service of my needs and desires.

(7) What I can *do* with an object is an important part of a representation of that object.

(8) Aspects of the environmental energy which my perceptual systems pick up on come to be informative by being used as parts of representations. They become informative through my history of using this environmental energy, encoded as constituents of representations, to facilitate my interactions with certain sorts of object.

(9) Each encoded piece of information which is a constituent of a representation stands in a semantic relationship which I will call "indicates" with some aspect of an object or state of affairs. The relation "indicates" is also something which comes about through the way this constituent of the representation has been used.

(10) Some patterns of activation encode (and indicate) perceptually detectable information about aspects of objects. These are aspects like how the object looks, what it smells like, what it tastes like.

(11) Information about these properties is picked up by being recognised by pattern recognisers processing the "first level" outputs of the perceptual systems.

(12) These patterns of activation can in turn activate other patterns of activation, which can activate other patterns of activation. These "deeper" patterns of activation encode information about properties of objects too. (Depending on the level at which these are encoded, these are properties like being made of metal, being a cat, being raw, being spherical, being a cricket ball, being a meeting of the local Rotary Club.)

(13) These "deeper" patterns of activation are activated by their own pattern recognisers, which pick up on information encoded in combinations of "shallower" patterns of activation.

(14) Representations are the structures which link perception with actions, so representations need to be connected to patterns of activation whose activation initiates actions.

(15) We are *agents*. Indeed we're agents in two senses. Not only are we not passive absorbers of information but active gatherers of information, but also we're active in non-perceptual ways. This means

we have to accommodate the whole range of human actions into our picture of what it is to be a representation.

(16) The associations between encoded actions and encoded perceptions which constitute my representations embody habits, or hard-won rules of thumb, for using perceptually picked up information to interact with objects.

(17) The semantic relationship which a whole representation has with an object is "piggyback" on the semantic relationship which each of its constituents has with a property of an object. Each of the representation's constituents indicate some property of an object, and the constituents all together *express* an abstract object with all these properties.

(18) The sort of object correctly represented by such a representation is characterised by the abstract object expressed.

(19) The representation's relationships could be diagrammed as in Figure 4.1.

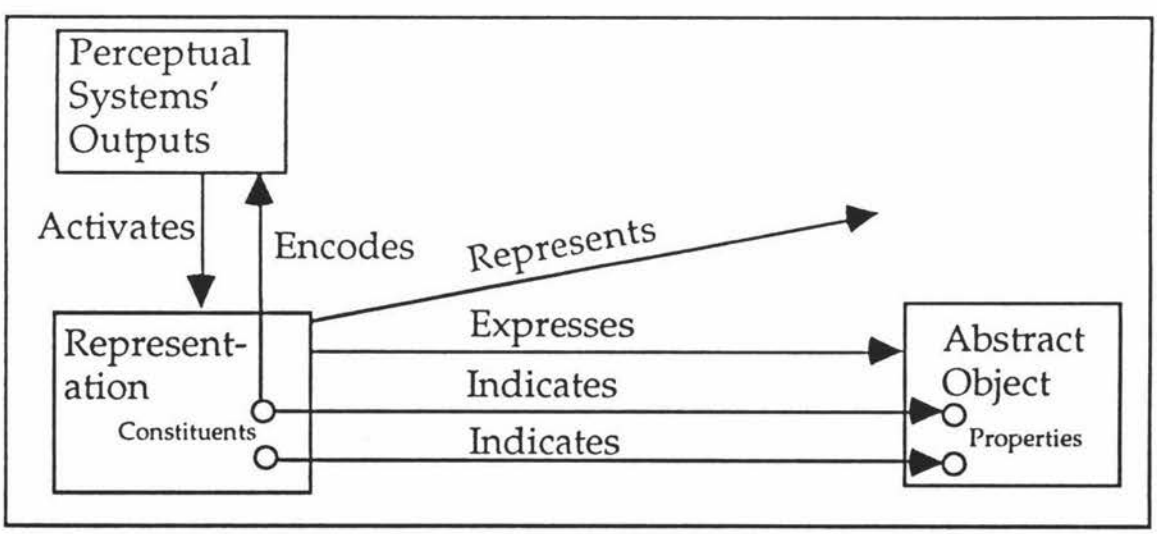


Figure 4.1 The representation's semantic relations.

Given that the above issues are so interdependent, it's difficult to discuss them separately. Nevertheless I shall attempt to discuss each piece, show where it fits into our picture of a representation, and when necessary revise this picture accordingly. When all these pieces are taken together and all the revisions have been made, we'll end up with a picture which shows how a PATTERN OF ACTIVATION can have constituents which encode information. Because they can do that, the PATTERN OF ACTIVATION can be a representation, and can express a particular abstract object which characterises the sort of object this representation represents.

4.2 What are representations for?

If I asked the question, "what do I have representations for?" a lot of the answers I'd receive would be something like, "so that I can make true statements," or "so that I can think true thoughts." This is barking up the wrong tree altogether. My having representations has little to do with truth. To be sure, representations are used to veridically represent, but that's not their main use. For the most part they're used by agents to *successfully interact*. I'm an agent, I do things, and I have representations so that I can use these representations to facilitate the things I do. I don't have a **coffee** representation purely for the purpose of being able to represent coffee, just so that I can know when coffee is present in my environment. I have one so I can use it to successfully interact with coffee; so that I can buy, grind, make, drink, serve and find coffee.

So in answer to the question of what I use my representations for, I have them to help my interactions with objects be *successful*. By "successful" interactions I don't mean "knowing what objects are actually present," or anything to do with the truth or veridicality of my representations. What "successful" means is managing to survive. *Mother Nature is interested in survival, not truth.* "Successful" means serving my needs and desires, meeting my goals and aspirations. One aspect of successful interaction is eating: crying when as an infant I needed to be fed, realising what things satisfy my hunger and which ones don't, deciding what foods I like and making my preferences known, realising that some things which look edible are poisonous, being able to go to the supermarket to buy ingredients for dinner. Successful interaction with my environment also means being able to get around: being able to crawl across a room to where something interesting is, being able to avoid obstacles while crawling across the room, learning what sort of things I can lean on in order to support myself when first attempting to walk, learning to ride a bicycle, learning to find my way to the bathroom in the dark, being able to negotiate my way across a crowded marketplace to get to the stall which is selling broccoli for the cheapest price. To do all these things, I need representations which enable me to react appropriately to perceptual information.

Being an agent has been severely undervalued in traditional accounts of representation, and in the account of a representation we've had up to the beginning of this chapter. Being a doer of deeds is a fundamental facet of human life, our representations are used by agents to successfully interact with objects. How a representations is used, and for what purpose, is vital to an explanation of what a representation represents. Ludwig Wittgenstein gives similar advice when talking about what a word means; he says:

> "For a *large* class of cases–though not for all–in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language."[2]

We could interpret this as advising that to know the "meaning" of a word, we should look at the way the word is *used*, rather than looking for the things the word stands for. We can apply this tactic to a representation as well as to a word: to know what sort of thing a representation "represents," we should look closely at the way the representation is used, rather than searching for what it "stands for." The idea that a representation's semantic relationships come from the ways the representation is used by agents to interact with objects is going to be central to establishing what the representation represents.

4.3 Representations connect perception with action.

To do this we'll need to examine some of the details of the ways I use my representations then. As I said earlier the main job representations do is to facilitate the causal relationships between perception and action. Representations *are* the structures which connect perception with action. I will be especially interested in two sorts of action: those that I perform in order to pick up information, and those I perform when interacting with objects.

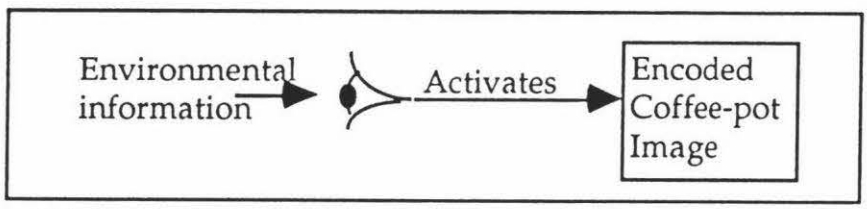
I stressed in Chapter Two that proprioception is an essential aspect of the pickup of information. But the way I stressed this, I down-played the way I perform actions to pick up information. As well as encoding information about the positions of my muscles I'm now talking about initiating actions in order to pick up information. I don't just passively *receive* information from my environment, I actively investigate my environment and seek out information. I look at, listen to, smell, feel, taste and otherwise *act* to pick up information. (Note that these are all verbs.) My perceiving my environment depends a lot on my acting so that I can attend to different aspects of my environment: moving my head so that I can point my eyes towards what I'm trying to look at; adjusting my eye muscles to focus my vision; breathing in through my nose to smell; putting my hand out to feel the texture of things. It's important to be sensitive to the fact that my perceptions of my environment are active. Because of this we must incorporate the actions involved in seeing, feeling, tasting, smelling and hearing into our picture of a representation. These sorts of actions are going to need to be encoded as constituents of the representations I use to perceive

---

2 Wittgenstein (1958) §43

my environment. By having patterns of activation which initiate actions when activated, physically connected to patterns of activation which encode environmental information, the actions can be co-ordinated with my perceptions of my environment. The activation of some piece of encoded information can cause the activation of a pattern of activation which initiates a certain action.

For instance, I have a pattern of activation which encodes the information I pick up visually when I go to the common-room to get coffee. This pattern can be activated when I actually see the pot of coffee on the hot plate in the corner.



4.2 The pattern of activation which encodes visually picked up information that there is a pot of coffee on the hot-plate .

Another way this pattern of activation can be activated, is by the activation of my **coffee** representation in certain situations. Suppose, for example, that I'm passing the common-room and the pattern (5,4,2,5,3,6) is activated over the nodes connected to the outputs of my olfactory perceptual system. This pattern encodes the information I've picked up that there is coffee present somewhere in my environment. The activation of this pattern (along with the context: where I am, what I'm doing etc.) causes the activation of my **coffee** representation, which in turn causes the activation of certain other encoded pieces of information. One of these pieces of information is this information I normally pick up visually when I walk into the common room and perceive a full pot of coffee on the hot-plate in the corner. Although this pattern of activation is normally activated by picking up the information it encodes from the optic array, in this case it was activated by my **coffee** representation being activated by the encoded smell information I picked up.

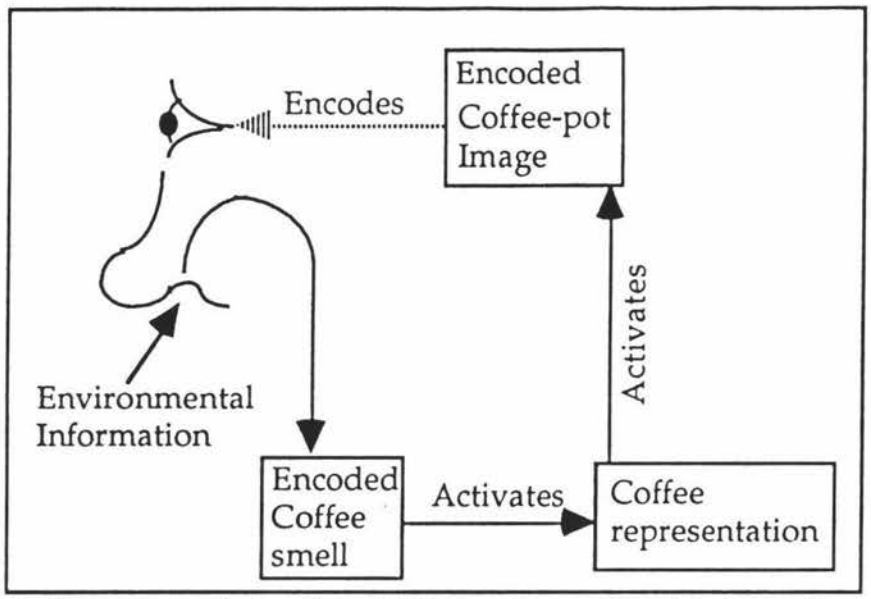


Figure 4.3 The pattern of activation which encodes visually picked up information is activated because a representation it is a constituent of is activated.

The activation of the pattern which encodes the visual information could cause the activation of a pattern of activation which encodes the actions involved in poking my head in the door and looking to see if there's any coffee in the pot on the hot-plate in the corner. This pattern of activation which encodes actions, would initiate these actions when activated.

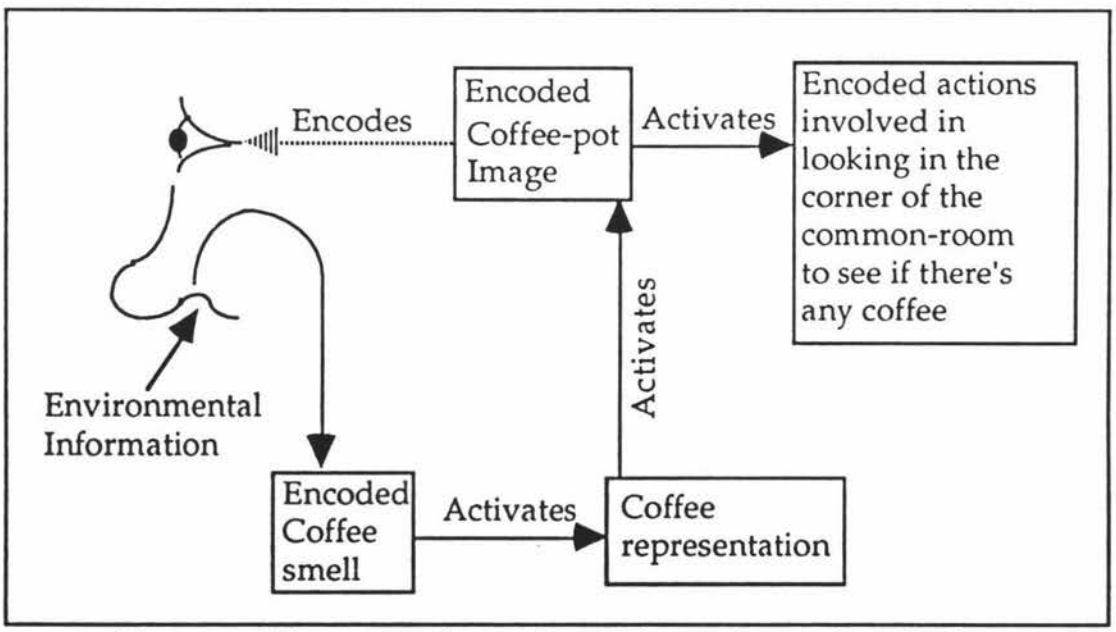


Figure 4.4 The pattern of activation which encodes visually picked up information in turn activates a pattern which encodes actions.

Having this pattern of activation which encodes the information I would pick up visually, activated "indirectly" through the activation of the pattern which encodes smell information, causes me to look to see if I would pick up the information that there is a pot of coffee on the hot-plate in the corner "directly" as it were, from the optical array. This is one way I use my **coffee**

representation: its activation causes me to perform actions which facilitate my picking up further information.[3]

This is not to say that there is a set action or response for every representation activated. Which actions will be activated and thus initiated depends on the context; on where I am, what I'm doing, what mood I'm in, how much coffee I've already had today, and so on. What behaviour an agent produces in a given situation is the result of a very complex interaction between *all* the agent's beliefs and desires. It's a gross oversimplification to say that in situation X, I will perform action Y simply because action Y is encoded as a constituent of representation **X**. Admittedly I'm not going to perform action Y unless **X** is activated (if **X** is the only representation connected to the pattern of activation which encodes action Y). What I mean to say, is that action Y is one of many possible responses I might make to my belief that X is the case. Whether I will perform Y or some alternative action or no action at all, depends on my current beliefs, desires, etc. all interacting to activate certain constituents of a representation. (Recall the flexibility connectionist representations have, which I discussed at the end of section 3.2, with regard to my **spaniel** representation.)

But I don't have representations simply in order to use them to *perceive* my environment and understand what I perceive; I'm an agent, I *interact* with objects. I use my representations to facilitate these interactions with objects. The sorts of actions we've been talking about so far are ones like looking closely, feeling, smelling and so on. These are legitimate actions to be sure, but a class of actions which centre around simply *perceiving* objects. I don't have a **coffee** representation simply to enable me to find and identify coffee; I have a **coffee** representation so that I can also use this representation to successfully interact with coffee once I've found it. For example, I use this representation to make, drink, or buy coffee.

Because I use my representations to help me interact successfully with objects, many of my representations of objects must also have constituents which encode the actions I perform when interacting with the sort of object represented. The activation of these patterns of activation cause me to perform those actions. For example, my **coffee** representation would have encoded as constituents the actions I must perform in order to make, drink, and buy coffee, somehow encoded as routines to enable me to perform them. Thus when my **coffee** representation is activated certain encoded actions can be activated as well. Drinking coffee is a simple routine which breaks down

---

3 What I've been trying to illustrate with this paragraph and these diagrams, is how the activation of an encoded action can be fitted into a connectionist network. I realise I've only sketched this. The idea could certainly use more developing. Unfortunately I'm not sure how to do this. (Colin Martindale (1991) takes a promising stab at this problem on pp193-199.) I think, personally, that incorporating actions into a connectionist network is one of the hardest problems connectionism now faces.

into muscle movements such as bringing the cup to my mouth, slurping, swallowing and so on. Making coffee is a bit more complex, and may involve sub-routines by which I boil the jug, get the coffee-tin from where it's kept, find a clean mug and so on. Buying coffee requires another sort of action to be encoded as a constituent of this representation. To buy coffee I need to be able to produce certain words (among other actions); I need to be able to say "A cup of coffee, please." So as well as physical muscle movements and the like, I must also have the actions involved in the production of certain words encoded as constituents of this representation.

Such representations thus expanded, *are* the connections between patterns of activation which encode information picked up from my environment and patterns of activation which encode actions. More generally, representations are the neurological structures by which actions are coordinated with perception. All the actions I perform when going about my daily business are performed because they are prompted by the activation of representations of which they are constituents. And the activation of these representations is, for the most part, a result of my perceptions.

4.4 The co-evolution of perception, action, and information.

I've been looking at two important relations in the previous section. One is the relation between the information carried in environmental energy and organisms picking up that information, and the other is the relation between picking up information and the production of actions. (These relations together are what makes certain features of the ambient energy *informative*, and they're also what gives representations the ability to represent.) Now I want to turn my attention to how these relations come about.

The fact that I have certain perceptual systems by which I can pick up on environmental information useful to me is a result of the evolutionary path my ancestors took. Sharks, rattlesnakes, bats, and electric eels each have perceptual systems by which they pick up on environmental information which is useful to them. I cannot pick up on this information. This is because their ancestors took different evolutionary paths to that which my ancestors took. The perceptual systems of each species, evolved because each individual of that species needed to be aware of certain facts about their environments to ensure that they survived and reproduced.

Ecological psychologists use the concept of a *niche* to refer to how an animal lives within its environment.[4] A niche is a set of things an organism needs which its environment provides, such as shelter, food, hiding places,

4 Gibson (1979) : p128-129.

methods of locomotion, and so on. Gibson calls these things a niche provides for an organism's use *affordances*. (I'll discuss these affordances in more detail later on.) So as Gibson puts it, an organism needs to have perceptual systems which enable it to pick up on information about the affordances an environment provides.

A recent development of Gibson's ideas by Varela, Thompson, and Rosch points out that as well as a species' niche shaping that species' perceptual systems over successive generations, the niche is simultaneously shaped by the species that inhabit it; that is, the species and its niche are "structurally coupled." Because of this complementary process, the environmental information available for an organism to pick up has *co-evolved* with the perceptual systems by which those organisms pick up on that information. The colour-vision of bees provides a good example of such co-evolution. It seems that the ultraviolet reflectance of flowers was not always the way it is now. Once flowers didn't reflect ultraviolet light at all. Now they reflect patterns almost advertising themselves to bees. The way flowers reflect ultraviolet light has co-evolved with bees' ability to pick up on information contained in the ultraviolet light. This happened because of the mutuality between bees and these flowers; flowers need to be able to attract pollinators by their food content and to remain distinguishable from other flowers, while bees need to be able to recognise the flowers which afford the best supplies of food.[5] Thus the information available to the bees, and the bees' ability to pick up on this information have produced each other. Varela et al. claim that this illustrates that environmental information is not independent of an organism and always available whether or not anyone picks up on it, as Gibson and his followers maintain. Rather this information is brought forth, or in Varela et al.'s terminology *enacted,* by histories of such structural coupling.[6] Through this history a species evolves with perceptual systems capable of picking up the information individuals of the species need to survive in their niche.

As well as the relations between perception and information co-evolving over successive generations, the relations between perception and action have also co-evolved. In some organisms, such as insects and so-called "lower" forms of cognitive animals these relations are the product of genetic patterns they are born with. The patterns having developed over the species' history of structured coupling within its environmental niche. In other organisms, like humans for example, a significant proportion of these

5 Lythgoe (1979)

6 Varela, Thompson et al. (1991) : p201-202.

relations between perception and action are not inborn, but have developed together over each individual organism's lifetime. We call this learning.

This is a very important point. The connections between encoded actions and encoded information which constitute my representations weren't magically "given", but developed slowly over my lifetime. The representations I had when I was one year old evolved into the representations I had when I was six, which developed further into the representations I had when I was sixteen, and these evolved into the representations I have now. My representations have developed progressively over my lifetime, as I gradually learned to use my perceptions to guide the actions I perform, so that my actions have become more successful more often. This learning amounts to the connections between encoded actions and encoded information which constitute my representations being adjusted, revised, added to and expanded. This took time and experience.

Varela et al. emphasise that perception and action are "fundamentally inseparable in lived cognition",[7] claiming that perception and action are not merely contingently linked in cognitive individuals, but have evolved together. This mutuality between perception and action (as well as the mutuality of organism and environment, by which the organism comes to have the perceptual systems it has) brings forth a view of cognition as *embodied action*. By embodied action, they mean to emphasise that cognition depends on the kinds of experiences that come from having a body with various sensorimotor capacities. Perception alone cannot produce an organism capable of picking up *information*. It's through an agent's interactions within its environment, that patterns in the environmental energy which already encode information come to be informative. These patterns in the environmental energy might already convey information to others, but they come to convey information *to me*, by *my* interactions. Through my experience in interacting with objects I learn to pick up information about objects and how they might be useful to me.

