

Chapter 5

Intentionality: Insides

In direct contrast to German philosophy which descends from heaven to earth, here we ascend from earth to heaven. That is to say, we do not set out from what men say, imagine, conceive, nor from men as narrated, thought of, imagined, conceived, in order to arrive at men in the flesh. We set out from real, active men, and on the basis of their real life-process we demonstrate the development of the ideological reflexes and echoes of this life-process. The phantoms formed in the human brain are also, necessarily, sublimates of their material life-process, which is empirically verifiable and bound to material premises. . . . [We do] not explain practise from the idea but the formation of idea from material practise.

— Marx and Engels, *The German Ideology*

In the beginning was the deed.

— Goethe, *Faust*

5.1 Opening the Black Box

Modern philosophy of psychology started with Freud's (re-)discovery that there is more going on in our heads than we are consciously aware of. Our conscious selves are not masters in their own house, in his patrician phrase. This left us with a problem, since it implies that if you want to understand what is going on in someone else's head then it is not sufficient to just ask them what they were thinking. (And of course the same argument applies to ourselves: *we* do not always know what *we* are thinking.) If you want to do psychology then first person introspection is not enough. Freud's solution to this problem was to develop his theory of the unconscious and the techniques of psychoanalysis, but this ended up as a degenerate form of hermeneutics devoid of any empirical rigour. Freud was a novelist who missed his true vocation, not a scientist. Skinner's response to this malaise was to re-assert, with a vengeance, the primacy of third-person observation. In future all talk about the insides of heads was to be abolished, to be replaced by constructions over directly observable behaviour.

It is hard to overstate the extent to which behaviourism has influenced the subsequent philosophy of psychology. If we define behaviourism purely operationally as acceptance of some form of the Turing Test (1950) then the term includes not just the militant analytical and psychological behaviourism of Carnap, Hempel, Skinner, and Ryle, but also the more sophisticated

empiricist, pragmatist and instrumentalist versions of Quine, Putnam *nouveaux*, Davidson, and Dennett¹. What all these have in common is an agreement that the only way to settle disputes about psychology is by reference to third-person observations of behaviour. In short, we should treat the brain as a black box: we may speculate about what is inside, but never open it up.

This attitude has always seemed quite mysterious to me. Psychology is the science of discovering what is going on in people's heads. Therefore if you want to settle disputes in psychology then surely you should *look* insides people's heads; i.e. try to understand the neurological mechanisms underlying their observed behaviour. It is time to open the box. Chomsky, for example, hypothesised an innate language organ within the brain to explain certain patterns in the way in which we learn and use language. But, as Sampson (1997) argues, this evidence is not, in itself, sufficient to settle the argument one way or another. The same patterns of linguistic behaviour can be explained without recourse to hypotheses about innate language organs. But if we want to know whether the brain contains an innate language organ then surely the obvious strategy is to look inside to see if we can find one? Of course in day to day practice we never know what is going on inside people's skulls. But we should not make a virtue of necessity. After all, if everything worked the way it appeared to then there would be no need for science, as Marx put it. In the last chapter I discussed some of the practical problems with opening the box, and in the next two I discuss some of the philosophical problems.

Of course I am not the first to suggest that we open the box. The most notable exceptions to the behaviourist trend have been Smart and the identity theorists, Fodor, Stich, Block, Kim, and the Churchlands, who all argue that intentional psychological states (beliefs, desires, hopes, fears, assumptions, and the rest) must, by definition, be realised in entities inside the head which *represent* the outside world in some way. The disagreements are then over what kind of thing these internal representational states are, and what it means to say that they 'represent' the outside world. Are they particular neuronal firings, or do they exist at a higher level of organisation like the functional patterns of a computer program? Must they take the form of atomistic linguistic symbols, or can they be more fuzzy and distributed? However all these theorists face a problem: if psychological states are entities inside the head then how can the fact that they represent the world make a difference to the behaviour that they control?

