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.

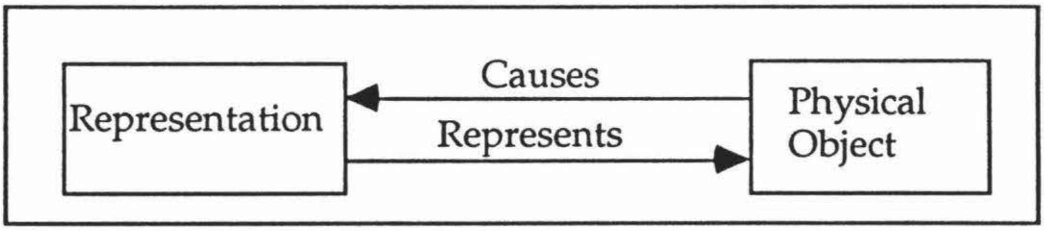

Figure 1.1: Crude Causal Theory's account of representation.

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.

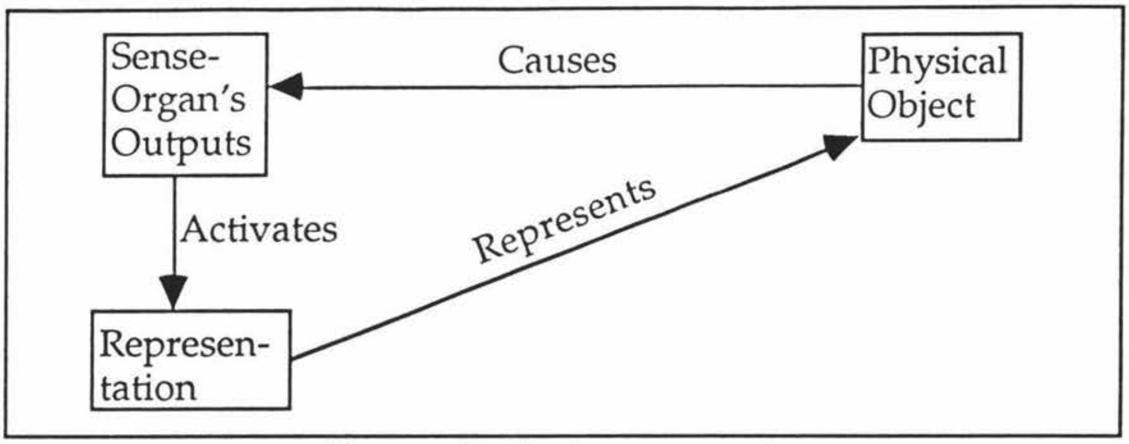

Figure 1.2 Putting the sense organs in.

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.