Varela et al. similarly argue that the cognitive structures by which actions are guided by perception (representations[8]) develop through using perception to guide action. They use as an example Held and Hein's experiment[9] to illustrate this point. In this experiment pairs of kittens who

---

7 Varela, Thompson et al. (1991) : p173. A note in praise of Varela et al.'s book: I had been trying to articulate the way perception and actions develop *together* to form my representations themselves for ages, without being able to nail down exactly how to express what I was trying to do. Then when I read this book I found written down just what I had been attempting to say.

8 Varela et al. don't call these structures representations. They actually hold that a lot of such interactions using perception to guide action don't involve representations at all. Calling these structures "representations" is something I as well do with reservations, as I'll explain in Chapter Five.

9 Held and Hein (1958)

had been kept in the dark were exposed to visual stimulation under special conditions: one of each pair was put in a special carriage which allowed it little freedom of movement, while the other of the pair was put in a harness by which it pulled the first kitten's carriage around. Each of the kittens therefore had similar visual perceptions, but only one of each pair had experience of *acting* according to the visual perceptions. When the kittens were released from this apparatus after two weeks, only those who'd had experience at guiding their actions according to their perceptions behaved normally. The kittens who had been pulled around, and had had no such experience, acted as though they were blind: they fell off edges and bumped into things. This experiment illustrates elegantly the idea that cognition comes from embodied action, from *perceptually guided action*. It seems that because the patterns in the luminous energy had not been *used*, the kittens had not learned to pick up the *information* encoded in these patterns. Varela et al. use this to show that "cognitive structures emerge from the kinds of recurrent sensorimotor patterns that enable action to be perceptually guided."[10] Using the terminology I've been developing, this translates as: the ability to pick up information conveyed in the ambient luminous energy co-evolves with the ability to use that information to guide actions. This co-evolutionary process develops the connections between encoded information and encoded actions which make up the organism's representations. Representations are formed by using perception to guide action.

The same conclusion can be drawn from Jean Piaget's investigations, not with kittens, but with human children. The way Piaget describes the different stages of children's cognitive development provides an excellent illustration of the progressive development of the interconnections between perception and action in representations.

Piaget calls the first two years of a child's life the "sensorimotor period,"[11] because during this period children learn about the world and express this knowledge through their senses and motor skills. During Piaget's "First Stage", the relations between the infant's perceptions and his or her actions are basically constituted by the reflexes "hard-wired" at birth into the sensory neuron-motor neuron interconnections.[12] For example. when an infant perceives a stroking or tickling sensation on her cheek, she will automatically act to move her head towards the side which was stroked,

---

10 Varela, Thompson et al. (1991) : p176.

11 The following description of Piaget's theory of cognitive development is taken from a summary of it in Berger (1988) : pp 122-131.

12 Kalat (1988) discusses this and other infant reflexes, (pp209-210), and also explains the way the sensory neuron interacts with the motor neurons , via interneurons in the spinal column (pp26-27).

and when she feels anything touch her lips she will suckle. These reflexes are useful (even essential) for her survival; reacting to a touch on the cheek by turning the head and suckling helps the infant find a nipple to get sustenance from. Other useful reflexes are gripping anything which touches her palm, and staring at anything which comes into focus.

In Piaget's "Stage Two" (1-4 months) the infant's reflexes begin to adapt to the environment. The infant begins to coordinate actions and perceptions in this stage, and attempts to tailor her actions to her perceptions: for instance, on hearing a noise she can turn her head to try to look around to find its cause, although seldomly in the right direction. She can also see an object and try to touch it, also rarely successfully. The infant also begins to classify objects; for example into those which are sucked for nourishment, those sucked for pleasure, and those not to be sucked at all. At about three months old a baby will spit out a pacifier when hungry, and will suck a pacifier differently from a nipple. Infants in this stage have also been shown to grasp more frequently at objects which are the right size and shape for grasping.

In "Stage Three" (4-8 months) infants become much more aware of how their perceptions convey information about objects and people. They develop ways of interacting with people and objects to make interesting and exciting experiences. They vocalise a lot more as they realise that other people can respond; they enjoy making a noise, listening for a response, and answering back. Their interactions gradually become more successful during this stage, presumably as they get better at picking up information, and at encoding that information in such a way that actions appropriate to the situation can be produced more reliably.

In "Stage Four" (8-12 months) some really interesting things begin to happen. One of the most important for our purposes is the development of what Piaget calls *object permanence*. The child realises that objects continue to exist even when out of sight. Before 8 months old, if a child drops a favourite toy out of sight, she will find something else to be interested in. Even if this stage three child sees the toy being hidden under a blanket, she will lose interest in it: "out of sight" is literally "out of mind." After eight months old however, the child realises that a toy which is out of sight is still there; she can look under the blanket the toy is hidden under, or crawl after the rolling ball and get it back out from beneath the chair it has rolled under. Here we might say the child is learning to pick up information about occlusion; about how objects can hide other objects. In this stage the child also develops more deliberate interactions with objects; she no longer shakes everything, sucks everything, or drops everything. She begins to perform different actions depending on the object she's interacting with. The child

also learns to anticipate events, based on her perceptions. Infants in "Stage Four" have been known react to the information they pick up, by, for example, squealing with delight on hearing the bath water running, and by keeping their mouth shut on seeing spinach on the end of the spoon at dinner time.

In "Stage Five" (12-18 months) the child becomes what Piaget called "the little scientist." The toddler actively experiments with objects, asking herself "what happens if I do this?" and "what else can I do with this?" This stage is characterised by the toddler actively investigating objects and their uses.

"Stage Six" (18-24 months) is also an interesting one from our perspective. In this stage infants learn to try out various actions mentally before they act, are able to invent new ways to achieve a goal, can reproduce behaviour seen in the past, and demonstrate an ability to create mental "images" of things and actions that are not actually in view. These abilities are illustrated by the child's developing what Piaget called *full object permanence*. With full object permanence the child is convinced that objects can't just disappear, and will search for objects when they go out of sight. Most importantly, they are able to imagine where objects could be hidden, even if the object has never been discovered at this hiding place before. Piaget performed the following experiment on his daughter Jaqueline: he showed Jaqueline a coin, and then put it in his hand, put his hand under a cloth leaving the coin there, and then withdrew his empty hand. Earlier on, Jaqueline always only looked for the coin in the place where it was seen to disappear: her father's hand. But by the age of 18 months Jaqueline looked first in her father's hand, but on seeing that the coin wasn't there, without hesitation looked under the under the blanket. Her ability to "use mental combinations," as Piaget put it, enabled her to imagine where to search for the coin, even when she didn't see it being hidden there. The ability to make some sort of connections between different representations is demonstrated in a "Stage Six" child's beginning to pretend. Pretending typically involves combining actions from one context with actions in other different contexts: making car noises while pushing a toy car around the floor, for example.

It can be seen from Piaget's description of a child's cognitive development that an infant's interactions with the world start out clumsy and uncoordinated, and are rarely successful. But gradually infants learn to act in ways which will get their needs and desires fulfilled more successfully. The important point to notice for our purposes is that this learning–this progressive refinement of the child's representations–happens through the child's *experience in being an agent*; through her experience in attempting to use her perceptions to interact with objects and refining the connections

between encoded information and encoded actions to make these interactions progressively more successful. The same is true of the rest of us.

To develop my representations to the stage where I have a usefully coherent set of representations which I can use to successfully interact with my environment, has taken a long history of using my representations to attempt to interact with my environment, and striving to do this more successfully. My representations, as interconnections between encoded perceptual information and encoded actions, have developed through my lifetime of *embodied action.*

4.5 What representations' constituents encode information about

When talking about the aspects of objects which patterns of activation encode information about,[13] there is a temptation to concentrate on properties like being a certain size, mass, shape, colour, position, and so on: the sorts of "objective" properties scientists can measure and describe. This temptation is a hangover from viewing representation as concerned principally with truth (as opposed to successful interaction), which requires an objective viewpoint we can all agree on.

The sorts of properties which my representation's constituents encode information about are not all this sort of "objective" property. A representation's constituents also encode information about the properties objects have *in relation to perceivers.* These properties are in two types: One is the properties which objects have in relation to perceivers' *perceptual systems;* properties like feeling rough, being too heavy to lift off the ground, tasting sour, looking round, blue and flat, being a size and shape which fits comfortably in my hand, feeling cold, and sounding like my telephone. The other sort of perceiver-relative information picked up and encoded as constituents of my representations, are properties objects have in relation to perceivers *who are agents.* These constituents encode information about the uses an object can be put to; they encode information about what I can do with an object. I'll come back to encoding information about the uses of objects soon.

Picking up perceptual aspects of objects is done by the pattern recognisers which process the outputs of the perceptual systems. These recognise smells, colours, visual shapes, sounds, how things feel and so on. The patterns of activation at the outputs of these pattern recogniser networks

---

13 I have mentioned briefly, and I'll explain in more detail later, how this pattern of activation comes to encode information *about* an environmental object. It comes from the way the pattern of activation developed along with patterns which encode actions, through being used in the production of behaviour. Chapter Five explains in more detail where this semantic property comes from.

encode information about perceptual properties: how things look, smell, taste, feel and sound.

But as well as these perceptual aspects we do need patterns of activation which encode information about "objective" non-perceptual properties. Information about these more "objective" properties is picked up from the patterns of activation which encode information about perceptual properties. For instance, I need to get from the aspects of looking spherical and feeling spherical to the property of *being* spherical. This is not a difficult problem though. Recognising that something is spherical is just a matter of detecting when either the pattern which encodes that the object I'm looking at looks spherical or the object I'm feeling feels spherical are activated. To do this I could have connections from the patterns of activation which encode this perceptual information, to patterns which encode the information that the object I'm looking at and/or feeling is spherical. This pattern of activation would be activated *automatically* if the pattern of activation which encodes the information that the object I'm looking at looks spherical, and/or the pattern which encodes the information that the object I'm feeling feels spherical were activated. The pattern of activation which encodes the information that the object is spherical could come to encode information about this property through my experience in interacting with objects which look and feel spherical. This pattern of activation and its connections to other patterns of activation have developed through being used to help guide my actions when interacting with spherical objects.

In a similar way, I can also have a pattern of activation which encodes information about the property of being made of leather, which developed through my practice in dealing with things which feel leathery, look leathery and smell leathery. A pattern of activation which encodes the information that something is made of leather could be connected to the nodes which encode how leather objects look, feel and smell. By these connections the activation of one of these patterns of activation could cause the activation of the pattern of activation which encodes the information that the object I'm looking at/smelling/feeling/etc. is made of leather.

In a similar way I can pick up information about the property of being a cricket ball from patterns of activation which encode information about objects being spherical, feeling hard, being made of red leather and being a size which fits comfortably into my hand. Again a pattern of activation which encodes this information could be connected in such a way that when these "shallower" patterns of activation are activated they activate this "deeper" pattern of activation, which encodes the information that this object is a cricket ball. The activation of this pattern of activation, and all

these patterns of activation which are connected to it constitutes the activation of my **cricket ball** representation.

In this way we can have information encoded at deeper and deeper levels, away from patterns of activation which encode information picked up from the outputs of the perceptual systems, to patterns which encode information about properties which are recognised through combinations in the activation of other shallower patterns of activation. (Recall the examples I gave in section 2.9 about the flow of information through many layers of pattern recognisers, like the ones which pick up information about sound features, then phonemes, then syllables, then words for example.)

As we get deeper we can eventually have patterns of activation which encode properties which are not perceptually detectable at all, and which are only activated after the whole representation of that object is activated. My record player, for instance, is an object I bought for $20 at a garage sale a couple of years ago. This is not a perceptually detectable property of my record player. Rather it's a property encoded as a constituent of my **my record player** representation, which is activated only when I'm thinking about my record player (and possibly also about what it's worth, or what a good deal it was, or where it came from) and this whole representation is activated. Presumably it was encoded as a constituent of this representation through my experience of going to the garage sale, seeing it, hearing the sound it made, and paying $20 to the guy who was selling it, and also through my recounting the tale to my flatmates. So sometimes this constituent of my **my record player** representation is activated, and I'm reminded about what a good deal it was.

This information is "brought to my attention" in much the same way that information about an object being a cricket ball or about my environment smelling like there's coffee present is brought to my attention: by the activation of the pattern of activation which encodes this information. It just happens that patterns of activation which encode information, are at different levels, and encode different sorts of information.

4.6 Encoding Affordances

Another sort of aspect we need to encode as parts of my representations comes from the purpose of having representations. As I've been repeatedly stressing, I don't have representations like this simply to be able to represent objects, to detect their presence and know their properties when they are present. I represent objects, so that I can use the information encoded by the constituents of these representations to achieve my aims, to carry out my

projects. I have representations to facilitate my interactions with objects. Because I'm use objects to carry out my projects, I need to pick up on information about the ways I can use objects. Thus some constituents of my representations must encode information about the *uses* the object represented can be put to.

I mentioned earlier the term *affordance* coined by Gibson to use to refer to the use provided by an object. An affordance is basically a use which an object lends itself to.[14] Gibson says that all objects which we perceive afford uses: the ground affords support; a house affords shelter; a stick affords leverage, wielding, leaning on, striking, raking and throwing; water affords drinking, swimming and washing; fruit affords eating; air affords breathing; a thread affords knotting, binding, lashing, knitting and weaving; another person affords company, friendship, competition, assistance, and a rich multitude of other affordances.

I make use of objects in my environment in countless ways: I eat them, sit on them, stand on them, drink them, make structures out of them, tear them down, cut them, cut things with them, get to work on them, work with them, throw them, catch them, tie them in knots, surf on them, swim in them, kick them, pat them, feed them, and wear them. I use various objects as levers, tables, weapons, steps, projectiles, paperweights, containers, and seats. We use objects to further our own ends, and use representations to help us do this. Often what matters most to us about objects is not what they *are* but what we can use them for.

An example of the ways objects can be used: Lemons are very useful objects, they can be used to make lemonade, they can be used as the ball in a game, to make the taste of pesto more "vibrant", to throw for the dog to chase and fetch, as projectiles to throw at the neighbour's randy cat which is howling on the back fence, and to make lemon icing for a carrot cake. I somehow need to be able to encode this sort of thing, the affordances a lemon can provide as constituents of my **lemon** representation.

The idea that some patterns of activation of representations encode information about affordances is pointed to strongly by the way our representations develop by being used to interact with objects. Early in life we are cued up to notice what objects afford, rather than what objects are. An infant in Piaget's "Stage One" has reflex associations between perceptions and actions so that she reacts to feeling a stroking or tickling sensation on her cheek by automatically moving her head towards that side, and when she feels anything touch her lips she will react by suckling. These are not merely associations of perceptions and actions, they are ways in which the baby is

---

14 See Gibson (1979), Chapter Eight, "The Theory of Affordances": pp127-143.

geared towards survival. She survives by being "hard-wired" to find objects which afford food. The ticking sensation on her cheek is used to help her take advantage of objects which afford food, by automatically causing the action of turning her head towards the side tickled. These sensations and actions are associated around the beginnings of a representation employed to find and use objects which afford food. They're not associated around the idea of any specific type of object, or around some property those objects might have. These perceptions and actions aren't centred around the beginnings of a **nipple** representation, or a **bottle** representation; the *identity* of the object isn't relevant, what's important is that it affords food. As the child develops, she begins to form many distinct representations from what once was one representation. As I said earlier, the infant in Piaget's "Stage Two" begins to differentiate between things which are sucked for food (nipples and bottles), things which are sucked for pleasure (fingers and pacifiers) and things which are not to be sucked at all (fuzzy blankets).[15]

The reflex of grasping with her hand whenever she feels pressure on her palm is another reflex action an infant comes "hard-wired" with. Picking up the information that something is touching her palm, and this grasping action are centred around the beginnings of a representation which we might call her **things-which-afford-grasping** representation, whose central pattern of activation encodes information about this affordance. By "Stage Two" the infant can also differentiate between things which can be grasped in the hand and things which can't be grasped, and will grasp more often at things which are the right size and shape for grasping.

Representations seem to start as a simple series of connections, like this:

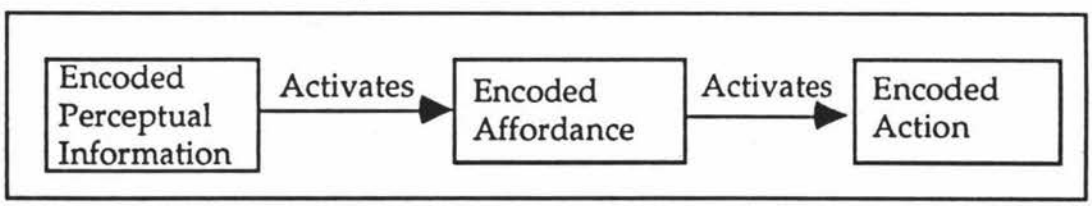


Figure 4.7 The beginnings of a representation.

Representations like these are the fore-runners of the complex representations the child slowly builds up. Recall Piaget's "Stage five", the "what else can I do with this" stage. In stage five the infant is amassing the beginnings of many representations, by investigating the affordances an object provides, and collecting perceptual information (principally images, I suppose) to associate with these affordances. The child is learning to pick up

15 Berger (1988) : p124

information about affordances by developing representations which have constituents which encode this information. Here the child also learns to coordinate the actions necessary to take advantage of the affordances provided by objects with the information picked up. Through this process, the child develops representations to encode the fact that objects which look like *this* afford *this* sort of use, which can be taken advantage of by performing *this* action. The representations developed come to have as constituents patterns of activation which encode information about affordances (like affords food, affords gripping, affords leaning on), as well as patterns of activation as constituents which encode information about properties of objects, and patterns of activation by which actions are initiated and guided.

It's important to notice is that what objects afford is often noticed first; what they are is secondary. My picking up on information about high, vertical, unmoving surfaces is what activates the representation used to find and lean on things which afford leaning on. We might call this my **things-which-afford-leaning-on** representation. This representation isn't activated by information about the identity of the object, different instantiations of it could be activated by information about a range of different objects. What is paramount is that the object affords leaning-on; what the lean-on-able object *is* is secondary. It's the affordance the object provided which was important to me as a toddler when this representation was being formed; this affordance of being lean-on-able is the central defining characteristic of the objects this representation represents.

What affordance is picked up depends as much on my needs and desires as it does on the object perceived. For example, imagine a rock which is flat, extended, and about knee high; this rock affords sitting on. When I see this rock I could pick up on information which would activate a representation with a constituent which encodes the information that the rock affords a seat, but this is most likely to happen *when I need to sit down.* I could pick up on this information about the rock if I need to sit, but in different situations other affordances could be picked up on–I might also pick up on information which causes the activation of my **obstacle** representation if I need to pass, or my **coffee-table** representation if I need something to put my cup of coffee on. (To Gibson the rock provides all these affordances, even if no one ever notices them.) What project I'm pursuing governs what affordances I'm more predisposed to pick up on, which affects what representation is activated. (This happens via the activation of contextual patterns of activation, as I explained in Chapter Three.)

As I've just illustrated, affordances are fundamental parts of my representations. As an infant, my representations began being centred

around affordances to do with food and the manipulation of objects. My representations multiplied and became more refined and developed in other ways as I learned to make distinctions, but it's not unreasonable to expect that patterns of activation which encode information about affordances still occupy central places in the physical constitution of my representations.

4.7 The overall shape of a representation.

There is a temptation to give affordances a *very* central position in the physical make-up of representations. Throughout Chapter Three (especially in figure 3.2), I illustrated and discussed networks like those in figure 4.5. In this chapter I've been explaining that the patterns of activation at the extremities of networks encode information and actions. These peripheral patterns of activation are centred around other patterns of activation at the "hubs" of these networks. The "hubs" are those nodes indicated by dots in the left-hand diagram, and by the patterns of activation over the central set of nodes in the right-hand one.

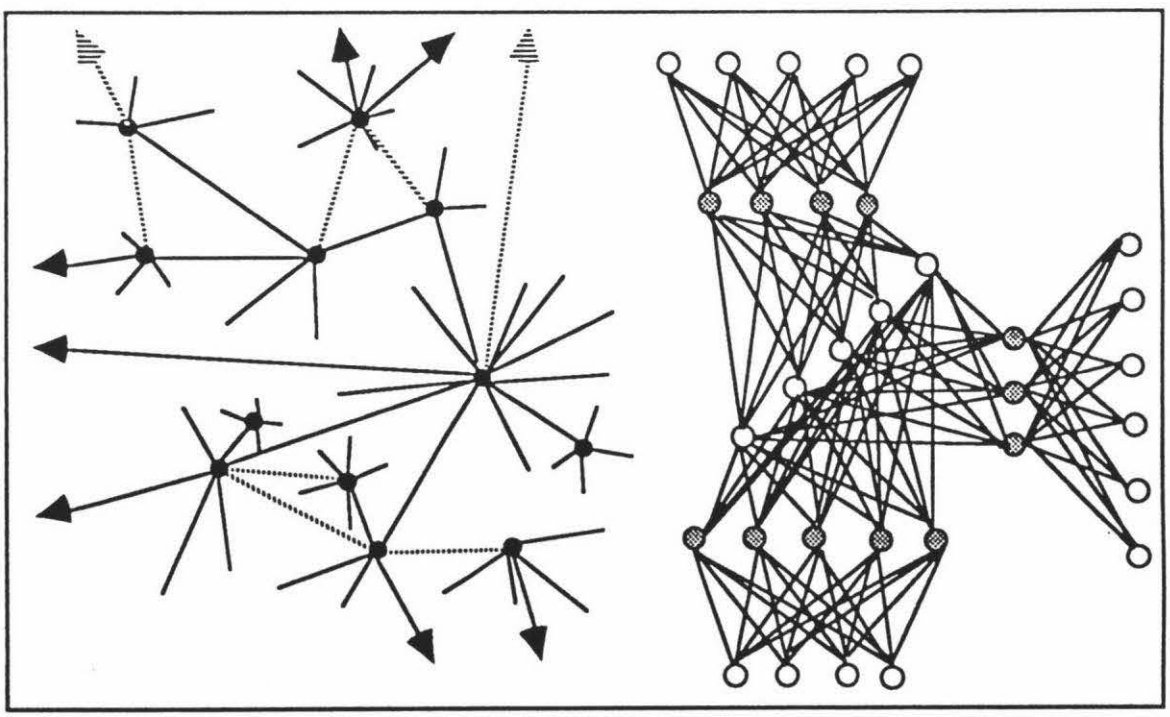


Figure 4.5

Initially I believed that the central patterns of activation, these hubs, encoded information about one thing. Later on I changed my mind, and for quite a while I was convinced that they must encode information about something alse. Nowadays I'm not very happy with either of these ideas. Unfortunately I don't have much in the way of knock-down reasons for rejecting these other two positions, other than that a third way of viewing

things seems to make more sense. For this reason I remain uncommitted to any of these three ideas, although most days I lean very strongly towards the third. Let me outline what these three alternatives are anyway.