The solution to this problem stems from the fact that in order to understand what is going on inside the head of an agent it is necessary to understand what is going on outside. We have to carve the insides and the outsides of an agent simultaneously in order to understand how its behaviour is produced. This also means that we cannot assume that the agent's environment will be carved in the same way as ours. When I look around my office, for example, I see computers and books and papers and mugs. My cat, on the other hand, only sees things to sleep on and things to eat. Therefore there is no point looking for 'mug' or 'computer' representations inside *her* head. My concepts may not fit her objects, and *vice versa*.

In the next two chapters I try to defend these basic intuitions. In this chapter I concentrate on the inside of heads (i.e. how we individuate representations), and in the next I concentrate on the outsides (i.e. how we individuate objects). The conclusion is a form of realism in which successful thought is based on some kind of correspondence between things in the head and things outside.

¹I don't care whether someone 'really believes' in behaviourism, just as long as they behave as if they do.

Now if one is a realist about anything then it is usually assumed that one must be a realist about physics: after all, physics is the most empirically accurate and successful of the sciences and so the one most likely to be ‘true’. But realism about physics usually carries a lot of Kantian metaphysical baggage about universal mathematical laws, essential and intrinsic types, objectivism, and so on. In short realism seems to imply that there is a determinate list of Objects as they Really Are, and that the point of knowledge is to bring our minds into correspondence with them. In these two chapters I try to present an alternative type of realism that does not carry this baggage but instead is consistent with what Fine (1984) calls the ‘natural ontological attitude’ in which the objects of our everyday lives — trees, washing machines, streams, and the like — are just real as, or even *more* real than, the abstractions of theoretical physics. This type of realism is based on the way that we carve out the objects in the world through our own activity, rather than a correspondence between things-in-the-head and a list of things-out-there revealed to us by the high priests of physics. Of course there is a reality — a world out there — prior to mind, but this world has no essential structure, no fixed set of types. According to this view truth boils down to the fact that some ways of carving the world are more successful for certain purposes than others.

5.2 Anti-Turing

How can we tell what is going on in someone’s head? In other words, what makes a psychological description of them true? In chapter 3 I argued that how we choose to describe something depends on what we want out of our description, and this applies to third-person descriptions of intentional behaviour as much as anything else. For example I could claim that ‘my car doesn’t like to go up hills’. Everyone would know what I meant, and would be able to make certain accurate predictions about its counterfactual behaviour. In this sense it is a perfectly good description. Yet everyone — apart from animists — would agree that it is not ‘really true’.

So what makes an intentional description ‘really true’, and in what sense? In chapter 3 I argued that if we want our descriptions to be ‘really true’ in the sense that it is ‘really true’ that planets follow elliptical orbits, that species evolve through natural selection, gravity acts through centers of gravity, elements are periodic, continents move with tectonic plates, gasses obey the gas laws, and governments are an expression of social forces rather than divine will, then our descriptions should also play an explanatory role. Therefore *if* we want our intentional descriptions to be, roughly speaking, ‘scientific’ then they should not just be empirically adequate, acceptable to our social peers, or even maximally predictive, but should also *explain* how the observed behaviour is produced. (On the other hand, if you regard psychology as a type of hermeneutics or literature or therapy, rather than as a science, then other criteria will apply.)

So, what does it take for a description to be capable of explaining how a behaviour is produced? In section 3.3 I argued that this depends on the status of any theoretical terms that the description uses which, in the case of intentional descriptions, means internal mental states such as beliefs and desires. A behaviourist regards these internal states as *abstracta*, mere constructions over observed behavioural data, whilst for the anti-behaviourist they are *illata*, posited entities which are instantiated in the underlying mechanism of the agent in a discernible way. In other words, if intentional descriptions are to be explanatory then beliefs and desires must be realised in *internal representations*: functional properties, processes or entities of an agent (the representational ve-

hicle) that plays a role in the intentional behaviour of that agent in virtue of the information that it carries about some aspect of the environment (the content)². For example, the rats discussed in section 4.2 were able to successfully negotiate a maze and find their food because the place cells in their hippocampus consistently fired when the rats were in a particular location. This mechanism underwrote their beliefs about their position within the maze, and so when we attribute those rats with a ‘sense of place’ we are really explaining how that behaviour is produced, not just describing the behaviour we observe.