It was initially tempting to say that the patterns of activation at the hub of these networks somehow encode the concept of the object which is the central theme of the representation; for example, it has been tempting to say that the pattern of activation at the centre of my **table** representation encodes my concept of a table. Because of this, it would be tempting to say that this central pattern of activation is used to encode the information that an object has the property of *being a table.* In figure 3.1 I illustrated my **coffee** representation in this way. There was a node at the "hub" of a large network of nodes, and this centre node was like the defining characteristic of my **coffee** representation; it could be said to encode the property of being coffee. As long as this central node was activated, we could label the PATTERN OF ACTIVATION made up of this node and whatever combination of peripheral nodes also happened to be activated as my **coffee** representation.

But after thinking about the way representations are used, I saw another candidate for the information encoded by this central pattern of activation (or node). These central patterns of activation could encode information about the aspect of the object most important to the perceiver as an agent who uses objects to pursue projects. They could encode information about affordances, the uses the object can be put to. The central patterns of activation of each representation encodes information not about an object *having a certain property,* but rather about an object *affording a certain use.*

Viewing representations in this way fitted better with the perspective that representation is about successful interaction. On this view I don't have representations of objects and *their properties* so that I can interact with those particular objects. I have representations of the *things I can do* with objects, so that I can achieve my goals using whatever object provides the affordance I need. In the example I used earlier, perceiving a high unmoving vertical surface can be used by a toddler to encode information about a thing which is lean-on-able. The identity of the lean-on-able surface–whether it's part of a wall, a door, a fence, a chair, a cabinet or even a person–isn't important to the toddler; what's important is that it can be leant upon. With this viewpoint, the central pattern of activation of this representation encodes the information that this object affords leaning-on.

But some objects provide lots of different affordances. It's hard to say what affordance the central pattern of *some* representations should encode information about. For example. a stick affords a lever, a walking stick, a weapon, a tent-pole, a fire-poker and a multitude of other uses. Which of

these is the central affordance of my **stick** representation? On this way of looking at representations, it could be that I don't actually have a **stick** representation , but that the uses I need the stick for prompt the activation of, for instance, my **fire-poker, lever, tent-pole, walking stick,** or my **weapon** representations. Each of these representations has a central pattern of activation which encodes information about an affordance, while the property of being a stick is merely a constituent of whatever representation is activated.

But in a similar way it could be that the central constituent encodes the information that this object has the property of being a stick, and this representation has peripheral patterns of activation which encode information about all these affordances the stick provides. And which of these constituents is activated depends on my needs at the time.

A third way of looking at the physical make-up of the PATTERN OF ACTIVATION solves the dilemma by taking some inspiration from Wittgenstein's *Philosophical Investigations*. Here we can bite the bullet, kick away the ladder, and maintain that there is *no central pattern of activation*. Rather than a representation looking something like the spokes of a wheel, with a uniting essential pattern of activation at the hub of the wheel, we could view a representation as looking more like a plate of sticky spaghetti. On such a view there are connections all over the place between patterns of activation, but there is no central pattern of activation which all the constituents of a representation are centred around. These two views are contrasted in Figure 4.6.

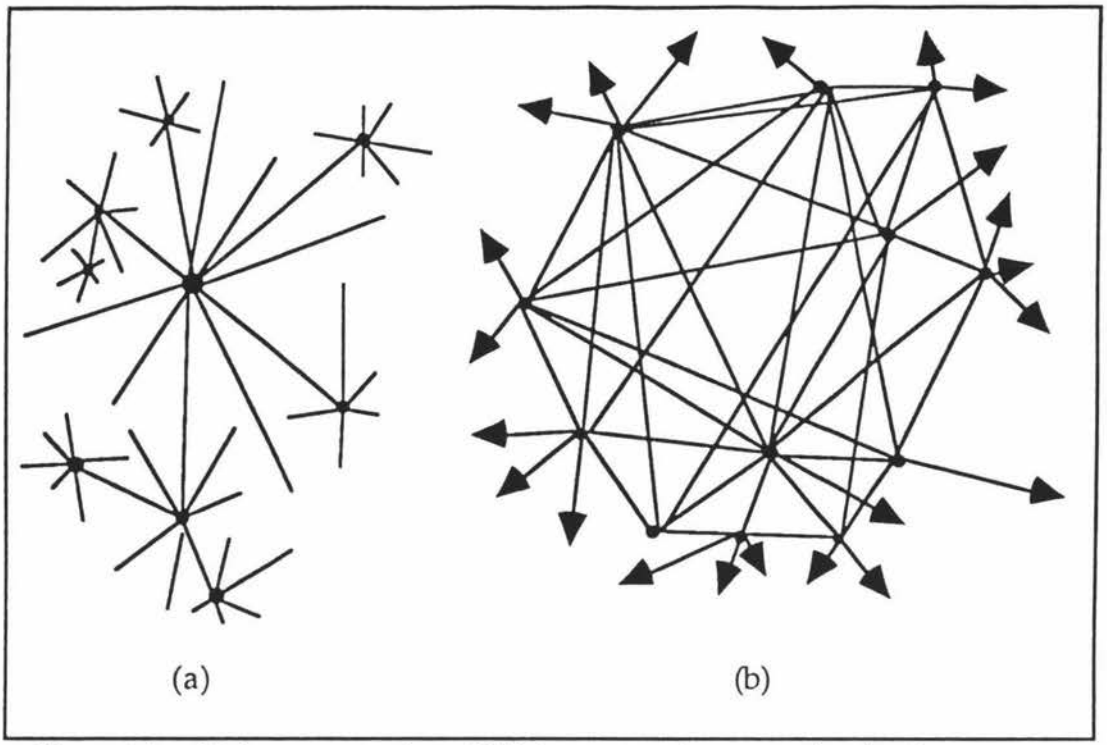


Figure 4.6 **(a) A representation which has a central pattern of activation.**
(b) A representation which doesn't have a central pattern of activation.

Wittgenstein maintained convincingly that there is no defining characteristic by which we can apply a word like "table" to an object. There is no essential property which makes an object the sort of thing which we can call a table.[16] There is not even an essential defining affordance (an idea which tempted me for quite a while); a painting of a table, and a doll's house table could be referred to as tables, but I can't use these to put my cup of coffee on. Likewise, I'm tempted to hold the view that there is no pattern of activation which must be a constituent of a representation for that representation to be my **table** representation. On this view, rather than there being a central "defining" pattern of activation which makes a PATTERN OF ACTIVATION an occurrence of my **table** representation, there is instead simply a "family resemblance" between each PATTERN OF ACTIVATION which we would like to call my **table** representation. The representation is merely a large, multiply interconnected network of patterns of activation. Each of these patterns of activation can encode information about affordances tables provide, the properties of tables in general, or the properties of specific tables. Any of these patterns of activation could be unactivated, and we could still call the PATTERN OF ACTIVATION activated my **table** representation.

I've now finished my survey of the constituent parts of a representation, and why it isn't a "list" of encoded sensations, as I first

16 Wittgenstein (1958) §66-67.

imagined it might be. Now I want to turn to the "deeper" question of how to bridge the semantic gap between a representation and an environmental object.

CHAPTER FIVE

BRIDGING THE "SEMANTIC GAP"

5.1 A different perspective on the way representations are related to objects.

I used to think, as traditional theories often do, that a perceptual experience is meaningless without some knowledge of objects, how they behave, and how they affect my perceptual systems. On this view, a representation's job is to help me figure out what objects are present in my environment, through some sort of inference from my perceptions (based on abduction; see footnote 1, Chapter Four). I need answers to the question: "How would the world have to be to explain my having this perceptual experience now?" According to the traditional perspective, in order to infer the location, presence, properties and behaviour of objects from our perceptual experiences, we need knowledge of objects, how they behave, and how they affect our perceptual systems. On this account, representations embody this knowledge; a representation is something like a proposition, which is used to reason out the state of the world from the state of our perceptual systems. Thus a representation incorporates propositions like "objects with property X cause sensation A."

I've been leading up to another perspective: what I have representations for, is not to represent truthfully, but to interact successfully with objects. Contrast these:

1) I need *knowledge of the ways objects affect my perceptions*, so that I can have good reason to expect that an object with a certain property is present by virtue of my perceptual experiences.
2) I need to have some *habits for the use of perceptually picked up information*, so that I can successfully interact with the objects I pick up information about.[1]

---

1 There are some people who take a third point of view: that Mother Nature doesn't select for truth, she selects for "good enough". For example the rabbit who activates fox often in situations where there isn't a fox present. The aim of the activation of this representation isn't veridical representation of the rabbit's environment, but a representation of its environment which is good enough to ensure the rabbit's survival. As long as errors are on the side of minimum harm or maximum benefit, that the representation's activations are "good enough" without needing to be true.

These folks hedge back from holding that the point of having representations is to *veridically* represent the representing organism's environment, but they are still working within the perspective that view (1) comes from; that representation is about being aware of your environment, rather than seeing this awareness as merely a means to the higher end of interacting within the environment in a manner that ensures prosperity.

What I shall now try to show, is that (2) is the job that a representati does. I'll also show that this job is doable by an interconnected set of encoded perceptions and actions, and also that looking at the problem in this way shows how a PATTERN OF ACTIVATION can stand in the desired range of semantic relationships with an object.

An example is a good place to start showing this. Imagine I'm a toddler walking unsteadily across a room, needing some support to keep me on my feet, such as a wall to lean on. I have a visual perception caused by seeing a wall on my left. From the perspective that my representations embody knowledge of objects and their behaviour, and the ways objects affect my senses, I need something like "a high, unmoving, vertical surface causes perceptual experiences like this." Great! Now I know *what* the object which caused my perception is: it's a high, unmoving, vertical surface.

But as a toddler needing support, I don't need to *know what it is*, I need to *lean on it*. A better perspective is provided by seeing this perceptual information I pick up as being encoded as a constituent of a representation which also has constituents which encode the actions involved in leaning. For me to be able to consistently use perceptually picked up information (encoded as a constituent of a representation) to successfully interact with a certain sort of object, I need a set of *habits or rules of thumb for the use* of the information picked up. And my representation, as the interconnections between encoded perceptual information and encoded actions implements exactly this habitual use.

The visual information that *here* on my right is an object which looks like *this*, is encoded as one constituent of a representation. Another constituent encodes the information that most often things which look like *this* can be successfully leaned upon if I'm in the right spatial relationship to them. A third constituent encodes the actions of putting out my arm to lean on the object. The connections between these three patterns of activation, mean that the activation of the first two patterns of activation causes the activation of the third pattern of activation, which initiates the leaning actions. (Of course this grossly over simplifies: Leaning is very tentative. There's going to be a lot of feedback about the stability of the objects guiding the leaning actions. Also the actions involved in getting to the right spatial relationship with the object need to be involved. So although I only spoke of three patterns of activation, there may need to be considerably more involved.) These connections implement my rule of thumb, or habit, for using this information to locate and lean on objects which afford leaning on.

According to view (1), the patterns of activation which encode properties and affordances of *objects* could be said to constitute my representations. These patterns of activation encode information about the

object I'm perceiving and its properties. But from perspective (2), it's important to include the patterns of activation which initiate *actions*, for two reasons. Firstly because it's through connections to these patterns of activation that I become able to pick up information, and secondly, to connect and coordinate perception with action are what we have representations for. I've already talked about the second, now I want to discuss the first for a while.

5.2 Representations' semantic relations come from their use.

It's through the use of perceptions to initiate and guide actions that we find semantic relations between constituents of representations and environmental objects. Wittgenstein advised us that the meaning of a word is, in most cases, its use in a language. In a similar way, the semantic relationship between a pattern of activation and some aspect of an environmental object is, in most cases, the pattern of activation's use in interactions with objects.

As I explained in section 4.4, networks are trained through trial and error. Through being used in guiding and producing actions, they slowly develop the ability to recognise patterns in their input and encode this as patterns of activation over their output nodes. For example, a representation which is used to pick up information about things to lean on and to use this information to guide my leaning, has developed through my attempting to walk, needing support, and leaning on things. Sometimes the leaning has been successful and sometimes I've fallen over. Through trial and error I've found that most often things which look like *this* can be successfully leaned upon. That is, by this history of use in the production of actions a pattern recogniser has developed the ability to recognise a certain pattern (looks like *this*) in the ambient optical array. That is, it develops the ability to pick up information about things which afford leaning on. It encodes this information as a pattern of activation on its output nodes. This pattern of activation has come to be connected to a pattern of activation which encodes my leaning actions. This, in turn, causes me to perform those actions.

Looking closely, we can see that picking up information and producing actions have co-evolved. Because it has developed by being used to guide my leaning actions, the pattern of activation which encodes my recognition of a high unmoving vertical surface has developed to encode information about objects which afford leaning. In turn, because it encodes the information that this object affords leaning, I habitually use the activation of this pattern of activation to initiate and guide my leaning

actions. The ability to produce actions and the ability to pick up information have produced each other. They are "structurally coupled".

So a pattern of activation has developed to encode information *about* a certain property. Thus "encodes information about" is a semantic relation; a rough one, which I'll now sharpen. It has come to encode information about a certain property because this pattern of activation has developed to be useful in finding and interacting with objects with that property, and has gradually developed to be more useful in such interactions, through being used in this way. Because it encodes that information, and because that information is used to guide and initiate interactions with objects with that property, this pattern of activation *indicates* that property.

I've just invented the word indicates[2] to use to refer to the semantic relationship between a constituent of a representation and a property or affordance of an object (which I've been referring to roughly as "encodes information about" a property or affordance). Thus a pattern of activation, as a constituent of a certain representation, indicates a particular property or affordance of an object, because of the information it has developed to encode.

In the above example of finding things to lean on, the pattern of activation which encodes the information that the object I'm perceiving affords leaning on, indicates that this object can be leaned on. It indicates something which can be leaned on because it has developed *so that I can use it* to find and lean on objects which can be leaned on, and because it has developed *by my using it* to find and lean on things which can be leaned on. In a similar way the pattern (5,4,2,5,3,6) activated over the nodes connected to

---

2 After I'd "invented" this term, I found that Dretske(1988) already had a "patent" on it. Dretske also refers to what a neurological mechanism *indicates* about environmental conditions. Dretske makes a distinction between three types of "indicators". (pp 53ff) His "Type III" indicator is very similar to the way a pattern of activation indicates some aspect of an environmental object.

Type I indicators have no intrinsic powers of representation. What they are supposed to indicate and their power to successfully indicate what it is their function to indicate are derived from another source. Road signs, maps, written and spoken language are type I indicators. They must have a meaning assigned to them through conventions.

Type II indicators are "natural signs", like bird tracks in the snow, which mean that a bird was here. These indicators also are not "intrinsic" indicators; this type of indicator gets its intentionality through conventionally exploiting the sign's natural (unconventional) meaning. Fuel gauges, fingerprints, melting ice indicate the amount of petrol in the tank, murderer's identities, and the temperature only if someone takes them to indicate this. The indicator would not indicate what it does if no-one observes them and makes the appropriate inference.

Type III indicators are the ones which have intrinsic powers of representation. These are natural structures, events or signs that somehow indicate how things stand elsewhere in the world. These are part of every organism's biological heritage. They have intrinsic intentionality. To Dretske they come ready-made, with functions defining what it is they are supposed to indicate about the world. (Type I and II indicators derive their representational powers from systems which already have type III indicators, and thus already have the full range of intentional states.)

Dretske also holds (p99) that type III indicators get their representational powers through their use as "movement switches" by which movement is initiated because of what these structures indicate about environmental conditions. "Only by *using* an indicator in the production of movements whose successful outcome depends on *what is being indicated*" can we have a principled, non arbitrary way of saying what the indicator has the function of indicating. (p70).

my olfactory perceptual system's outputs, as a constituent of my **coffee** representation, encodes information about the presence of coffee. Because it has developed so that it can be used to help me find and interact with coffee, and because it has developed by being used in this way, we can cay that this pattern of activation indicates the presence of coffee.

This indicates relation is important. The semantic relation the whole representation has with some piece of the environment is based upon what the representation's constituents indicate. This "larger" relation is the relation of representation. I'll be getting to the semantic relationship of representation in a moment, but first I want to concentrate on this "smaller" indicates relation a bit longer. I want to break this indicates relation down into a list of points, to detail precisely what it is, and how it comes to be a semantic relation between constituents of representations and properties of objects.

- Semantic relations (like meaning, aboutness, represents, indicates) which hold between an entity and some aspect of the world come from the way this entity is used.
- Representation **R** has evolved, over time, to have parts useful for successful Φing with things which can be Φed with.
- These things which can be Φed with turned out to be Xs.
- For example, x is the constituent of **R** which encodes information picked up visually about the presence and location of Xs. x has evolved as the pattern of activation it is, with the connections it has to other patterns of activation, so that it can be used to help find, identify, and Φ with things which can be Φed with.
- x has come to encode information about the presence and location of things which can be Φed with through the agent's history of attempting to Φ, and needing some way of picking up information about things which can be Φed with.
- The way x is encoded, the connections it has with other patterns of activation, and the way these patterns of activation encode what they encode, have all developed through the agent's attempts to Φ more successfully with things which can be Φed with.
- The use of the information encoded by x is implemented in the relationships (weighted connections) between the patterns of activation which encode this information and the patterns of activation (also parts of **R**) which encode the actions involved in Φing with things which can be Φed with.
- x stands in a genuinely semantic relation with the aspect of things which can be Φed with, i.e. Xs which it encodes information about

because of the way x is used, and the way it has developed by being used.

- I've decided to call this semantic relation "indicates". Thus x indicates the presence and/or location of things which can be Φed with (Xs).

So representations are made up of parts or constituents, and each constituent indicates some aspect of the perceiver's environment. Now I want to begin looking at the semantic relations the representation as a whole stands in. I'm going to explain what a representation represents in terms of what is indicated by all the representation's constituents taken together.

5.3 What does a representation (correctly) represent?

To be honest, I'm answering this question under duress. I said earlier that Wittgenstein advised us that the meaning of a word is found in its use in the language. That wasn't quite right. What Wittgenstein meant, is that words don't really have meaning, they have uses in a language, and if anything has "meaning" then it's the use of a word in a language. So Wittgenstein really counsels us that if you *insist* on talking about the meaning of a word, then rather than looking for the word's referent, a less pernicious place to look is at the word's use in the language. In a similar way, I'm reluctant to talk about what a PATTERN OF ACTIVATION represents. If anything does actually stand in the represents relation with an object it's the *use* of a PATTERN OF ACTIVATION rather that the PATTERN OF ACTIVATION simpliciter. But if you *insist* that a PATTERN OF ACTIVATION does actually represent, and want to specify what such a "representation" represents, then rather than looking for what the representation stands for, a less pernicious place to look is at the way the representation is used in interactions with objects.[3]

To be blunt, what a representation represents is given by the aggregation of all the indicates relationships of the representation's activated constituents.[4] These indicates relations, remember, have developed to be

---

3 The fact the actions are also constituents of PATTERNS OF ACTIVATION is another pointer that we're worrying overly about what a PATTERN OF ACTIVATION stands for, when what's important is how it is used. For instance, the actions of moving three feet to the right could be a constituent of the "representation" I use to guide my leaning. But the action of moving three feet to the right isn't a part of my wall representation.

Encoded actions don't indicate anything at all. Rather than standing in any semantic relationships themselves, encoded actions are part of implementing the "hard won rules of thumb", or habits in the use of the other constituents, and thus help determine how the representation is used, which (if you insist on there being semantic relations here) is where the semantic relations must come from,.

4 There is, of course, nothing new in the idea that one entity gets its semantic relationship through the aggregation of all the individual semantic relationships its parts have. The early Wittgenstein famously wrote that a sentence's semantic relations with a fact rides piggyback on the aggregation of all the semantic relationships which the individual parts of the sentence have. A sentence has a semantic relationship with a fact through the semantic relationships which its nouns have with objects and which its verbs have with actions. A similar relationship exists between a representation and an object, through the semantic relationships of the representation's constituents.

used in a certain way, through being used in that way. A representation is a PATTERN OF ACTIVATION over a large network of nodes, which is constituted by a group of patterns of activation, which each encode information about the properties of objects or about affordances, or initiate actions. Each of these constituent patterns of activation stands in a semantic relationship with a property of an object or an affordance provided by an object: they indicate this property or affordance. Thus some constituents of the representation indicate perceptual aspects of an object such as how it looks, feels, smells, sounds, or tastes. Other constituents indicate properties of the object, such as being spherical, being made of leather, being raw, or being bought for $20 from a garage sale. Still other constituents indicate the various affordances the object provides, such as that it affords sitting on, throwing, levering, reading, talking long distance to other people, or friendship.

Because of all the indicates relations a PATTERN OF ACTIVATION's constituent patterns of activation have, the PATTERN OF ACTIVATION can be a representation. Such a representation veridically represents an object which provides all the affordances and which has all the properties indicated by the representation's constituents. It stands in this represents relation because the representation is *used* to find and interact with objects which have all the properties and affordances indicated by its constituents.

For example a PATTERN OF ACTIVATION whose constituents indicate that the object is yellow with a greenish tinge and a few brown spots, looks spherical, feels spherical and bumpy, tastes sweet and acidic, and affords throwing and enlivening the taste of pesto, is a representation which *represents* an object to which this description applies. For any PATTERN OF ACTIVATION we can assess exactly what it represents by taking the aggregation of all the properties and affordances indicated by the individual patterns of activation which constitute this overall PATTERN OF ACTIVATION.

5.4 What about misrepresentation then?

So far it looks like the sort of object correctly represented by a representation is the object which has all the properties and affordances indicated by the representation's constituents. Considering misrepresentation is going to make me put a few more complicating details in before I can say this properly.

(The fact that the later Wittgenstein rejected this view that words and sentences have semantic properties could perhaps be taken as another pointer to the fact that the search for what a representation stands for is misguided.)

Cases of misrepresentation are useful in that they provoke us to look closer at this represents relation. They illustrate the idea that an account of representation must involve something other than the representation and the physical object which caused the perceptual experience. Let me recap why there has to be something else at work here.

The semantic relations a representation's constituents have are set in place through the representation's history of being used in the production of actions when interacting with objects of a certain sort. What a pattern of activation encodes information about is due to the pattern of activation's history, not to what is presently being perceived. A pattern of activation encodes information about the sort of object the pattern of activation has been used to interact with, and it indicates that this sort of object is being perceived.

Cases of misrepresentation are often cases where the information trransduced is either not picked up or is not transduced very well. In Chapter One I gave an example of genuine misrepresentation: that of my perceiving my grey jersey on a chair when I didn't have my glasses on and thinking that the object on the chair was my cat Madison. When light patterns which encoded information about my grey jersey were trandsuced by my visual pattern recogniser network, the output pattern they produced encoded information about Madison. The visual information I picked up was information about the object on the chair, but the pattern recognisers which processed the outputs of my visual perceptual system activated a pattern of activation which best fitted this (poor quality) information. They activated a pattern of activation which encodes information about Madison. (It encodes this information because of the way it developed through its history; through using it to interact with Madison.) This pattern of activation is a constituent of my **Madison** representation, so this activated the rest of that representation. Now my **Madison** representation has constituents which indicate a specific set of properties, affordances and so on: they indicate a grey tabby cat with a big appetite, who loves chocolate and cheese, who sheds hairs all over the furniture, who likes having her tummy rubbed, and who nags for food whenever I go into the kitchen.

If the previous section told the whole story, then my **Madison** representation would presumably represent an object with all these properties. But in this case there isn't such an object, and the object which caused my perceptual experience, the object which is there on the chair has a quite different set of properties. So something's not quite complete in the story I told in the previous section. This is why we have to look at misrepresentation carefully.

Thinking of a representation as standing in a different semantic relationship –other than encodes information about, indicates, or represents– is going to help us look closely at cases of misrepresentation, and to explain what the represents relation looks like. I got hung up on trying to base the represents relation on the indicates relationship. It's going to be much easier to base the represents relationship on a semantic relationship which I'll call "expresses".

This *expresses* relationship is in turn based on the indicates relationship. A representation expresses, rather than represents, an object with all the properties and affordances indicated by all the representation's constituents. So in all cases where a representation with exactly the same constituents is activated, the same object will be *expressed;* so the same object will be expressed whether or not any physical object with those properties is present and being perceived. This means that the object expressed cannot be the physical object perceived, because in cases of misrepresentation this object has different properties from those indicated.

The object expressed by the representation is going to be an abstract object then. This abstract object is an object with the properties, affordances and so on indicated by the representation's constituents, and is an object which has these properties and affordances whether or not it exists. For instance, when my **Madison** representation is activated, this representation expresses an *abstract* grey tabby cat with a big appetite, who loves chocolate and cheese, who sheds hairs all over the furniture, who likes having her tummy rubbed, and who nags for food whenever I go into the kitchen. This object is expressed whether or not the object I'm perceiving actually is Madison.

(Some people object to the idea of abstract objects figuring in an account of perceptual representation. But as I'll show in the next chapter, we *need* to see a representation as expressing an abstract object, if we're going to be able to explain exactly what the *representation* relation consists of, what it does, and what happens to it in a case of misrepresentation.)

To provide an account of representation which also explains misrepresentation, the tactic I applauded in Chapter One, was is to specify in a non *ad hoc.*, non-circular way how a representation comes to correctly represent certain objects but not others. If we can do this, then we can identify cases of genuine misrepresentation as those cases where this representation is activated because an object caused incomplete or imperfect perceptual information, and where this object is one which the representation doesn't correctly represent. By starting from scratch, looking at the perceptual and representational powers new-born infants have and

how their representations' semantic relationships develop by the representations being used to interact with a certain sort of object, I am providing an account of the way a representation comes to stand in certain semantic relationships in a principled, non-circular way. I've explained the encodes information about relation in this way, and then based the indicates relation on that relation. Now I've outlined the expresses relation and how it is simply the aggregate of all the indicates relationships. Now I need to explain the semantic relationship we're principally interested in –the represents relation– and how it is based on the expresses relation. This is the job I take on in the next chapter.

CHAPTER SIX

THE REPRESENTATION RELATION

6.1 *"Are we nearly there yet Dad?" – Where we've got to so far, and a little white lie about how we haven't far to go.*

One of the major concepts to come out of the discussion so far, is the idea of connectionist representation systems. I've brought the concept of patterns of activation to light, showing how they encode the information picked up by a cognitive organism's perceptual systems. I've illustrated how they can form large PATTERNS OF ACTIVATION made up of smaller constituent patterns of activation. I've also revealed how each of these constituent patterns of activation has the function of indicating certain properties and affordances. A PATTERN OF ACTIVATION made up of such constituents expresses an abstract object: an object which has all the properties and affordances indicated by the PATTERN OF ACTIVATION's constituents (and which has these whether or not it exists).

The last major step to cover then, is to address the question "If a PATTERN OF ACTIVATION is a representation (which is disputable), exactly what does a PATTERN OF ACTIVATION represent?" I'll refer to this from now on as the Question.

That all we have left to do is to answer this Question is the "little white lie." The truth (don't tell the kids) is that this isn't as easy a question to answer as it might seem. In this chapter I'll be working towards an answer to the Question which can account for what a PATTERN OF ACTIVATION represents in what I take to be the three crucial situations:

(a) veridical representation,
(b) non-veridical representation, or misrepresentation, and
(c) non-perceptual representation, when a PATTERN OF ACTIVATION is activated by my thinking about an object rather than perceiving one.

I can see four points of view on the answer to the above Question of what a representation represents, each of which tells a different story about which object is represented in these three situations. I'll deal with these four views in sequence, explaining how each deals with what is represented in non-veridical and veridical representation, and also how each deals with situations where I'm just thinking about an object rather than perceiving one. I'll also outline the

difficulties with each of these viewpoints, showing how the next account either overcomes or avoids these difficulties. With the last account, we'll have a viewpoint which I believe provides a satisfactory answer to the Question of what a PATTERN OF ACTIVATION represents.

One hurdle we'll have to pass over along the way is the relation between the abstract object expressed by the PATTERN OF ACTIVATION and the physical object which caused the perceptual state which activated my representation. This relationship is still to be explained. The diagram so far could be that of Figure 6.1.

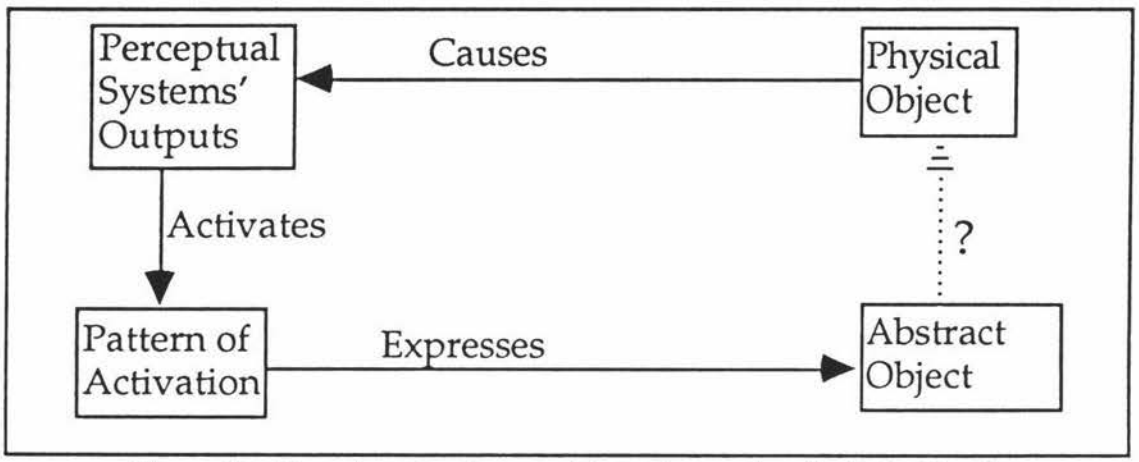


Figure 6.1 How does the abstract object relate to the physical object?[1]

This relation will need to be laid out somewhere along the way, as the relationship between these two entities is going to be integral to finding an answer to the Question of what a PATTERN OF ACTIVATION represents. So to arrive at an explanation of representation, I'll spend some time explaining how the physical object and the abstract object of Figure 6.1 relate to each other.

---

1 Standing back from this I get worried. Since I've been talking all along about information being encoded, one might be tempted to say that the relation on the top line should be "encodes information about." the problem is that the information encoded in the environmental energy and the information encoded by the perceptual systems' outputs can be different. For example, when I perceived my jersey and activated a constituent of my **Madison** representation. The information encoded by the perceptual systems' outputs was information about Madison, not about the jersey which I perceived. So here I want to get away from talking aboyuut encoding information and get back to talking about causes (everyone else talks about them).

I say the physical object perceived"causes" the perceptual organs to activatre a representation because I want to avoid any talk of patterns of activation encoding information, because the information they encode in cases of misrepresentation wouldn't be information about the physical object *presently* being perceived. This is because the semantic relations a representation's constituents have are set in place through those constituents' history of being used in interactions with objects of a certain sort. And it's *that* sort of object which they encode information about. Hence what the representation's constituents encode information about is already set in a sense, by their history of use. Thus in cases of misrepresentation the information encoded by the outputs of the pattern recognisers is about a different object to the one which is presently perceived. This is why the encodes relation doesn't feature on this diagram of what is presently happening.

One of the aims of this chapter is to flesh out the encodes information relation in terms of the other three semantic relations: indicates, expresses and represents. Thus the physical object causes the perceptual systems to encode information which activates a representation. What the representation's constituents indicate is given by the abstract object expressed. How any semantic relations point back to this physical object (and, in fact, whether they do point back to this physical object) is the topic of this chapter.

I made a big fuss about misrepresentation back in Chapter One. How an account of representation handles misrepresentation is an important yardstick with which we can measure its adequacy as a theory of representation. In spite of the fact that many of the accounts I'm about to illustrate in this chapter fail to provide an adequate account of misrepresentation I still want to work through each of them. I'd prefer to work slowly towards the final version, trying each account out, figuring out it's shortcomings and then progressing to a better view, rather than leaping straight for the final version. This way we can see the advantages and disadvantages of each version, and then discard each for an improved version which doesn't suffer from these disadvantages. If we go too quickly, we might miss some vital piece of the puzzle along the way. Another reason for progressing this way is that this methodical way of doing things is, in some sense, autobiographical; it illustrates the succession of ideas I've run through in trying to uncover a solution to this puzzle. In the following sections of this chapter I'll work slowly through this succession of viewpoints towards the version I currently hold in favour, which I believe does provide an adequate account of (a) representation, (b) misrepresentation and (c) non-perceptual representation.

This chapter's answers to the Question of what a representation represents take inspiration from the work of Brentano, Meinong, and Frege, whose accounts of the meaning of linguistic expressions gave abstract objects a respectable place in explanations of meaning in language. In fact they showed that abstract objects must be a part of the story if we're to get a coherent account of linguistic expressions' meanings. I aim to show that the same applies for any account of the aboutness or intentionality of mental representations: abstract objects are a respectable addition to a story of representation and misrepresentation, and we're not going to get very far without them in the picture. The way that the different accounts I'm going to run through differ, is *how* abstract objects fit into the picture.

6.2 A "Brentano-ish" theory of representation.

One way of attempting to explain what is represented by a representation, and how a representation can misrepresent is to make a distinction between veridical perceptions and non-veridical ones. That is, we make a sharp distinction between situations in which the object which caused the representation-activating perceptual experience is the sort of thing the representation should represent, and other situations in which this object is not

the sort of thing the representation should represent. ("Should" here is based on the representation's content.)

We can say that a representation represents the object which caused my perceptual experience, but it represents that object only in veridical cases; that is, only in cases where this object *is* the sort of object the representation's content correctly applies to. In non-veridical cases the representation instead represents the abstract object. In such cases the abstract object is the sort of thing the representation should represent.[2]

So when my **cat** representation is activated by my seeing a cat, the representation does "fit" the object I'm perceiving; this object has all the properties indicated by the representation's constituents. Thus **cat** represents the cat which caused my perception.

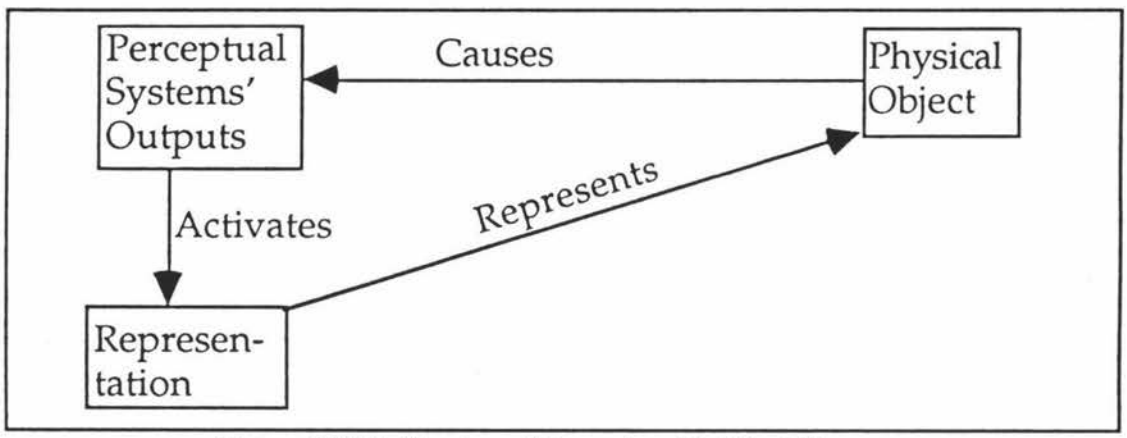


Figure 6.2 (a) Brentano-ish version: Veridical Cases

With cases of misrepresentation the situation is quite different. An example of misrepresentation (of type (ii), which does qualify as a case of genuine misrepresentation–see Chapter One): suppose I see a small furry animal up a tree while out walking on a dark evening, and this causes my perceptual systems to be in a state which causes the activation of my **cat** representation. This PATTERN OF ACTIVATION expresses an abstract object which has properties like being an animal which stands on four legs, is about calf-high

2 I call this a "Brentano-ish" version, because this is the sort of view Brentano held, or at least he might be interpreted as holding it (before about 1904, when he seemed to change his mind). On this interpretation, when I'm thinking about an A, this A I'm thinking about can be an actual A, existing in reality, or it can be a contemplated A, which can continue to be thought about even after all actual A's cease to exist.(Brentano, 1930: p27). Thus on this interpretation of the early Brentano, a thought can be about either an existent physical object or about an intentional object.

Here, to aim for better than my usual C^+ in historical accuracy I should add: Brentano did not think of Diogenes' thinking about an honest man (assuming there is no such man) as entailing that Diogenes is related to an object which doesn't exist. (Brentano explicitly rejected the idea that there are "inexistent objects".) Rather this relation is a psychological relation which is unlike any purely physical relation; a relation between Diogenes and an intentional object, to which no existent physical object corresponds. See Chisholm (1957) , Chapter Eleven, "Intentional Inexistence", especially pp 169-70.

to me, which feels soft and furry, which I'd like to pat, which usually likes being patted, which makes a purring noise when I pat it, and is the sort of thing which often comes to "Here Kitty" (uttered in a friendly voice).

If a possum-on-a-dark-night caused the perception which activated this representation, we would like to say that this is a non-veridical perception in which the representation has been activated inappropriately (because the sensory information isn't good quality as a result of the lack of light). In such non-veridical cases the cause of the representation's activation isn't the sort of thing we intuitively feel that the representation should represent. We'd like to say that **cat** must only represent cats. My **cat** representation shouldn't represent possums-on-a-dark-night. In such a situation, if someone asked me what I believe is up the tree I'd reply that it's a cat. I'd say this because the abstract object expressed by my representation is an abstract cat: an animal which stands on four legs, is about calf-high to me, which feels soft and furry, which I'd like to pat, which usually likes being patted and which makes a purring noise when I pat it. Saying that this representation represents a possum-on-a-dark-night, which has hardly any of these properties, seems to run counter to the content of the representation. When **cat** is activated by a possum we would like to say that this is a case of misrepresentation, rather than maintaining that **cat** can somehow correctly represent possums-on-a-dark-night.

We can do this by making a distinction between veridical cases and non-veridical cases. Non-veridical cases are cases where the properties of the physical object which caused the perceptual systems' outputs which activated the representation are different to the properties indicated by the representations constituents. If we wish to maintain that a representation represents only things which have all the properties individually indicated by each of the representation's constituents then the representation must represent something other than that physical object which caused the representation's activation. My **cat** representation must represent something other than the possum, because **cat**'s constituents indicate properties quite different from those possessed by the possum.

The best candidate for the "something else" represented in such a non-veridical case is the abstract object expressed by the representation, which has all the properties individually indicated by each of the representation's constituents. So where the physical object has properties different from those expressed by the representation's constituents, the representation represents the abstract object it expresses. Thus in non-veridical cases **cat** represents this abstract cat rather than the possum. By making a distinction between veridical

cases and non-veridical ones, we can maintain that my **cat** representation always represents cats; either physical ones in veridical cases or abstract ones in non-veridical cases.

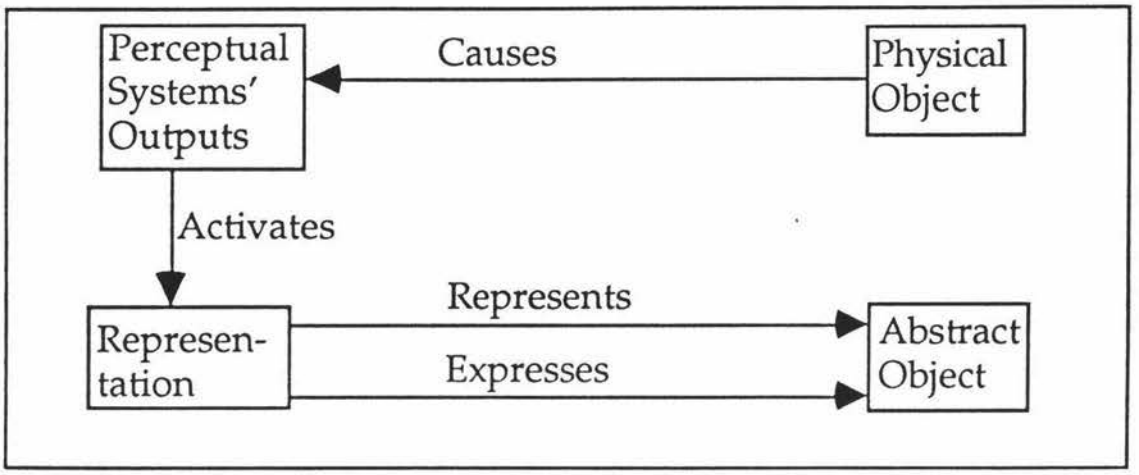


Figure 6.2 (b) Brentano-ish version: Non-Veridical Cases

As well as accounting for what's represented in both veridical and non-veridical cases of perception, a theory of representation should also able to account for non-perceptual representation. By "non-perceptual" representation I mean cases where my representation is activated by means other than the current state of my perceptual systems' outputs. For example, sometimes my **lasagne** representation is activated because I'm thinking about what to cook for dinner, rather than by my being confronted perceptually by a piece of lasagne. Here I'm thinking about the lasagne I'm going to cook for dinner rather than perceiving anything which activates this representation. The activation of this representation wasn't due to any perceptual experience. It was caused by activity in non-perceptual areas of my brain rather than by any physical object. A semantic theory should explain what lasagne represents in such a case.

The obvious answer here is that **lasagne** represents the abstract object it expresses, an object with all the properties indicated by each of my lasagne representation's constituents. My **lasagne** representation's constituents indicate things like being made from pasta, cheese, herbs, tomatoes and vegetables, tasting delicious, and being my flatmate's favourite food. The abstract lasagne expressed has all these properties. Also, the lasagne I'm thinking about doesn't exist, and it may not ever come into existence if I change my mind and cook spaghetti instead. So the fact that the abstract lasagne expressed has all these properties, and it has these properties whether or not it exists is another advantage. In cases where the representation is activated by non-perceptual

activity, the representation represents this abstract object; it's this abstract object which I'm thinking about.

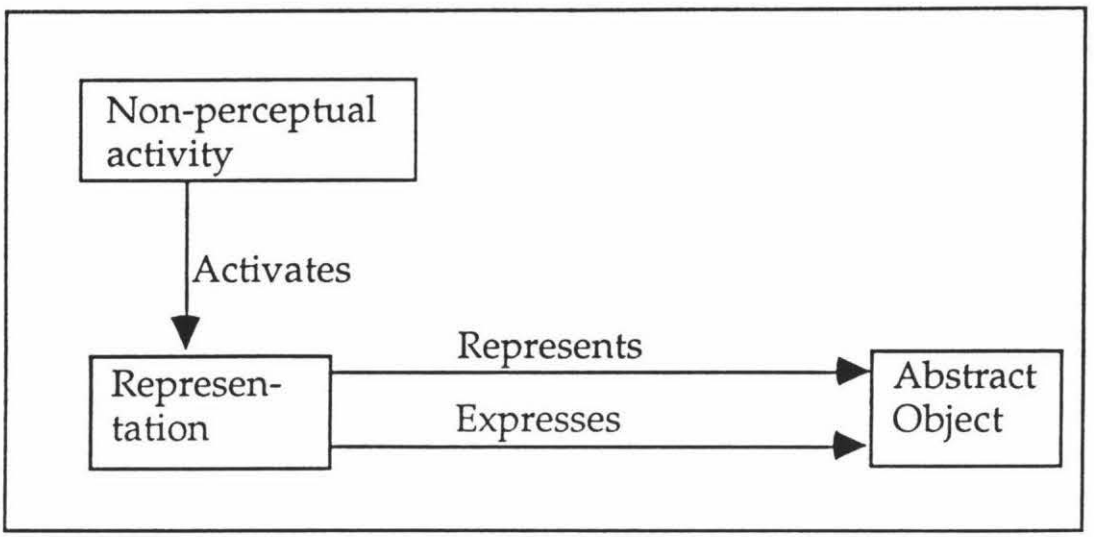


Figure 6.2 (c) Brentano-ish version: Non-Perceptual Cases

So on this view, when a representation is activated in (a) veridical cases, it represents the physical object which caused the perception, and in (b) non-veridical and (c) non-perceptual cases it represents the abstract object expressed by the PATTERN OF ACTIVATION.