Of course representations can be realised in the brain at levels of organisation higher than that of single neurons. This possibility is often associated with computationalism but I want to avoid using this term, for two reasons. The first is that computationalism usually includes assumptions about languages of thought, symbol systems, and the modularity of mind, and I will later argue why these are not necessary. The second is that there is no contradiction between understanding representational vehicles as computational states, and as states of the underlying physical mechanism. Computational states *are* physical states, just at a higher level of description. Therefore I will talk in general about ‘brain states’, whilst making no claims about their level of instantiation.

But why should our intentional descriptions depend on what is going on inside the head of the agent, apart from fitting into our general scientific ‘idea of the good’? What are the implications for the philosophy of psychology?

The first implication is that treating intentional states as brain states makes them causally efficacious. Behaviourism defines intentional states as constructions over observable behaviour, or as dispositions to behave. And, as with other dispositional properties, problems arise if we understand them in terms of actual or counterfactual outcomes (section 3.3). Recall Carnap’s argument that if the dispositional property of ‘being soluble’ is defined as ‘dissolving when in water’ then the claim that ‘X dissolved because it was soluble’ is tautologous. Similarly, if intentional states are defined solely in terms of behaviour then we are not making a substantive claim when we subsequently cite those states as a cause of that behaviour. Of course the solution is that ‘solubility’ describes a property of a substance *in virtue of which* it dissolves, and intentional states describe brain states in virtue of which behaviours are produced.

For example, if we describe a rat as having a ‘sense of place’ iff it can traverse a changing maze then the claim that ‘the rat found the food source because it had a sense of place’ is tautologous. But if by ‘having a sense of place’ we mean that the hippocampal place-cells of the rat accurately correspond to its location then we have a truly causal explanation of its behaviour. (It was considerations such as these that forced Tolman to abandon Skinner’s behaviourism and laid the foundations for cognitive psychology in the first place (1932).)

Davidson (1980) objects to this line of reasoning. He argues that although statements like “the cause of A caused A” may be uninformative or tautologous, that does not necessarily mean that they are false. However such statements only become informative when it is possible to identify that which fulfills the role of ‘the cause of A’ independently of that description (Morris, 1986). It may be the case that we only discover which substances are soluble by putting them in water, and we may only discover someone’s intentional states by observing behaviour, but we should not confuse the way that we measure a property with the facts in virtue of which an entity holds it.

²So far this looks just like an argument for an identity theory. In section 5.3 I show why it is not.

If intentional states are grounded in brain states then we can also account for *errors*. Suppose that we are unable to produce a coherent intentional description of a system that fully accounts for its behaviour. We have two choices. The first is to revise the list of beliefs and desires that we attribute to the system in an attempt to make the behaviour rationally explicable. Such a retrospective revision is always possible, though possibly at the cost of ascribing wildly implausible intentional states to the agent. For example, suppose I use a pound coin to buy a newspaper that costs 45p, and the shopkeeper gives me the wrong change. Why did he do it? One possible explanation is that he *really* believed that £1 minus 45p was 35p, or that the pound coin I gave him was worth 80p. If so then his reasoning was perfectly rational, but his beliefs were bizarre. (This issue is discussed between Stich (1981) and Dennett (1987, ch4)). This way of accounting for errors is equivalent to the pre-Copernican practise of retrospectively adding epicycles to our Ptolemaic descriptions of planetary orbits in order to get a fit; a practise that produced empirically accurate descriptions but at the expense of vastly convoluted explanations.

The alternative that Dennett discusses is to admit that the system was acting irrationally, but that “mistakes of this sort are slips in good procedures, not manifestations of an allegiance to a bad procedure or principle.” In other words the shopkeeper simply made a mistake. This is equivalent to noting departures from Kepleran ellipses but not abandoning his laws as a result, since these errors can be *explained* as being due to a departure from the usual law of gravity on which they are based. As Dennett puts it,

we must descend from the level of beliefs and desires to some other level of theory to describe his mistake, since no account in terms of his beliefs and desires will make sense completely. At some point our account will have to cope with the sheer senselessness of the transition in any error.