Don Bradshaw put this sort of view forward (not entirely seriously, it seems), as a view which might solve Fodor and Pylyshyn's problem of lack of misrepresentation. To Fodor and Pylyshyn[3] it seems that the object represented must be the object which plays the central causal role in activating the representation. Thus their problem is that in misperception an abstract object can't be represented, because abstract objects can't cause perceptions. Bradshaw points out that accepting non-actual, or "merely intentional" objects of misperception isn't wholly in conflict with Fodor and Pylyshyn's views. He suggests that they could maintain that:

> "Perception is a process that is mediated causally. In veridical perception the object that plays the central causal role is the intentional object. In misperception this is not the case."[4]

That is, in non-veridical cases the object represented is something other than the object which caused the representation's activation. (Fodor, of course wouldn't agree with this idea.)

---

3 Fodor and Pylyshyn (1981)
4 Bradshaw (1991) : p432.

For several reasons I'm not happy with this "Brentano-ish" view at all. The idea that what is represented "flip flops" back and forth between the physical object which caused the representation's activation and the abstract object seems a very *ad hoc.* solution. We wanted to justify our describing certain cases as "cases of misrepresentation." To do this we decided that for each representation there is a certain sort of thing which that representation should represent, and then stipulated that in cases where the physical object which is perceived is different from the object the representation "should" represent, an abstract object is represented. Thus whether my **cat** representation represents an existent physical cat or not changes depending on whether the perception is what we would term "veridical" or not.

It seems that if we want a non-ad hoc solution, it should really be the other way around. We should be able to specify what the representation represents in some principled way , and then on the basis of whether or not the object represented is the "right" sort of thing, we decide whether or not we should call the representation "veridical." The representation's being veridical or not should depend on whether or not the representation represents the physical object being perceived, rather than the other way around, as this "Brentano-ish" theory would have it.

This solution seems especially *ad hoc.* when we note that I can't tell whether my perception is veridical or not, and hence what object is represented. Only from a "God's eye view" is it possible to know whether a perception is veridical or non-veridical. All we mortals have access to is the outputs of the perceptual systems, the representation activated and the abstract object expressed by the representation. We don't really have access to the true identity of the physical object which caused the representation to be activated, so we have no ability to make the distinction between veridical and non-veridical perceptions on which this solution is based.

6.3 A "Meinongian" view.

What we really want is a principled, non-ad hoc way of specifying what is represented, rather than specifying whether or not a perception is veridical, and then stipulating what is represented on this basis. And since all we have access to are the outputs of the perceptual systems, the representation activated and the abstract object expressed by the representation, it would be nice if we could find a way of accounting for what is represented in all three sorts of representational cases with an apparatus which only involved these three parts

of the diagram in explaining what is represented. It would also be an advantage if what is represented could be explained using only one diagram instead of three. This is just the sort of account a "Meinongian" view provides. On this view what is represented is not at all ad hoc. What is represented is always the abstract object. In all cases of perceptual and non-perceptual representation, the representation represents the abstract object expressed.

So even in veridical cases the representation doesn't represent the physical object which caused the perceptual systems' outputs, instead it represents the abstract object expressed by the representation. Since a representation always represents the abstract object expressed, whether a perceptual representation is veridical or not is given, not by what is represented, but by the relationship between the abstract object represented and the physical object which caused the perception. This relationship is important here, so now is a good time to examine what this relationship is all about.

One of the purposes of activating a representation is to identify the object I've encountered. Because my **cat** representation expresses an abstract cat, when this representation is activated I identify the object I've encountered as a cat. So I interact with the small furry animal up the tree in accordance with my belief that it's a cat: I try to coax it down from the tree by calling "Here Kitty," in a friendly voice.

Ideally the physical object which is the cause of my perception has been correctly identified, and has all the properties which the object I believe I've encountered has. That is, ideally the physical object has all the properties possessed by the abstract object; if this is so, then my interactions will usually have a good chance of being successful. In general, being able to correctly identify objects in my environment, and thus knowing all the (relevant) properties of the physical objects I'm dealing with makes it much easier to plan and carry out actions which will serve my needs and desires. If the abstract object has properties which the physical object does not have then my interactions are possibly going to be unsuccessful. For example, if the small furry animal up the tree is actually a possum then the physical object which caused my perceptual experience doesn't have properties of enjoying being patted, purring, often being friendly, and being the sort of thing which often comes to "Here kitty" (uttered in a friendly voice). In such a situation my attempts to coax the "cat" down from the tree by calling "Here kitty" would be unsuccessful, or even disastrous.

It can be seen that the abstract object and the physical object which caused the perception having properties in common is important to the

appropriateness of my beliefs, and thus to the success of my interactions with my environment. The relationship between abstract object and physical object is tied up with their sharing properties. It's the fact that the physical object perceived has all the properties the abstract object has which makes my perception veridical and thereby makes my actions more likely to be successful.

So we can specify on the diagram the relation between the abstract object and a physical object perceived as one of correspondence. The properties of the abstract object correspond with those of the physical object.

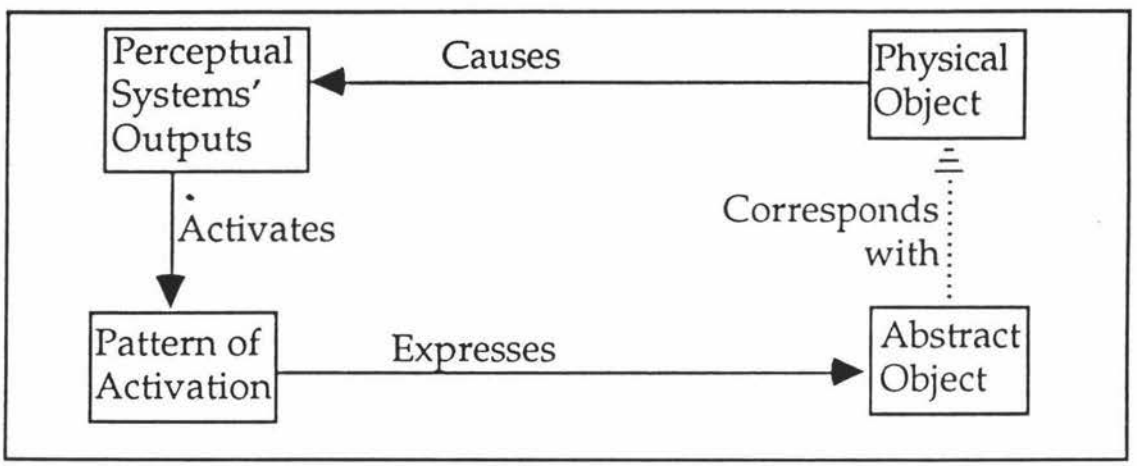


Figure 6.3 The Relation between the Abstract object and the Physical object

Note that the corresponds relation is a one-way relation; the abstract object corresponds with the physical object but not vice versa. The abstract object corresponds with the physical object in the sense that the abstract object adembrates, or partially describes, the physical object. The physical object will probably have more properties than just those expressed by the abstract object, because the abstract object may be indeterminate about some properties. In a veridical case the abstract object corresponds with the physical object perfectly; it's a perfect, but partial, description of the physical object. The physical object has at least all the properties the abstract object has, but it will also have other properties not possessed by the abstract object. Imagine that the small furry animal I see up the tree really is a cat. My cat representation is activated, a representation which expresses an abstract cat. The physical cat has all the properties possessed by the abstract cat expressed by the PATTERN OF ACTIVATION. The physical cat and the abstract cat are both animals correctly described by "stands on four legs, is about calf-high to me, feels soft and furry, is an animal which I'd like to pat, usually likes being patted, usually makes a purring noise when I pat it, and is the sort of thing which often comes to "Here Kitty" (uttered in a friendly voice). The point is that the physical cat also has properties which

the abstract cat doesn't have: properties like being an animal with exactly 178 965 hairs on its body, being an animal who lives with a woman named Diedre, and being an animal which had chicken giblets for dinner. Thus the physical cat has all those properties possessed by the abstract cat my PATTERN OF ACTIVATION expresses, but the abstract cat is indeterminate about some properties, and thus does not have all the properties possessed by that physical object. So the abstract object corresponds with the physical object, but the converse isn't the case.

If we take the "Meinongian" view that what is represented is always the abstract object expressed by the representation, then what is represented doesn't depend on whether we would judge the representation to be veridical or not (as it should be, if we're to have a non-ad hoc. solution). Instead whether or not a representation is veridical is given by this correspondence between the abstract object and the physical object. When there is a perfect correspondence between the properties of the abstract object represented and the properties of the physical object which caused the perception, we have what is traditionally called a veridical case of perception. In such a situation this physical object has all the properties possessed by the abstract object. In a non-veridical case the abstract object represented does not correspond perfectly; the abstract object has some properties which give it a character the physical object doesn't have.

One virtue of this story is that this correspondence isn't an all or nothing affair; there can be a whole range of levels of correspondence. For example, a three-legged cat (seen from an angle where I don't notice this anomaly and thus activate my **cat** representation rather than a **three-legged-cat** representation) will correspond almost-but-not-quite perfectly with the abstract object expressed by my **cat** representation. The abstract object will have the property of being four-legged, where the physical object doesn't have this property. But apart from having a different number of legs, practically all the other properties of the abstract cat will be possessed by this three-legged cat. This perception is non-veridical, but only very slightly so. If the physical object which caused my perception is a possum, then the perception is even more non-veridical. The physical object is still an animal which stands on four legs, is about calf-high to me, and which (probably, though I've never touched one) feels soft and furry, but it's not the sort of thing I'd like to pat, and it doesn't make a purring noise when patted; it also wouldn't come to me if I call "Here kitty", even if uttered in my friendliest voice. Here the correspondence between the abstract object and the physical object is much less than perfect.

This means that the veridicality of perception is also not an all-or-nothing dichotomy. Perceptions can be veridical, just slightly non-veridical, or

seriously non-veridical, depending on the level of correspondence between the physical object and the abstract object. Every case where this correspondence is less than perfect is a non-veridical case, but there is a spectrum of such less-than-perfect levels of correspondence. The perfect correspondence of a veridical perception is at the extreme pole of the spectrum, with non-veridical perception running the full range of less-than-perfect correspondence.

At the opposite pole of the spectrum of perfect correspondence with a physical object, are the abstract objects expressed by representations activated by activity in non-perceptual areas of my brain. With such non-perceptual representation there is *no* physical object involved, so the correspondence between the abstract object and any physical object is totally absent.

Because the same thing is represented in veridical, non-veridical and non-perceptual cases of representation, we can draw a single diagram which covers all three situations. Figure 6.4 illustrates what is happening in all cases where a PATTERN OF ACTIVATION is activated.

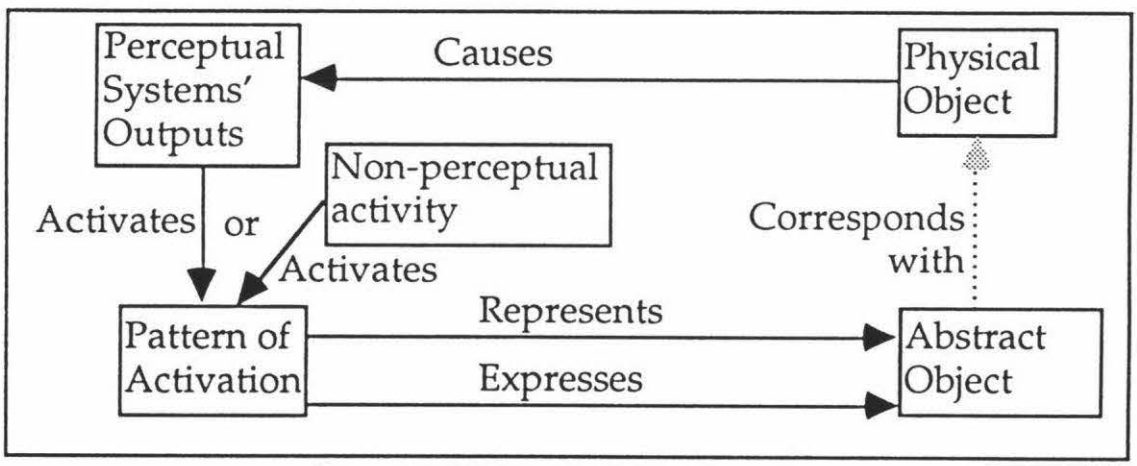


Figure 6.4 "Meinongian" theory's version

In all cases, the abstract object expressed is the object represented. In (a) veridical cases, where I correctly identify the object encountered, the object represented is the abstract object, and this abstract object corresponds perfectly with the physical object which caused my perceptual state. In (b) cases of non-veridical perception, the abstract object is also represented, but the correspondence between the abstract object and the physical object is not perfect. This correspondence may be nearly perfect, moderate or even quite poor. In (c) cases where the representation was activated by non-perceptual factors, by my thinking about associated things, for instance, the representation activated still represents the abstract object the representation expresses, but there is no

correspondence at all between this abstract object and any causally active physical object.

This account has the advantage that it gives an explanation of what is represented in each of the three situations we need to explain, and gives this account using a single diagram. Also we have a system which is not at all *ad hoc*. in stipulating what is represented in each case. What is represented is always the same thing: the abstract object.

But this account does have some drawbacks. Here, since the object represented has all the properties indicated by the representations constituents, the representation can never misrepresent in the sense of representing something which it shouldn't represent, something which has properties other than those indicated by the representation's constituents. If my **cat** representation is activated by a perceptual state caused by a possum-on-a-dark-night, this representation still represents a cat (albeit an abstract one). So even though the physical object which caused my perception shares relatively few properties with the abstract object represented, **cat** represents an abstract cat. The representation doesn't mis-represent, in the traditional sense of represent something which it shouldn't represent; my **cat** representation always represents cats, it never represents anything but cats. On this view of what the representation represents, **cat** certainly can't represent or, more to the point, misrepresent a possum-on-a-dark-night.

So if we accept this option, we're going to have to re-define what we mean by "misrepresent" and accept that we've been mis-using the language of representation. We can't say that I misrepresent the possum as something which it isn't, as a cat. What we must really mean by "misrepresent" is that I represent a cat when I should represent a possum. Misrepresentation happens when the properties of the abstract object represented don't correspond to the properties of the physical object which caused the perceptual state. When my **cat** representation is activated by a possum, and thus the abstract cat represented has properties which don't correspond with the properties of the possum, then this is a case of misrepresentation.

We could perhaps be happy to say that what we mean by "misrepresentation" is that the abstract object represented has properties different to those possessed by the physical object, but there is a problem: this doesn't sit well with our ordinary ways of *describing* cases of misrepresentation. With this "Meinongian" view, we would say that I just misrepresent, period; I represent a cat when I should represent a possum, so this is a case of misrepresentation. But this is not how I would normally describe the situation.

I would normally describe the situation above by saying that I misrepresent the possum as a cat. I interact with the object up the tree, the possum, in the belief that it's a cat. That is, I represent the possum, by misrepresenting it as a cat.

It seems that we should be able to account for the fact that when I misrepresent, I misrepresent the physical object perceived, and I misrepresent it by representing it as something which it is not. This "individual directedness" element, the representation relation being directed towards a physical object, so that I represent or misrepresent that object, is missing from this account, and is dealt with by the next version.

6.4 A "Fregean" View.

This way of answering the Question of what a representation represents comes from comparing the intentionality of mental representations with Frege's treatment of the meaning of referring expressions. Frege held that referring expressions (like names or definite descriptions) in general have a *sense* and a *reference*.[5] The reference is the physical object the name is used to refer to, and the sense of a referring expression is the mode of presentation, or the way of determining, that reference.[6] The sense presents the reference in a certain light, or from a certain point of view. The reference of "The morning star" is the object this referring expression refers to, the chunk of rock and gases orbiting around the Sun. The sense of this referring expression is the way this expression presents that object; it presents this piece of matter as the morning star. "The morning star", "the planet Venus," "the evening star", "the brightest celestial object in the morning sky," and "the second planet from the Sun" all have the same physical object as a reference, but they each express different ways of presenting this same object–they express different senses. I'm of the opinion that the abstract object expressed by a representation presents the physical object which caused the perception in a very similar way.

Frege employed the distinction between sense and reference to explain how identity statements could be true and still be informative. The tradition at the time was that there is only one component to the meaning of a referring expression: the object the expression refers to. That is a referring expression means the thing the expression "stands for." So in "the morning star is bright", the object referred to has the property of being bright ascribed to it. If there was

5 Frege (1892/1966)
6 Frege (1892/1966) : p57.

only this component of meaning, then we'd have problems. Take the identity statement (1) for example:

(1) "The morning star = The evening star"

If meaning was given entirely by the object the referring expression stands for, then we could replace "the evening star" in (1) with any expression which stands for the same object, and the resulting expression would not change its truth-value. This is known as the principle of intersubstitutivity of co-referential expressions; co-referring expressions can be substituted for each other in statements without changing the statements' truth-values. "The morning star" is an expression which stands for the same object as "the evening star", so we could apply this principle and substitute this for "the evening star" in (1) to get (2):

(2) "The morning star = The morning star"

Now although the *truth* of (1) is preserved in (2), (2) seems to be much less *informative* than (1). Frege shows that (2) actually means something quite different from (1). To show how the meaning of (1) is changed in (2), and thus to also show how identity statements like (1) can be true and informative (in the sense that it says something more than X is self-identical), Frege presented the idea that the meaning of a referring expression like "The morning star" has two components. As well as standing for a *reference*, it also expresses a *sense*. This sense is the way the expression presents this object.

By dividing meaning up into sense and reference, we can see that (1) is a true and informative statement. The two referring expressions in (1) express different senses. The identity statement is informative in that it states that these two ways of presenting a physical object, these different senses, present the same object. The object presented by "The morning star" is the same object as that presented by "The evening star". This is much more informative that (2) which merely proclaims self-identity of "The morning star"; a trivial claim, since everything is identical to itself.

So when applying the principle of the intersubstitutivity of co-referential expressions, although the truth value of the identity statement isn't changed, its meaning is changed because we substitute an expression for another with the same reference but a different sense. The substitution we made in going from (1) to (2) changed the meaning of (1), even though it didn't change its truth value, because there is more to meaning than just referring; an expression with a quite different sense was substituted.

As well as solving the problems of the informativeness of identity statements, this piece of apparatus also solved the apparent problem that names for non-existent objects would have to be meaningless. If a referring expression is taken to mean the physical object it refers to, then "Rudolph the red-nosed reindeer", since it does not refer to any such physical object, is would have to be meaningless. Frege used the distinction between sense and reference to show that expressions like "Rudolph the red-nosed reindeer" are not meaningless because even though they don't have any reference, they still express a sense. Frege held that every genuine referring expression expresses a sense, whether or not it has a reference.[7] This sense is what allows a referenceless referring expression to still have meaning. Everyone who knows the language can understand what I mean when I mention Rudolph, because we all grasp the sense of this referring expression. We can realise that I mean something different from the tooth fairy and from my new red ferarri, because "The Tooth Fairy", "my new red ferarri", and "Rudolph the red-nosed reindeer" each express quite different senses, even though none of these referring expressions pick out references.[8]

The similarity between Frege's sense and my abstract object is quite heartening. We can take a similar view with representations as Frege took with referring expressions. Frege showed that there is more to meaning than just referring, by separating out the two components of the meaning of referring expressions, the sense and reference. In a similar way, we can maintain here that there is more to aboutness, or intentionality, than just representing. We can separate out the two components of the aboutness of representations: the abstract object and the physical object. A referring expression refers to a physical object and expresses a sense, and in the same way a representation represents a physical object, and expresses an abstract object.

---

7 Frege (1892/1966) For example, p58: "The words 'the celestial body most distant from the Earth' have a sense but it is very doubtful if they also have a reference... In grasping a sense, one is not certainly assured of a reference."

8 An interesting aside: The following appears to be a problem for Frege's account: Frege would have to say that "Superman" and "Clark Kent" would express different senses, and neither has a reference. Because of this, they must mean different things. Thus against our intuitions, one of two things would be the case: it could be that the statement "Superman is (identical to) Clark Kent" would be false, even though Superman and Clark Kent are supposed to be the same person. However, I think (though I must admit to not being sure about this) that Frege maintains an alternative to this: that since this statement contains the referenceless expression "Superman" the statement is truth-valueless. Thus (again counter-intuitively) it has the same truth value as "Superman is president of the United States". (Maybe Frege did have a solution to this problem which I don't know about. Tnd this is a definite possibility, since I'm by no means an expert Frege scholar.)

(Pavel Tichy has a solution to this problem, however: "Rudolph the red nosed reindeer" refers to a role, or office, which happens to be vacant. And thus the statement "Superman is (identical to) Clark Kent" says that if there was anything which filled the role referred to by "Superman," then this object would also fill the role referred to by "Clark Kent". See Tichy's (1978) 'De Dicto and De Re' (Philosophia 8(October): pp 1-16) and also his (1987) 'Objects and Offices' (a translation of sections 1,2,3,8 of 'Einzeldinge als Amtsinhaber' from Zeitschrift für Semiotik, vol 9, (1987), pp 13-50).)