However we can only use the lower level theory to describe mistakes if we can use the lower level theory to describe successes. We cannot use knowledge of the workings of the mechanism to analyse how a system has gone wrong unless we know what it should have done in order for the system to behave correctly³. It is the ability to account for errors in this way that differentiates between systems that are rational but error-prone, and systems that are logical but bizarrely stupid. To err is human, after all.

Of course in one sense the modern behaviourists are absolutely correct: in everyday life we form and judge intentional descriptions on roughly hermeneutic or instrumental criteria. If we can make sense of someone’s behaviour and roughly predict their future actions then this is all that matters. Moreover I am not necessarily arguing that we should change these criteria in practice. Rather it is a question of how we *regard* the intentional descriptions that our hermeneutics generate. We can attribute the empirical success of an intentional description to the role of internal representational vehicles even whilst we have no direct experience of them, in the same way that Kepler could attribute the empirical success of the elliptical orbit to an undiscovered heliocentric force.

Anti-behaviourism does imply, however, that discoveries about the internal mechanism of an agent can affect our intentional descriptions. In one respect this is common sense. For example,

³The problem of differentiating between success and failure will be discussed in chapters 7 and 11; the problem here is to account for the ones that we identify.

the entire plot of *Cyrano de Bergerac* is based around a form of Turing Test, in which Roxanne is fooled into believing that Christian is as smart as he is beautiful by his ability to parrot poetry fed to him by Cyrano. The play hinges on the fact that, were the trick to be revealed, then Roxanne would realise that it was our misshapen hero that she loved, not the dumb Christian. Would the behaviourist argue that Christian really was a poet, just because he passed the Roxanne test? Surely to be a poet it is not enough to produce poetry; one must produce poetry *in the right way*.

Block makes the same point by imagining a machine that uses a crude look-up table to pass a Turing Test (1981). A look-up table works by simply mapping each possible input vector to a suitable output, $\mathbf{L} = \{i_j \mapsto o_j | j\}$. For example, $I = \{i_j | j\}$ may be a complete list of all questions in English of less than 100 words (including mis-spellings), and $O = \{o_j | j\}$ a list of suitable responses. Even though \mathbf{L} would be able to answer any question that we give it, Block argues that if we looked inside the black box then we would realise that it wasn't *really* smart, it just *acted* smart. But he fails to give a reason why such a system is not intentional, despite its ability to pass any Turing Test; and nor does he define a condition on how a mechanism works that would convince him that it were. It may be that *any* discovery about the workings of a brain would lead Block to reject a psychological description — ‘oh look, it's not really intelligent, it's just a bunch of neurons and nerves’ — just as we reject the idiom of magic whenever we work out the conjuror's trick.

The problem with a look-up table is that it does not have any internal states that can act as causally efficacious representational vehicles: it is a pure stimulus-response engine. Therefore, although we may usefully attribute it with beliefs and desires in order to make sense of its behaviour, these internal states do not carve its mechanism ‘at its joints’. However, it may seem that, by carving it in a suitably contrived way, we could re-describe a look-up table such that it apparently operates using internal states without making it any more intelligent. One way would be to create a new set of vectors and incorporate them into the mechanism, $\mathbf{L}' = \{i_j \mapsto s_j \mapsto o_j | j\}$. We could then form an internal pseudo-state by grouping together all those internal vectors, $S \subset \{s_j | j\}$, that subserves acts to which we would normally ascribe a particular belief, b . For example, suppose there is a subset of all questions whose correct answers involve the belief, b , that the earth moves round the sun:

$$\begin{aligned} i_{100} = \text{‘What is the third planet from the sun?’} &\mapsto s_{100} \mapsto o_{100} = \text{‘Earth’} \\ i_{101} = \text{‘Does the earth move round the sun?’} &\mapsto s_{101} \mapsto o_{101} = \text{‘No’} \\ i_{102} = \text{‘Does the sun move round the earth?’} &\mapsto s_{102} \mapsto o_{102} = \text{‘Yes’} \end{aligned}$$