As I just mentioned, Frege held that the sense of a referring expression is the mode of presentation of a reference. The reference is referred to by the referring expression, and it is referred to via the sense expressed.[9] Similarly, we can say that a representation expresses an abstract object, which is a way of presenting a physical object. This physical object is represented by the representation, and it is represented via the abstract object expressed. The sense characterises the reference: meaning is about superimposing the referring expression's sense onto its reference. In a similar way the abstract object expressed by a representation characterises the object represented. Representation is about "superimposing" the abstract object expressed by the representation onto the object represented. This object is represented as being the sort of object characterised by the abstract object. We might diagram this situation like this:

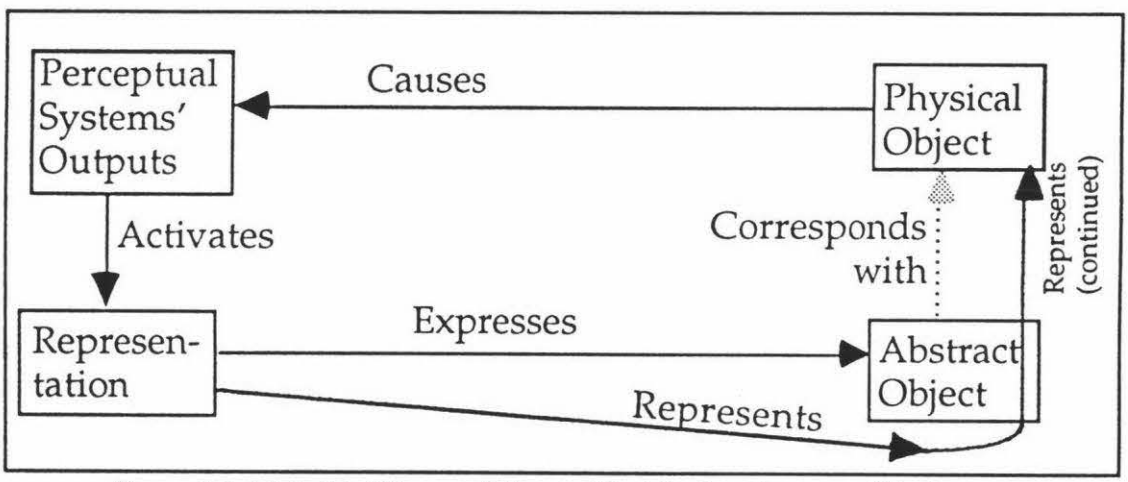


Figure 6.5 (a) & (b) "Fregean" theory: Veridical and non-veridical cases

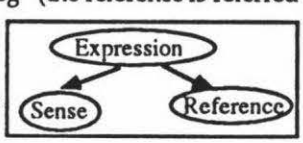
Here the "Represents" arrow points to the physical object, but it points to it *through* the abstract object that the representation expresses.

An advantage of this system is that it accounts for each of the two components Mohan Matthen rightly claims any theory of perceptual representation must address: the *individual-directedness* component, and the

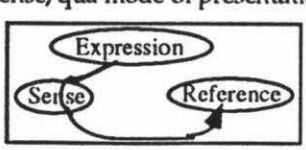
---

9 Frege actually says that "connected with a sign...[is] also what I should like to call the sense of the sign, wherin the mode of presentation is contained." (Frege (1892/1966) p57) So the sense *contains* the mode of presentation, rather than *being* the mode of presentation. I take this to imply that perhaps the sense does some other job as well as presenting the object.
Anyway, it's hard to tell whether Frege would diagram the relations between referring expression, sense and reference as two separate relations (the expression expresses a sense and also refers to a reference) or as a "dog-leg" (the reference is referred to via the sense, qua mode of presentation).

That is, would he prefer this:



or this:



I'm assuming that he would prefer the latter.

mode of presentation component.[10] The individual-directedness component relates the representation to the object I would point to when asked what I represent, it identifies the object my actions, beliefs, etc. are directed towards: namely, the physical object which caused my perception. The mode of presentation component is like the description under which the object pointed to is represented, it's the character of the abstract object expressed by the representation. The representation's constituents indicate the properties the abstract object has. These are the properties the physical object is represented as having. Thus I point through the abstract object to the physical object which caused my perception; the physical object is represented as something which has the properties possessed by the abstract object. So the individual-directedness component of the representation stipulates that the representation represents the physical object which caused the perception, and the mode of presentation component defines the sort of thing this object is represented as.

Note that this "representing as" isn't a new relation–the represents relation shouldn't become a "represents as" relation. I just mean that the represents relation points through the abstract object. R represents physical object Y as an X, is diagrammed by the represents relation pointing from representation R through an abstract X to physical object Y.

This account then, specifies which particular physical object my actions and beliefs are directed towards, and it also specifies the properties I believe that object to have: those possessed by the abstract object expressed by the representation. For example, when I see an animal up a tree, and this experience activates my **cat** representation, I represent *that* animal (pointing) as a cat. I'd say (if anyone happened to ask) that I think *that* animal is the sort of thing which stands on four legs, is about calf-high to me, which feels soft and furry, which I'd like to pat, which usually likes being patted and which makes a purring noise when I pat it. It's *that* animal which my actions are directed towards; I attempt to coax that animal down from the tree calling "Here kitty."

Whether or not the object my representation is directed towards actually is a cat is unknown to me, but this doesn't make any difference to what this object (whatever it is) is represented as. What the object really is makes no difference to what I believe, nor to how I act. In both non-veridical and veridical situations I represent the object up the tree as a cat. I represent it as the sort of thing which purrs, drinks milk, and usually enjoys being stroked on the back. If this were a veridical case, and the animal up the tree was actually a cat, then the correspondence between the abstract object expressed by this

10 Matthen (1988) : pp6-8.

representation and the physical object represented would be perfect. The only difference between such a veridical case and a non-veridical case would be in the degree to which the properties of the abstract object correspond with those of the physical object up the tree.

However, one possible problem with this view is that it requires accepting a perhaps unattractive viewpoint about what is represented in non-perceptual cases. The trouble is that there's no physical object to represent, let alone to represent as anything. Because of this I'm forced to maintain that in such cases the representation doesn't represent anything.

To solve this problem we can borrow Frege's tactic (again). Frege showed that referenceless referring expressions aren't meaningless by showing that every referring expression expresses a sense, even if this sense doesn't present any reference. The sense the referring expression expresses gives a referenceless referring expression like "Santa Claus" meaning, even though it doesn't refer to anything. Here I'll bite the bullet, and hold that in the same way that a referenceless referring expression doesn't refer to anything, a non-perceptual representation doesn't represent anything. (In which case we should call it a "PATTERN OF ACTIVATION" rather than a "representation".)

Frege maintained that meaning unpacks into two parts, sense and reference, where the latter position may be vacant. If it's acceptable for Frege to do this, then it doesn't seem unreasonable to suggest that the semantic relation a representation stands in unpacks in a similar manner, into the abstract object expressed and the physical object represented, of which the latter position may be vacant. In cases of non-perceptual activity activates a PATTERN OF ACTIVATION, there is no physical object for the PATTERN OF ACTIVATION to represent, but the abstract object expressed gives the PATTERN OF ACTIVATION something to be *about*. The activation of the PATTERN OF ACTIVATION is about the abstract object, it just doesn't represent anything.

When I activate my **lasagne** PATTERN OF ACTIVATION while thinking about what to cook for dinner tonight, what should this PATTERN OF ACTIVATION represent? The lasagne which I'm going to cook tonight? This isn't a physical object, as at the moment it doesn't exist, and it may not ever come into existence if I change my mind and cook spaghetti instead. What about it representing the abstract lasagne this PATTERN OF ACTIVATION expresses then? This seems equally doubtful, since in perceptual cases a PATTERN OF ACTIVATION represents a physical object as the sort of thing characterised by the abstract object. Representing an abstract object as the sort of thing characterised by itself seems pointless nonsense, and also quite contrary to the spirit of this solution. We want a

PATTERN OF ACTIVATION to represent a physical object as the sort of thing characterised by the abstract object, if it represents anything that is. I don't see any great problem then, with the idea that **lasagne** *doesn't represent anything* in cases when it's activated by non-perceptual activity. In the example above, I'm thinking about the lasagne I'm going to cook for dinner, but I'm not representing one.

Remember that we've had two relations all along: the expresses relation and the represents relation. If we take away the represents relation, we still have the expresses relation left. Remember that Frege maintained that we can take away referring and we still have meaning; the expression's meaning is given by the sense expressed in such cases. Here we can take away representing and still have aboutness; when activated by non-perceptual activity the PATTERN OF ACTIVATION's activation is about the abstract object expressed.

So in non-perceptual cases, the diagram is a lot simpler than in perceptual cases. The only relevant parts are the PATTERN OF ACTIVATION and the abstract object expressed. Because there's no physical object involved, there's no object for the "represents" arrow to end up at.

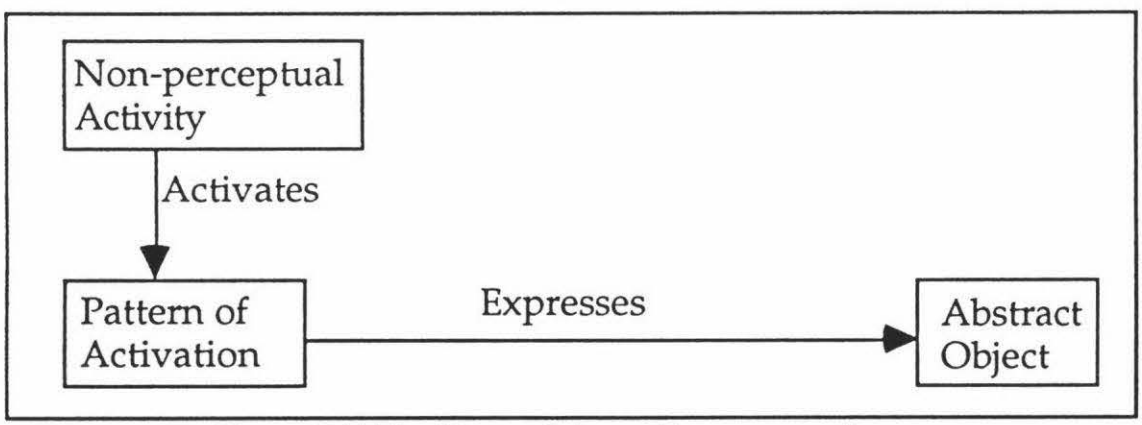


Figure 6.5(c) The "Fregean" theory: Non-perceptual cases.

This is not to say that my **lasagne** PATTERN OF ACTIVATION, which is activated by non-perceptual activity, is necessarily different to my **lasagne** representation, which would be activated by the outputs of my perceptual systems. The PATTERN OF ACTIVATION activated by thinking about what to cook for dinner may or may not be the same PATTERN OF ACTIVATION as that activated in situations where I believe I'm perceptually confronted by lasagne. These two PATTERNS OF ACTIVATION will almost certainly share some constituent patterns of activation. The only relevant difference here is that the PATTERN OF ACTIVATION activated when I'm thinking about what to cook for dinner isn't being used to do the job of representing.

So in perceptual representation, whether veridical or non veridical, the PATTERN OF ACTIVATION activated represents the physical object which caused my perception, representing it as having the properties possessed by the abstract object expressed by the PATTERN OF ACTIVATION. In some cases (a) the abstract object corresponds perfectly with the physical object; these cases we call "veridical." In (b) cases there is a less than perfect correspondence; this sort of case we call "non-veridical." In (c) cases where the PATTERN OF ACTIVATION is activated by non-perceptual activity, it expresses an abstract object which is the thing I'm thinking about, but this PATTERN OF ACTIVATION doesn't represent anything. (This is why it's called a PATTERN OF ACTIVATION rather than a representation.)

Note that what is represented isn't the same thing in every case, as it was with the "Meinongian" view; with this "Fregean" view sometimes there is no object represented, depending on the circumstances of the PATTERN OF ACTIVATION's activation. But what is represented still isn't decided in an *ad hoc.* fashion, as it was with the "Brentano-ish" version. The "Brentano-ish" view was ad hoc. because what is represented depended on whether or not the perception was veridical; something I cannot know, and which has to be imposed on the system from a "God's eye view". But with this "Fregean" view of representation, nothing depends on stuff I cannot know. I can tell whether my PATTERN OF ACTIVATION was activated by perceptual activity or not. When I'm looking at an object and trying to identify it, then the PATTERN OF ACTIVATION activated does represent the object I'm looking at. And when I'm not looking at an object, but just thinking about it, the PATTERN OF ACTIVATION activated doesn't represent any object, though it does express an abstract object which is the object I'm thinking about. Whether a particular PATTERN OF ACTIVATION represents any object or not depends on how I'm using it, which I am perfectly aware of. This system is not *ad hoc.* then.

So this "Fregean" view seems to have several advantages over the views I discussed previously. It seems to account for the way we normally describe cases of misrepresentation: I would normally describe the non-veridical situation discussed above by saying that "I misrepresent the possum as a cat," and according to this account this is exactly what does happen. Thus the representation does misrepresent in the sense that it represents something it shouldn't represent ("shouldn't defined by the correspondence between the abstract object expressed and the physical object perceived). And it does do so because it was activated by the outputs of the perceptual systems where a different representation would have been activated if the activating perceptual

information was of better quality. This view also has the advantage of accounting for the individual-directedness as well as the mode of presentation components of representation. It also allows us to stipulate in a non-*ad hoc.* fashion what is represented in veridical, non-veridical and in non-perceptual situations.

6.5 "Mason's final and utterly complete" view –(an amended "Fregean" view)

There is one problem with this "Fregean" view though. It seems that I've gone full circle, and come back to the view that what is represented is the object which caused my perception, *because* it's the object which caused my perception. This is (almost) the Crude Causal Theory's assumption, which I mentioned back in Chapter Two. Recall that the Crude Causal Theory subscribes to the following two assumptions (among others):

(1) physical objects themselves cause my representations' activation,
(2) because of this a representation must represent the physical object which caused its activation.

Most theorists take issue with the second assumption, claiming that sometimes the representation *misrepresents* the object which caused the representation's activation. I'm taking issue with both. I've rejected the first, showing that the perceptual systems' outputs, and not physical objects, cause the activation of my representations. I've already rejected the second assumption in one way too, the "Fregean" system rejected this by adjusting it to read "a representation represents the physical object which caused the activating *perception*", rather than the object which caused the representation's activation.

There is one further step I need to make. There is a question still to be answered: "how do I tell which object caused my perception?"

So far we've taken the physical object represented to be the physical object which caused my perceptual system to pick up and encode information as a pattern of activation which caused the PATTERN OF ACTIVATION's activation: that is, the physical object which caused my perception. This is the mistake made by the Crude Causal Theory. Taking the roundabout route to the physical object so that it is represented via a certain mode of presentation, the abstract object, means that this "Fregean" view doesn't make the Crude Causal Theory's mistake of claiming that representations *correctly* represent the object which caused my perception. (The Crude Causal Theory had "representations correctly represent whatever object is capable of causing my representation's activation"). But we still have the disagreeable idea that a certain physical object is

represented because it caused my perception. The idea is disagreeable because (again) it appeals to my knowing things I cannot know. I can sometimes be mistaken about the identity of the object which I pick up information about. The following two examples show how taking the object which actually did cause the PATTERN OF ACTIVATION's activation as the physical object represented can lead to problems:

(i) Sometimes perceptions I take to be caused by a single object were actually caused by more than one object.

A ventriloquist's act is a good illustration of this source of problems. The ventriloquist sits on the stage apparently having a conversation with a short wooden person–a "dummy". Imagine I'm in the audience, and don't realise that the point of the act is to make it appear that the dummy is talking, while the ventriloquist is actually the source of the dummy's voice. I mistakenly believe that the dummy is the cause of the voice I hear. I believe this because the information my visual and auditory perceptual systems are picking up is taken to be information about the same object (this is the point of the illusion, after all). The representation activated by this perceptual experience expresses an abstract object whose properties include being rather vocal and argumentative, being wooden, having a bad haircut and so on. The question is: what object do I represent as being characterised by this abstract object? So far I've taken the object I represent as the actual cause of my perceptual experience. But here there are two actual causes of my perceptual experiences. The actual cause of the visual information I pick up is the dummy, but the actual cause of the auditory information I pick up is the ventriloquist, not the dummy. Which object is the relevant actual cause to represent then: the ventriloquist or the dummy?

Before I answer, another situation where taking the actual cause of my perception as the object represented is problematic:

(ii) Sometimes it's hard to tell exactly which physical object is the one which actually caused my perception

Imagine I'm standing outside with my friend Diedre on a clear summer night in a spot away from the city lights, where we can see the stars. Suddenly a star I'm looking at flashes twice as bright as it was, and then return to its normal brilliance. I turn to Diedre and excitedly relate what I saw. "Which star?" she asks. I look back at the stars, and point to one, firmly believing that the star I'm pointing to is the one I saw flash. As it happens the star which flashed briefly was Beta Centuri. But because I turned away and then looked back, when I

looked back I looked at the wrong star, and am now pointing to Alpha Centuri. Here my representation expresses an abstract object with the property of having flashed brightly a moment ago, so I represent a star as having this property. But which star do I represent as having this property? Beta Centuri is the actual cause of my perception, but I believe that Alpha Centuri was the star that I was looking at.

In example (i), the obvious answer to the question "What do I represent as having the properties possessed by the abstract object?" is the dummy: the object I believe is the cause of all my perceptions, the object I take all the information I pick up to be information about. I represent the dummy as being the source of the voice I hear, being wooden, having a bad haircut and so on. In example (ii), the information about which object really caused my perception is something I have no access to now. Whatever the actual cause was, my representation is directed towards the star I believe was the cause of my perception. I represent Alpha Centuri as a star which flashed. The object which actually caused my perception is quite unrelated to what I represent now and how I represent it.

From this we can see that the correspondence relation shouldn't be between the abstract object and the physical object which is the *actual* cause of my perceptual state. Rather the correspondence relevant to the representation relation is the correspondence between the abstract object and the physical object which *I believe* is the cause of my perceptual state. In example (i), the correspondence isn't between the abstract object and the ventriloquist, it's between the abstract object and the object I presume is the cause of my perceptual state: the ventriloquist's dummy. With example (ii), the correspondence which matters to the veridicality of my representation is that between the abstract object and the star I believe I saw flash, the physical object which I presume caused my perception.

My representation is non-veridical in example (i) because the object I presume is the cause of my perceptual state, the dummy, doesn't correspond perfectly with the abstract object my representation expresses; the dummy doesn't talk, but I represent it as something which talks. In example (ii), my representation is non-veridical because (although I did see a star which flashed) a pattern which indicates the perceptual property of being seen to flash just a moment ago is also a constituent of this representation (this constituent activated the pattern of activation which indicates the property of *being* a star which flashed), when it isn't actually an object I saw flash. In each of these cases

the abstract object doesn't correspond perfectly with the object I presume caused my perception.[11]

Getting the object I presume is the cause of my perception to be the same object as the actual cause of the perceptual experience is important. The object I believe (correctly or not) to be the cause of my sensations is the object I attribute properties to, and it's the object I interact with on the basis of the information picked up by my senses. In the case of the ventriloquist's dummy I would point to the dummy itself, saying "That's something which talks." Since looking at something which is making a noise gets both ears pointing directly at the source of the sound for optimum sound pickup, I (incorrectly) look at the dummy when it's "talking" in the (mistaken) belief that in doing so I'll hear it best. The object which is presumed to be the cause of my perceptions is the object properties are attributed to, and the object my actions are directed towards.

Ideally the object I presume is the cause of my perception is the object which actually did cause my perception. Accurately figuring out the correct cause of your perceptual experiences is one of the essential elements to survival. Mother Nature selects for the ability to accurately tell (or to learn to accurately tell) which object in your environment is the actual cause of your perceptual state. If you can't do this well, then evidently the chances of your surviving long enough to have descendants are slim. This seems to be part of the reason why we've evolved with such useful faculties as binocular vision, to detect distance to the object viewed, and two ears to enable us to tell which direction a sound came from. With these faculties we're able to be more successful in correctly figuring out which object is the actual cause of our perceptions. (We're also able to be more successful because we're better equipped to get sound information married up with the correct aspect of the visual information we pick up.)

Anyway, we're going to have to revise the diagram to incorporate the fact that the object I presume is the cause of my perceptions is the physical object relevant to representation. The abstract object corresponds with this presumed cause of my perceptions rather than the actual cause. Thus the object represented is the presumed cause rather than the actual cause of my perceptual systems' current state.

---

11 Most of the way through this paper, I've used "non-veridical perception" and "non-veridical representation" as meaning basically the same thing; and haven't drawn any significant distinction between them. However, this might be the place to draw a line between non-veridical perception and non-veridical representation. We could say that my perception was veridical, as I did see a star which flashed, but my representation is non-veridical as I represent Alpha Centuri as having a property which it doesn't have.

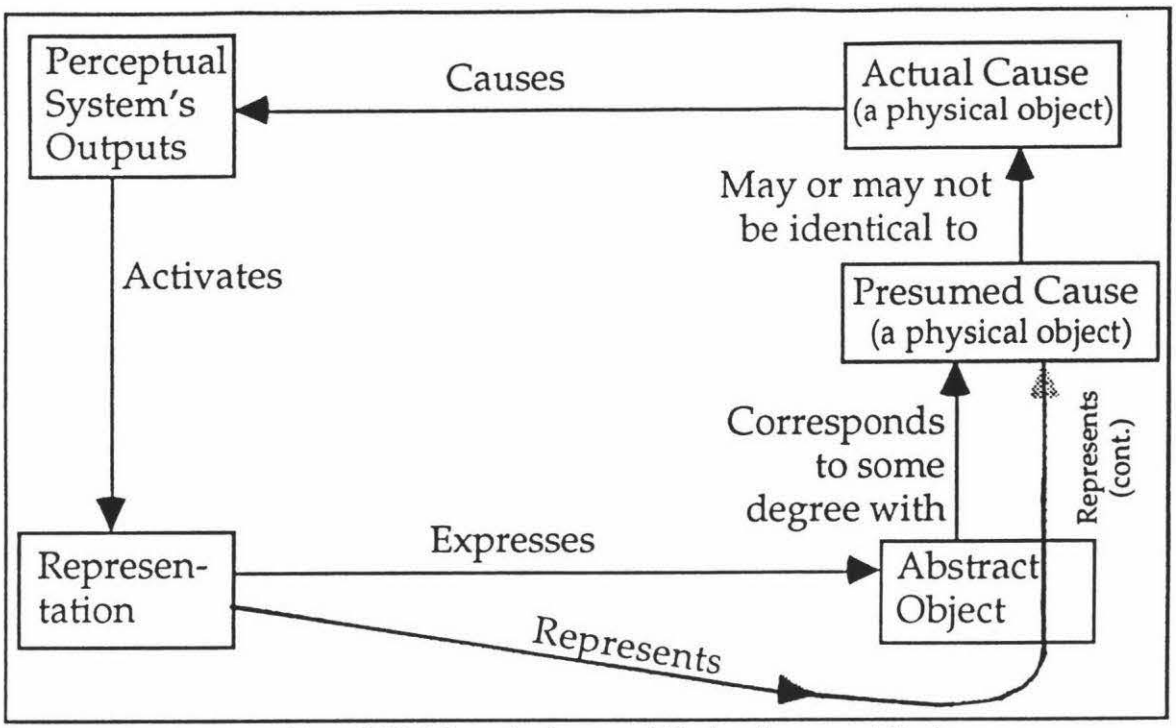


Figure 6.6(a) and (b) Mason's view: veridical and non-veridical cases

Some points need to be made clear here. The presumed cause and the actual cause are both physical objects, and these physical objects are both present in my environment. Ideally they're the same object. The may or may not be identical to relationship between them is not the sort of relationship which might apply between two separate but identical objects; between two peas in the pod, so to speak. It's not an "is the same as" relationship, as in the presumed cause has identical properties to those of the actual cause. Rather it's a relationship which states that the object I presume is the cause of my sensation either is or is not *the same object* as the actual cause of my perception.

Someone might take this diagram to mean that a veridical perception relies on getting affirmative answers to the two questions:

(i) "Am I representing the actual cause of my perception?" and
(ii) "Am I representing it as the sort of thing it really is?"

This is misconstruing the situation though. The wrong questions are being asked. I would answer (i) with a straight-out "No!" in all cases of perceptual representation. I never represent the actual cause of my perception, I only ever represent the presumed cause. Question (i) needs to be re-stated as:

(i) "Is the object I'm representing the same object as the object which actually caused my perception?"

Then the right question to ask is:

(ii) Am I representing the object I presume is the cause of my perception as the sort of thing it really is?"

If the answers to each of these questions is "yes" then I'm representing veridically.

An interesting point should be about these questions. A representation can be non-veridical because the answer to (ii) is "no", or because the answers to both (ii) and to (i) are "no". I find it very hard to see how a representation can be non-veridical if (ii) is answered "yes, you are representing the presumed cause as the sort of thing which it really is", but the answer to (i) is "no, the object you're representing is not the same object as the cause of your perception." I find it hard to see how I could be representing the wrong physical object, but representing it as the right sort of thing.

A candidate example could be if two stars flashed simultaneously, and I only saw one of them, but then mistakenly pointed to the other star which flashed, representing this other star (which did flash, but which I didn't see flash) as the sort of thing characterised by the abstract object expressed by my PATTERN OF ACTIVATION. This could be seen as being a non-veridical representation because the answer to (i) is "no" as the presumed cause is not the same object as the actual cause. The problem is that this representation is non-veridical because the answer to (ii) is also negative. The problem is that the property of being a star which I saw flash is indicated by a constituent of the representation activated. Here my perception is non-veridical for two reasons. Firstly because I'm representing the wrong star; the presumed cause isn't the same object as the actual cause of my perception. But it's also non-veridical because the star I'm representing doesn't have the property of being a star which I saw flash. It is a star which flashed, but it isn't a star which I saw flash, I saw a different star flash. Thus this object I presume caused my perception doesn't correspond to the abstract object expressed. The answers to (ii) and to (i) are both "no", and so the representation is non-veridical. It seems then, that (i) alone can't be answered "no". Either (ii), or (i) and (ii) together could have negative answers, but it seems that the answer to (ii) alone can't be "no". (Of course I could be proved wrong about this, but it isn't a crucial point.)

With that sorted out, let's look at how "Mason's final and utterly complete view" handles the three cases a semantic theory must deal with: veridical, non-veridical, and non-perceptual representation.

(a) In a case of veridical perceptual representation, I represent a physical object via an abstract object. The "represents" arrow points through the abstract object. I represent the physical object which I presume caused my perception as

the sort of object characterised by the abstract object expressed by the representation. In veridical cases the physical object I presume to be the cause of my perception is correctly characterised by the abstract object. And this object is also the same object as the object which caused my perception. This was pictured in figure 6.6

(b) In non-veridical cases, I misrepresent the physical object I presume is the cause of my perception. I do this by representing it as something which it is not. In such a case, the diagram is exactly the same as in the veridical case. The difference between veridical and non-veridical cases is that in non-veridical cases the abstract object's correspondence with the physical object I presume is the cause of my perception is less than perfect. In a non-veridical case we also have the possibility that question (i) is answered negatively. In addition to the presumed cause being represented as something which it isn't, it's also possible that the physical object which I presume caused my perception is not the same object as the physical object which actually caused my perception.

(c) Since the only changes from the "Fregean" view are to the identity of the physical object represented, and in non-perceptual cases no physical object is represented , the diagram for non-perceptual cases is the same as in the original "Fregean" view. The only relevant parts are the PATTERN OF ACTIVATION and the abstract object expressed. Here the abstract object is what the PATTERN OF ACTIVATION is *about*, but nothing is represented.

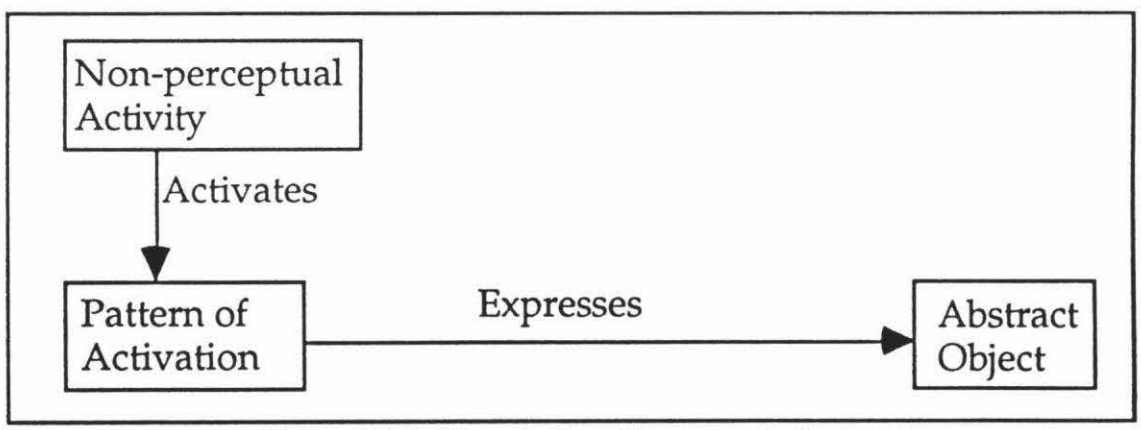


Figure 6.6 (c) Mason's view: Non-perceptual cases.

This system seems to provide an desirable account of representation for several reasons then. Firstly it provides a coherent account of the two components of a representation: the individual-directedness and the mode of presentation. That is, is shows what object the representation represents, and shows what this object is represented as. Also this system illustrates what is represented in veridical, non-veridical and in non-perceptual cases, in a

manner which is not *ad hoc*. It doesn't require a "Gods eye view" to tell what is represented; in situations where the PATTERN OF ACTIVATION doesn't represent the object I believe caused my perceptual state, this is because I don't believe any object caused the perceptual state which activated this PATTERN OF ACTIVATION.

Before I submit this account as a solution to the problem of representation however, I'll first put it through a "reality" test. In the next chapter, I'll look at how this account explains how this system of PATTERNS OF ACTIVATION would work in practice, and demonstrate how such a system could implement some systems which account for features of the way I successfully interact with objects in my environment. I'll also demonstrate that this system meets Kim Sterelney's four criteria for a theory of the "content" of mental representations.

CHAPTER SEVEN

TESTS OF AN ADEQUATE THEORY OF REPRESENTATION

7.1 Summary: The story so far

Over the last five chapters I've outlined a rough sketch of the way a connectionist-based system of representations could work. In Chapter One I gave an indication of the sort of problems theories of representation have faced: chiefly the problem of how to explain how a representation can represent certain things and misrepresent other things. Here I also gave a general tactic which has been used in attempts to solve the problem of misrepresentation. This tactic was to explain in a principled, non-circular, non-intentional way, how a representation could come to represent what some things, and not represent other things.. This could then be used to identify cases of misrepresentation. A case of misrepresentation is a case where the object represented had different properties from those the representation should represent.

In Chapters Two, Three, and Four I built up connectionist-inspired way of viewing what a representation is, how it's activated, and how it's parts are related. This concept of what a mental representation could be like is a refreshing and revolutionary change from the way a representation has been viewed in traditional accounts of representation. Here we have a representation which is constituted by a collection of parts; each part, or constituent, encodes some piece of information about the environment. In Chapter Five I showed how each constituent of a representation indicates some property, perceptual aspect or affordance of an object. Our picture of a representation is now a picture of a dynamic entity which can be subtly different every time it's activated. Now my **cow** representation can be a different entity each time it's activated; it can have a different set of constituents which have a "family resemblance" to other instantiations of my **cow** representation.

In Chapter Five I also showed how each constituent of a representation comes to indicate the property, aspect or affordance it indicates. This

happens because the way these constituents encode information evolves from the fundamental perception-action connections an infant is "hard-wired" with at birth. These constituents develop through their use in producing behaviour. The point of having representations is so that the agent's perceptions of their environment can be used to initiate and guide actions. The constituents of representations develop so that the actions produced will be more successful; they will be more appropriate to the way the agent perceives their environment and will be more effective at achieving the agent's goals. This co-evolution of perception and action forms the neurological structures which we call "representations". The connections between encoded information picked up from the environment and encoded actions which enable appropriate behaviour to be produced, develop through the agent's history of embodied action.

Because of this developmental history, the constituents of the agent's representations have evolved so that each instantiation of a representation expresses an abstract object. In Chapter Five I illustrated how this abstract object relates to a physical object. This abstract object sometimes corresponds with the object represented (in the sense that all the properties the abstract object has, the physical object has also), and sometimes it does not correspond so well. When the abstract object doesn't correspond perfectly with the object represented, this is a case of misrepresentation. Here I also described what the representation looks like. In all cases the object represented is represented as being characterised by the abstract object. The abstract object is like a description of the object represented; the object represented is represented under this description. In cases of misrepresentation, the object is represented under a description which incorrectly characterises this object. In cases where the PATTERN OF ACTIVATION is activated non perceptually the abstract object characterises the object I'm thinking about, but no object is represented.

So it seems that I've given an outline of an account of representation which explains both representation and misrepresentation. Before I offer this account as a successful account of representation, I'll show how it provides an implementation-level explanation of two similar cognitive phenomena noticed recently, and measure it up against four criteria which Kim Sterelney suggests any adequate theory of representational content must satisfy.[1]

---

1 Sterelney (1990) : p114.

7.2 *Perceptually guided action*

Sometimes I can pick up on some piece of information which specifies an affordance an object provides, react appropriately, and *then*, after reacting to the information which prompted my reaction I pick up on other information about the object. For example, I can visually pick up the information that something is flying through the air towards my head, this information being encoded as patterns of activation which indicate the trajectory and speed of the object. The activation of these patterns can activate a pattern of activation which indicates that this object affords *collision*, and this can in turn activate patterns of activation which initiate and guide evasive actions. Thus when I perceive something flying through the air towards me, I can react immediately by ducking.

Because the neurological structures which underlie cognitive processes are developed through an agent's history of using perceptions to guide action, Varela et al. claim that in situations like this when performing familiar, well ingrained actions guided by our perceptions, this can be carried out without representation playing a role.[2] Their point is that I can sometimes duck to avoid an object on a collision course without activating a representation of exactly what it is I'm ducking out of the way of. According to Varela et al., representations are simply not needed in a lot of cases; a significant proportion of our interactions with objects can be carried out without representations entering into the picture. They maintain that this sort of *perceptually guided action*, rather than representation, is the central component of cognition.[3] (Note that this account uses "representation" to mean some cognitive structure which is used to stand for an object or state of affairs. So Varela et al.'s claim that representation does not enter into many cases is a claim that a neurological item which *stands for an object* isn't used in such cases. This is a similar worry to the one I expressed in section 5.3, where I worried that PATTERNS OF ACTIVATION are used to interact with objects rather than to stand for objects.)

Thus in a significant amount of my interactions with objects, I can successfully take advantage of the affordances objects provide without the need to explicitly represent every object I interact with. When negotiating my way across a crowded marketplace, I don't need to use a detailed explicit representation of every person I pass. I merely need to recognise the fact that

---

2 Varela, Thompson et al. (1991) : pp201-213. They also defend this claim by citing Rodney Brooks' (1987) finding that when attempting to design cognitive "creatures" which perform low-level tasks such as exploring their environment, "explicit representations and models of the world simply get in the way. It turns out to be better to use the world as its own model." See Daniel Dennett's(1984, discussion of the frame problem for an elaboration of the idea that explicit detailed representations are difficult and unwieldy.)

3 Varela, Thompson et al. (1991) : p173.

that person affords a barrier, that that gap between people affords passage, and so on. All I'm interested in is situations which afford my passage or prevent it.

A similar case of perceptually guided action happens when I'm deeply involved in what I'm *doing*. I can be so involved that I don't need to employ representations of the objects I'm using. Even when using objects as tools, and taking advantage of what uses they afford, the tool itself doesn't *need* to be represented. Take the example of hammering in a nail, for instance. There is a sense in which I don't have to employ a representation of the hammer in order to use it to bang in a nail. My ability to act successfully comes from my experience and familiarity with hammering, not from my knowledge of the hammer. Winograd and Flores summarise the idea as follows:

> "To the person doing the hammering, the hammer as such does not exist. It is a part of the background of *readiness-to-hand* that is taken for granted without explicit recognition or identification as an object. It is part of the hammerer's world, but is not present any more than are the tendons of the hammerer's arm.
>
> The hammer presents itself as a hammer only when there is some kind of breaking down or *unreadiness-to-hand*. Its 'hammerness' emerges if it breaks or slips from grasp or mars the wood, or if there is a nail to be driven and the hammer cannot be found... As observers, we may talk about the hammer and reflect on its properties, but for the person engaged in... unhampered hammering, it does not exist as an entity." [4]

A similar experience of readiness-to-hand occurs when I'm riding my bicycle. Here I don't need to have an explicit representation of my bicycle activated. The central cognitive concept here is perceptually guided action, not representation. When I'm riding it down the street, I don't need to represent my bicycle as anything separate from me; all there is is "me" whizzing down the street. The bicycle is ready-to-hand; while I'm engaged in "unhampered cycling" I turn corners, brake, avoid obstacles, change gears, and respond in many other subtle ways to a vast range of perceptual stimuli without the need to explicitly represent the bicycle. I have the potential to activate a representation of my bicycle if something calls my attention to the bicycle explicitly: something like, for instance, a strange noise coming from the back wheel, the pedal falling off, or the brakes failing. In normal unproblematic riding however, the bicycle doesn't exist as a represented entity. I'm using the bicycle with regard to the fact that it affords transport without any explicit representation of the bicycle.

This concept of perceptually guided action illuminates the idea that there are *some* aspects of interaction with environmental objects whose

---

4 Winograd and Flores (1986) : p36. This view takes its root from Heidegger; "readiness-to-hand" is a translation of Heidegger's concept of *Zuhandenheit*. See for instance Blackham (1952) .

success does not seem to require operations involving explicit representations of every aspect of the objects interacted with. I'm going to show how the connectionist representation system I'm outlining can explain the mechanisms which facilitate perceptually guided action like this (although hopefully it's becoming obvious already). But first I'll discuss a point of view which is closely related to this concept of perceptually guided action. This view attempts to draw a distinction between affordances we can pick up without the need to identify the object which provides the affordance, and affordances which aren't picked up until the object is identified. (Compare the property my record player has of being about waist-high with the property of being something bought for $20 at a garage sale.) The system I'm outlining will clarify this line of thought too.

7.3 Encoded affordances at many levels

Ulric Neisser also noticed a phenomenon similar to enactive cognition's concept of perceptually guided action without explicit representation. Neisser claims that some affordances can be picked up and reacted to without the need to identify the object which provides the affordance. He calls these "physical" affordances.[5] Neisser contrasts these with affordances which do require identification of the object which provides the affordance to pick up; these Neisser called "cultural" affordances. Cultural affordances are uses which objects afford only through our cultural conventions; they are affordances which can't be recognised until we have identified the object which provides the affordance. A telephone's affording long-distance communication is such a cultural affordance. We don't pick up information about the telephone's affording long-distance communication from the physical qualities of the telephone; we need to recognise the object as a telephone before we activate any pattern of activation which indicates that this object affords long-distance communication. (Think about those novelty telephones which look like shoes, bananas or model cars.)

Neisser postulates the existence of two separate processing systems to perform what he saw as two different functions. The first processing system is responsible for picking up on information about physical affordances, so that I can react appropriately to the information I pick up without having to identify the objects I'm interacting with. The second processing system Neisser postulates is the system which is used to identify objects so that I can pick up on cultural affordances. If Neisser is right, and there is a division

---

5 Neisser (1989) This is an unpublished paper delivered at a conference. I'm using this as cited and explained in Bechtel (1990, p272 ; Bechtel and Abrahamsen (1991, p268)

between these different sorts of affordance, then I don't believe we need different processing systems to account for our ability to pick up on affordances without identifying the object which provides the affordance. I'm going to show that this can be explained by different "levels of processing" within the representational system I'm outlining.

This explanation, as I mentioned earlier, also accounts for perceptually guided action without activating explicit representations. In certain situations, when performing certain tasks it could be said that PATTERNS OF ACTIVATION are being employed in coordinating perception and action, but not as representations which "stand for" a particular object. The PATTERN OF ACTIVATION activated when I duck to avoid collision might have constituents which encode information about the trajectory I perceive the object as having, the speed with which it's coming toward me, and which indicate that this object affords collision. But the PATTERN OF ACTIVATION activated need not have constituents activated which encode what the object smells like, exactly what shape it is, what it's made of, what colour it is, and so on. I don't need to know what the object is in order to react appropriately. This information isn't necessary to my ducking to avoid getting hit by this airborne projectile. After I duck I *could* pick up this sort of information, and activate patterns of activation to encode what the object is and so on, but in order to act appropriately I don't *need* to do so.

An example will help illustrate this point. Say I'm standing in my back yard, and a cricket ball comes flying towards me. Without recognising that the object is a cricket ball, I can notice that whatever-it-is which is flying towards me affords collision (what Neisser calls a physical affordance), and can react to avoid it by ducking. And after ducking I could observe the object as it bounces against the fence and comes to rest under a rose bush, and identify the object as a cricket ball. I could then realise that because the object is a cricket ball, the object affords cricket-playing (a "cultural" affordance).

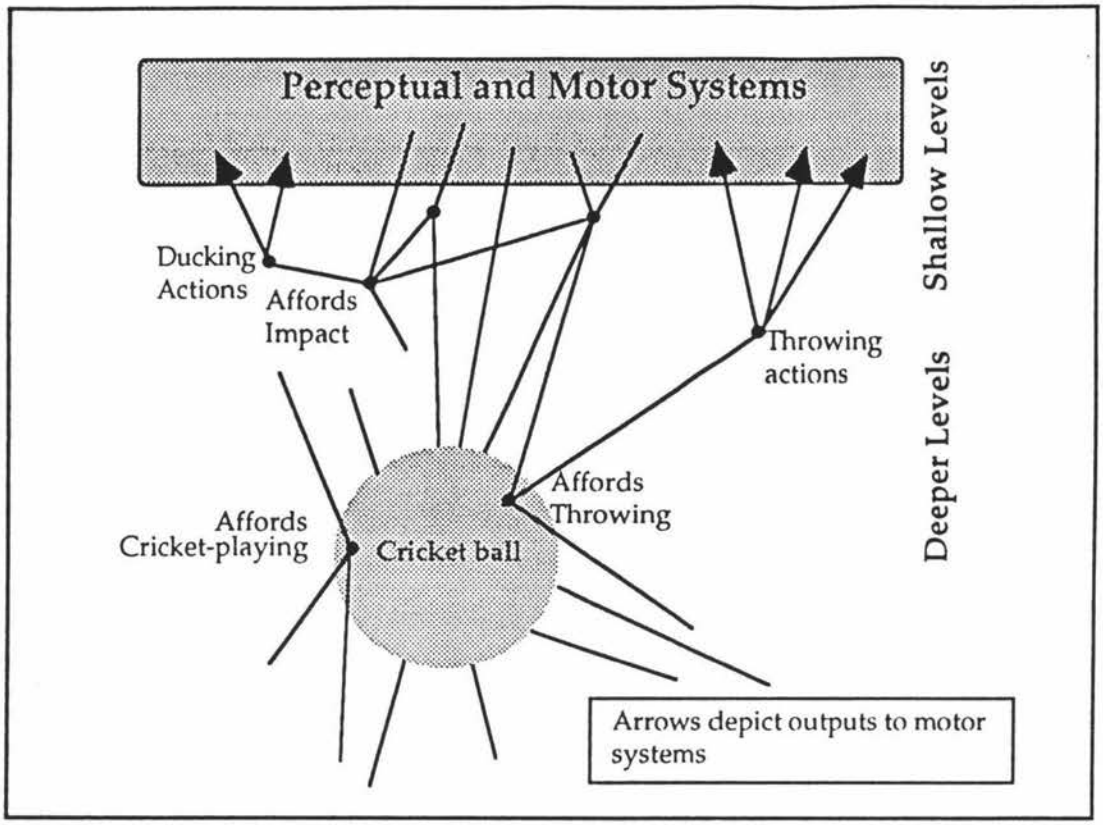


Figure 7.1 Affordances at many levels.[6]

In situations like this we can see that there can be connections between patterns of activation which indicate perceptually picked up properties of the object and those which indicate affordances at many "levels" of a representation. Figure 4.3 illustrates this idea.

This diagram illustrates what is happening in the above example. At shallow levels I pick up information about the object's trajectory and speed, which activates a pattern of activation which indicates that the object affords collision. Then after I react (or while or before, but anyway separately) I can perceive the object and identify it–the information I pick up activates patterns of activation which activate my **cricket ball** representation. The activation of this representation incurs the activation of a pattern of activation which indicates an affordance provided by this object: it affords cricket-playing. (A golf ball might also afford cricket playing, but this affordance seems more readily activated by the cricket ball. I might only notice that the golf ball affords cricket playing if I wanted something which provides this affordance, and a golf ball was the best thing I could find.)

Here we can see Neisser's two processing systems merely as deeper and shallower levels of activity in the same overall connectionist-based system. At shallow levels there could be connections from patterns of activation

6 Note that I've drawn the cricket ball representation as a sort of "black box". This is because I want the structure of the interconnections of the constituents of this representation to remain ambiguous. It could have a central constituent, but it might not (See section 4.7) I've only drawn in the patterns of activation which indicate the affordances relevant here.

which encode perceptually picked up information to patterns of activation which encode information about affordances. These patterns of activation which indicate "physical" affordances could be connected to patterns of activation which initiate and guide the actions performed in taking advantage of that affordance. This happens when I duck to avoid whatever-it-is which is flying through the air towards my head. Because of these shallow level connections, information about affordances can be picked up and reacted to without needing to identify the object which provides the affordance.

Perceptually guided action like this *can* often happen without going deeper, to activate entire PATTERNS OF ACTIVATION which represent the object which provides this affordance. This could also be happening with the example Winograd and Flores use of hammering in a nail. When hammering a nail I have a node activated which encodes information about the amount of nail still left to be hammered in; this node's function is to indicate whether I need to hit the nail again. This pattern of activation is connected to another pattern of activation which initiates my hammering actions. These two patterns of activation work together in a continuous input-output loop. As long as the first pattern of activation encodes the information that there is still some nail left to hammer in, I'll continue to hit the nail. That is, if interruptions or breakdowns don't occur. This simple apparatus shows how actions can be perceptually guided without the need to involve an explicit representation of the objects interacted with.

On other occasions however, we *do* need to represent the objects interacted with. On these occasions we get activation at a deeper levels; whole PATTERNS OF ACTIVATION are activated, by which the objects present can be identified and represented. If Neisser is right and in some cases we do need to identify objects, presumably by activating representations of them, before we can pick up information about some affordances (those he called "cultural" affordances), then the sort of system illustrated above can explain how this happens.

PATTERNS OF ACTIVATION which are representations of objects can have constituents which indicate "cultural" affordances like this. This happened when I recognised the object I'd ducked as a cricket ball, and realised that this object affords cricket-playing. The aspects of the ball were recognised and my **cricket ball** representation was activated. This had a constituent pattern of activation whose activation indicates that this is an object which can be used to play cricket with. Constituents like this, which indicate affordances could have connections to patterns of activation which initiate the actions involved in taking advantage of the affordance indicated.

Although the system I'm outlining shows how information about these so-called "cultural" affordances can be encoded, I'm reluctant to agree that it's only patterns of activation which encode such "cultural" affordances which are activated in this way. I've already shown that some properties of objects (like my record player being an object bought for $20 from a garage sale a couple of years ago) are not carried in the information about the object picked up perceptually, and which instead are encoded by patterns of activation activated only when the object is identified.

It seems that sometimes "physical" affordances are also not noticed until the object which provides the affordance is identified. Often information specifying a particular use an object lends itself to is not picked up until after the object is identified. Sometimes an object provides an affordance that isn't picked up on for weeks, or even years, because no one has been in a situation where they're looking for an object which can be used in this way.

This happened to me a few weeks ago, when the chain came off my bicycle. There's a special piece of metal by the cogs on the back derailleur which is *supposed* to prevent the chain coming off. In order to get the chain back on I needed to unscrew the screw securing this piece of metal in place, to get the chain back past this "keeper". I needed something which afforded a screwdriver. I didn't care what the screwdriver actually was, or what other affordances it might provide, I just wanted something with a rigid, thin blade-shaped end. My fingernail wasn't strong enough, and none of the coins I had in my pocket were thin enough. The solution to my problem came when I remembered something I had passed a moment before the unfortunate incident– a broken off windscreen wiper in the gutter. It would probably be rigid enough to turn the screw, and thin enough to fit in the screw's slot. It might afford a screwdriver. This was an affordance which wasn't immediately activated when I first perceived the windscreen wiper, but when I thought about the aspects of shape I picked up on, and identified the object at least as far as identifying that it is made of metal, I could imagine that this piece of metal might provide the affordance I needed. And when I went back along the road and got this object and tried to use it, I found that it *did* afford the use I needed. I could unscrew the screw, put the chain back on, screw the piece of metal back on and then ride home.

This affordance was a use the windscreen wiper "physically" afforded, because of its physical characteristics, rather than being any "cultural" affordance broken windscreen wipers are typically used for according to the conventions of the culture I operate in. Affording a screwdriver is not a constituent of my **windscreen wiper** representation (or at least it wasn't before this happened). But this is an affordance which wasn't picked up

straight away; it was picked up after I had identified the object, ridden past it, the chain had come off, and then I realised that I needed a screwdriver. Then I remembered that I'd seen a windscreen wiper in the gutter, and that it might be the right size, shape and rigidity to afford a screwdriver. If I'd been looking for something which affords a screwdriver, and *then* had seen the broken windscreen wiper, the information that this affords a screwdriver *could* have been activated without having to identify the object. But in this case I did identify the object, and through identifying it and being aware of its properties I realised that it provided the affordance I needed. This sort of case where physical affordances aren't activated until the object is identified might be seen as further evidence that we don't need two *separate* processing systems here; but that this is all just different levels of processing in the same overall system for implementing representations.

7.4 Sterelney's criteria

So this system looks like it can provide an implementation level explanation of perceptually guided action, and of Neisser's two processing systems. Now I want to conclude this thesis by quickly running this system past a set of conditions which Kim Sterelney rightly sets out as criteria of adequacy for a theory of the "content" of mental representations.[7] (For "content" read "the sort of thing a representation represents, sand how it can do that".) He says that we need a theory which meets four conditions. Sterelney's criteria are these:

(a) The account of the content of representations should not be circular, and nor should it attribute magical powers to the brain; it should be compatible with a physicalist theory of mind.

(b) It should explain how the reference of concepts are determinate. My **cat** representation must represent only cats, and it must apply to all cats. That is, it must be extendable, so that it can represent cats other than the cats I've met so far, but not be extendable so that it can represent things other than cats.

(c) It should allow for misrepresentation; I should be able to misrepresent a possum as a cat. It should also account for denotationless concepts, (they should not be just "neural static") and it should distinguish between different denotationless concepts; it should explain the difference between my concepts of **Santa Claus** and **the tooth fairy**.

---

7 Sterelney (1990) p114.

(d) It should be consistent with the fact that most human concepts are learned, rather than innate. It should explain how I *come to have* the concepts **cat** and **Santa Claus**.

Sterelney claims that satisfying all these demands at the same time is a difficult business. So if I can show that this system can meet all the above criteria, and as such is a useful piece of explanatory equipment, I'll feel quite justified in offering it as an account of representations and their contents.

Condition (a) has been met by showing how the representation, its constituents, and what these constituents pick up and encode information about, and the way the representation is used to guide and initiate actions all develop together, from the connections which human children are born with. They develop by being used to interact with objects. This system is *definitely* physicalisable.

Condition (b) refers to what is traditionally called "the qua problem." This theory must explain how my **cat** representation is specific enough to represent only cats, and also general enough to represent all cats. Assuming this representation has been developed through exposure to a wide enough sample, then it will be to a certain degree general, and to a certain degree specific. How general and how specific it is depends on what is indicated by the representation's constituents, which depends on the history of use which developed the representation and its constituents. The "quaness" of any representation–what the object represented is represented *as*–is given by the abstract object expressed. The object is represented as an object with all the properties indicated by the representation's constituents, which can vary in its determinacy, but will always be determinate to some degree. This gives the representation the ability to be very determinate in the way an object is represented.

I spent most of Chapter Six demonstrating how the system I'm outlining meets condition (c). It does allow for misrepresentation. It also allows fro denotationless concepts, and explains the difference between different "denotationless" concepts. The difference is that patterns of activation which are activated when thinking about such concepts will express distinct abstract objects.

This account is *dependent* on the fact that it conforms to condition (d), that most human concepts are learned. The way they are learned, through developing the connections between patterns of activation, and the way patterns of activation encode information, is how representations come to be able to stand in the representation relation with bits of the world.

This system I've outlined here meets all Sterelney's criteria for an adequate theory of representation, then. And with that final "thumbs up" I leave the topic (except for an appendix).

APPENDIX ONE

GENUINE AND NON-GENUINE CASES OF MISREPRESENTATION

In Chapter One I made a distinction between different cases which are used in the literature to illustrate accounts of misrepresentation.

(i) even though Xs and Ys *can* cause sensations which activate representation **R** , **R** *should* only represent Xs, and thus misrepresents when activated by sensations caused by a Y.

(ii') when representation **R** happens to be activated by sensations caused by something which **R** shouldn't represent, like a Y for instance, **R** misrepresents the Y.

There is an important difference between the activation of *vague* representations like those in type (i) and the *inappropriate* activation of representations, as we find with type (ii) situations. In Chapter One I insisted that type (ii) situations, in which the representation is activated because I get poor quality or incomplete sense-information from an object, are the only places where we'll find genuine misrepresentation. Type (i) cases where a representation can be activated by two or more different things, but should only represent some of these things which can cause its representation are not the sort of cases where we'll find misrepresentation. The problem is that representations to which (i) applied are vague. They aren't specific to tell the difference between information about Xs and information about Ys.
I gave the following as a test for an example of genuine misrepresentation:

- If the environmental information was of better quality or more complete, the same representation wouldn't be ACTIVATED.
- If I attempt to ACTIVATE the representation through other of its constituents, by looking from a different angle, by listening, smelling, feeling and/or tasting as well as looking, or by looking closely at features not inspected originally, the same representation wouldn't be activated.

Let me try to explain this in more detail, by putting some examples used in the literature, and some examples which illustrate the problems with the traditional examples, to this test. I'll start with some obviously

disjunctive representations, and following that, I'll tackle the more problematic ones.

- Frogs snapping at beebee pellets. The frog would always activate this same representation. The viewing angle, distance, lighting, etc. are probably the best the frog is going to get in the time the it has to react. The frog doesn't have the time or even the physiological equipment to have a "really good look". It seems from the frog's physiology, that it doesn't have any other representation to activate. It has a set of nodes in its optic tectum dedicated to picking up information about small moving dots. These are connected *directly* to the mechanisms which initiate its snapping reaction. The frog will snap at *anything* which causes the activation of it's small-moving-dot detector, so it seems that it isn't possible for the frog to activate a **don't-snap-at-*that*-small-moving-dot** representation. This same representation (if we can call it a representation) will always be activated. Also, since there doesn't appear to be any other constituents to this representation, it appears that the frog can't do anything more to check the identity of the small-moving-dot before the decision to snap is made (if it's even a decision). The point here is that all the frog has to go on is the activation of its small-moving-dot detector, which is triggered by flies and by beebee pellets. So the frog's small-moving-dot representation is vague then, and can't misrepresent beebee pellets.
- Some bacteria are equipped with magnetosomes, which indicate the direction of the earth's magnetic field. In the northern hemisphere these bacteria use these magnetosomes because by pointing north they *also* point to oxygen-free environments where the bacteria prosper. The magnetosome is taken to misrepresent if such a bacteria were moved to the southern hemisphere where the indicator would point *away from* oxygen-free environments. Here the magnetosome would always point in the direction it does, and the representation (if we can call it that) activated would always represent the way it does. The bacteria can't tell the difference between the magnetosomes pointing to beneficial oxygen-free (when in the southern hemisphere) and its pointing to harmful oxygen-rich environments; no matter where the bacteria is, the magnetosome is always used to point to oxygen-free environments. Something is going wrong here, but we can't call it misrepresentation. The magnetosome can't point any other way: it already has the best information possible and there isn't any independent check the bacteria can make using a different "pattern recogniser" (though I doubt the bacteria has anything resembling a set of neurons which recognise

patterns). This representation (again, if we can call it that) is vague, and thus can't misrepresent.

- The stuff with chemical formula XYZ, which is perceptually identical to the stuff with chemical formula H_2O (it looks, tastes, smells, the same, it quenches my thirst, puts out fires, and is used to make coffee in the same way), causes the activation of **water**. Here the traditional story tells us that the content of my water representation is the stuff which causes its activation. So since I've only ever encountered H_2O, and have never before encountered XYZ, the content of my **water** representation is *"the stuff with chemical formulaH_2O"*. Thus according to some accounts when I perceive XYZ, and this activates **water**, I misrepresent XYZ as H_2O. But since I can't tell the difference between XYZ and H_2O even with the most complete and best quality sense information possible, I would still activate the same representation whatever the chemical formula of the stuff I encounter. So my **water** representation's content is vague, and it can't misrepresent when activated by XYZ.
- This example (taken not from the literature, but from a James Bond movie) illustrates further the idea that we can only have genuine misrepresentation *if I can tell* which of two alternatives is the case: 007 awakes from a nap to find a big black spider crawling across his chest. He thinks, "Oh gosh! That *might* be a really astoundingly poisonous spider." He assumes that it is poisonous and acts accordingly. (Due to his incredible bravery, cunning etc. and the fact that the movie's not even half over yet, he eventually manages to get the spider off his chest and squash it with his shoe –after the requisite amount of suspense, alarming music, close-ups of the spider and of beads of sweat on Bond's forehead, of course.) If the spider turned out to have actually been harmless (if Felix Leiter had rushed in with a spider identification manual and compared the squashed remains), then we would be very tempted to say that Bond misrepresented the spider. We would like to say that he activated **poisonous spider**, when the spider was harmless. But the test I've been advocating would show that he didn't misrepresent the spider, because Bond didn't *know* how to tell whether this spider was poisonous or not; so he wouldn't have activated any different representation, even if he looked really closely from the best vantage point in the best lighting conditions, and if he had enough sensory information to directly activate *all* the other constituents of this representation. (If it had bitten him, *then* he'd have had sensory evidence of its poisonous-ness.) He didn't really activate a **poisonous spider** representation, he activated a **big-black-hairy-spider**, and *assumed* that the spider was poisonous and acted on that assumption. His **big-black-hairy-spider** representation was activated, but

the pattern of activation which indicates whether the spider is poisonous or not is neither definitely activated nor inhibited. Bond activated **big-black-hairy-spider**, he didn't activate **big-black-hairy-poisonous-spider** nor did he activate **big-black-hairy-harmless-spider**. So the representation activated has the content *that's a big black hairy spider*, which applies correctly both to poisonous big black hairy spiders and to harmless big black hairy spiders. This then is a type (i) situation; the representation ACTIVATED was therefore disjunctive. So Bond didn't misrepresent here, he just made an assumption, which turned out to be incorrect.

- A cat on a dark night activates **dog**. This is quite obviously a case of genuine misrepresentation. If the lights were turned on or I got close enough to see better, or if I went up and patted the animal and it purred when I patted it, I wouldn't still activate **dog**. I would probably activate **cat** instead. So this is a case of misrepresentation. The constituent of **dog** activated when I saw the animal was activated by the pattern recogniser producing the wrong output because of the poor quality information it had to process. This representation was inappropriately activated, and would not have been activated if better quality, or more, perceptual information was picked up.
- My seeing a cardboard cut-out of a cow standing in a farmer's field activates my **cow** representation. I do misrepresent here, because if I looked from a different angle, or looked a bit closer then I wouldn't activate cow, I might activate **cardboard-cow** instead. I can tell the difference between a cardboard cow and a real cow. In this situation my **cow** representation was activated by a constituent which we might call **cow-image**, which is a constituent of both representations and might be said to be vague. It's correctly activated by things which look like cows. But my *entire* **cow** representation only correctly represents cows, because cardboard cows don't give milk, go "moo" etc.
- A thin water-buffalo standing in a farmer's field causes the activation of my **cow** representation. This is a debatable one. Whether or not this is a case of misrepresentation depends on whether I've been trained to tell the difference between thin water buffalo and cows. If I don't know what the difference between cows and water-buffaloes is, (which, to be honest, is something I really don't know) I would still activate **cow** even if I looked closely with my glasses on, in good light, from lots of different viewing angles. Even if I listened to, smelled, poked and milked (if this is possible) the water buffalo and I probably still wouldn't realise that this isn't a cow but a thin water-buffalo. This couldn't be a case of genuine misrepresentation then. The representation I have is a vague one which doesn't misrepresent thin water-buffaloes. But on the other hand, if I had

been trained to tell the difference between cows and water buffaloes, then I would *not* still activate **cow** when I looked closely etc., and would activate **thin-water-buffalo** instead, then this is a case of misrepresentation.

- A wax apple activates **apple**. Purely visually, I can't differentiate between the wax apple and the grown-on-a-tree variety, so the constituent whose activation activated the representation is vague. It indicates *"something which looks like an apple"*. But if I inspected the wax apple thoroughly, tasting, smelling and feeling it, then I would not still activate **apple**. This would qualify as a case of genuine misrepresentation, then.
- A rabbit sees an animal moving on the hillside and runs away to hide. If the animal happens to be a harmless wombat, the rabbit could be said to misrepresent here. This case could go either way, because it's pretty hard to get inside a rabbit's head. One of two things could be happening. It could be that the rabbit activates **dangerous animal** as soon as it sees movement on the hillside. In which case the rabbit does misrepresent, since it seems safe to assume that if the rabbit looked a little closer the rabbit would notice that the moving animal isn't a dangerous animal, but a harmless one. This then would count as misrepresentation, because the representation wouldn't be activated if the rabbit had a closer look. The problem is the decision costs: if the rabbit didn't run until it had had a closer look and was *sure* that the animal is a dangerous one it could become dinner for a hungry fox.

  But on the other hand it could be that the rabbit is as paranoid as James Bond. It could activate neither **dangerous animal** nor **harmless animal**, but just activate a vague **movement!** representation and takes the safest move and *assume* the movement indicates danger. In which case the representation activated is vague, and doesn't misrepresent.

The cases used to illustrate misrepresentation don't always exemplify misrepresentation then. Using the right sort of examples to illustrate accounts of misrepresentation could have prevented a lot of misunderstanding.

edn.) 51-88.

Fodor, Jerry A. and Zenon W. Pylyshyn (1981). "How Direct is Visual Perception?: Some Reflections on Gibson's 'Ecological Approach'." *Cognition* 9: pp139-196.

Frege, Gottlob (1892/1966). "On Sense and Reference." *Translations from the Philosophical Writings of Gottlob Frege* Eds. Peter Geach and Max Black. Oxford, Basil Blackwell. (Second (1966) edn.) 56-78. ((Originally published in 1892))

Gibson, James J. (1979). *The Ecological Approach to Visual Perception.* Boston, Houghton Mifflin Company.

Held, R. and A. Hein (1958). "Adaptation of disarranged hand-eye coordination contingent upon re-afferent stimulation." *Perceptual-Motor Skills* 8: pp87-90. (As cited in Varela *et al.* (1991))

Hinton, G. E., J. L. McClelland and D. E. Rumelhart (1986). "Distributed Representations." *Parrallel Distributed Processing* Ed. D. E. Rumelhardt and J. L. McCleland. Cambridge Mass., MIT Press. 77-109.

Kalat, James W. (1988). *Biological Psychology.* (Third edn.) Belmont, California, Wadsworth Publishing Company.

Land, E. H. (1977). "The Retinex Theory of Colour Vision." *Scientific American* 237: pp108-128.

Lythgoe, J. (1979). *The Ecology of Vision.* Oxford, Clarendon Press. (As cited in Varela *et al.* (1991))

Marr, D. (1982). *Vision: A computation investigation into the human representation and processing of visual information.* San Francisco, W.H. Freeman and Co.

Martindale, Colin (1991). *Cognitive Psychology: A Neural Network Approach.* Pacific Grove, CA, Brooks/Cole Publishing Company.

Matthen, Mohan (1988). "Biological Functions and Perceptual Content." *The Journal of Philosophy* **LXXXV** (1, January): pp1-27.

McClelland, J. L., D. E. Rumelhart and G. E. Hinton (1986). "The Appeal of PDP." *Parallel Distributed Processing* Ed. D. E. Rumelhardt and J. L. McCleland. Cambridge Mass., MIT Press.

Millikan, Ruth Garrett (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism.* Cambridge, Mass., MIT Press. A Bradford Book.

Neisser, Ulric (1989). *Direct Perception and Recognition as Distinct Perceptual Systems.* Paper presewnted at the Eleventh Annual Meeting of the Cognitive Science Society, Ann Arbor, MI.
(as cited and explained in Bechtel (1990) and Bechtel & Abrahamsen (1991))

Ornstein, Robert and Richard F. Thompson (1985). *The Amazing Brain.* London, Chatto and Windus.

Peirce, Charles Sanders (1955). "Abduction and Induction." *Philosophical Writings of Peirce* Ed. Justus Butler. New York, Dover publications. 150-156. (a selection from articles written in 1896, 1901, 1903 &1908)

Poggio, J. (1984) "Vision by Man and Machine." *Scientific American* , (April)

Sterelney, Kim (1990). *The Representational Theory of the Mind.* Oxford, Basil Blackwell.

Varela, Francisco, Evan Thompson and Eleanor Rosch (1991). *The Embodied Mind: Cognitive Science and Human Experience.* Cambridge Mass., MIT Press.

Winograd, Terry and Fernando Flores (1986). *Understanding Computers and Cognition: A New Foundation for Design.* Norwood, New Jersey, Ablex Publishing Corporation.

Wittgenstein, Ludwig (1958). *Philosophical investigations.* (Third (1974) edn.) (G. E. M. Anscombe transl.) Oxford, Basil Blackwell.