There thus seems to be a well-defined internal state, $S = \{s_{100}, s_{101}, s_{102}, \dots\}$, that subserves the belief b . (Other beliefs, such as ‘the earth is a planet’, may also be involved in answering these particular questions, but they would also be implicated in others. Thus the sets of internal states would be distinct but not disjoint.) This system is obviously no smarter than \mathbf{L} — it has the same procedural semantics — and yet apparently uses perfectly well-defined internal states corresponding to any beliefs and desires that we may ascribe to it. The problem with this strategy is that, in order for an internal state to be accorded a causal role, it must be defined independently of the behaviour that it is invoked to causally explain; but the only thing that identifies the members of S is precisely their membership of S , which is defined by the fact that its members subserves behaviours on the basis of which the agent is attributed with belief b . Therefore a token internal

state, $s_j \in S$, does not have causal powers in virtue of its membership of the set — i.e. being a representational vehicle of a particular type — rather it is of that type *because* of its causal power⁴. Thus it would be inaccurate to claim that L' was able to answer questions about the solar system *because* the relevant internal states were members of S . We are back to claiming that ‘the cause of A caused A ’.

Consider another example. Suppose two children learn to do two-column subtraction. The first uses the look-up table strategy and just memorises each and every sum. The other memorises just the single-column facts ($3-1=2$, $8-5=3$, etc) and works out two-column sums using the rules of borrowing, carrying, and so on. Both children will be able to do the same sums, so it looks like they both must have the internal structures necessary to do subtraction. Now it is true that the look-up child knows how to do each particular subtraction — just as the look-up table ‘knows’ how to produce the right output in response to the right input — but does she know how to do *carrying*? Is the carrying rule one of her beliefs? Surely not, since the correct results were not produced *because* of an internal mechanism that instantiates this particular belief. For example she would conclude correctly that $81 - 35 = 46$, but would not reach this conclusion *because* of the carrying rule. Therefore the two children will show the same behaviour, but this behaviour should be described differently in each case. The differences between the two children are usually hidden, but may show up in the pattern of errors that they make. The look-up child will tend to make random errors as they forget particular answers. But children that learn how to do carrying tend to show clear patterns of error as they misapply particular rules, such as forgetting to subtract one from the tens column (Brown & Burton, 1978).

In short, intentional behaviour is not simply a matter of what something *does*, but also *how* it does it. When we try to work out what is going on in someone’s head — i.e. when we try to ascertain the beliefs, desires, motives and assumptions that lie behind their actions — we are doing literally that. We are not just laying bets about future behaviour, and nor are we necessarily trying to imagine what it is like to be in their shoes (or body), even though all these motives may be involved. Rather we are trying to determine, at a suitable level of description, the functional organisation of the physical mechanism that subserves their behaviour.

5.3 Externalism

Suppose someone avoids being hit by a train and we explain their actions by saying ‘she moved because she thought a train was coming’. In the previous section I argued that this claim is only (‘really’) true if there is a representational brain state that instantiated this belief and causes her to move. But if this is the case then it seems we could just as well say that she moved because that brain state was active; and even though we may choose to describe that state as ‘believing that a train was coming’, the semantic properties are strictly irrelevant to a causal explanation of her behaviour. The same argument applies to the *causes* of beliefs as well as their effects. We would normally say that she thought a train was coming because she heard it. But if that belief is instantiated in a particular brain state then we could equally say that she held that belief because of

⁴Note that this assumes a constructivist approach to set theory since, *contra* Frege, Russell, and Quine, I assert that S does not exist prior to the rule used to construct it. Therefore, in order to count as a causally efficacious internal state, S must be an example of what Frege and Russell called a *class*, rather than a set.

the stimulation of her ear drum, rather than because she heard what sounded like a train. Of course in order to explain how her behaviour is in intentional co-ordination with her environment we would have to supplement this story about the insides of the person with one about the outsides. In order to understand how she avoids trains, for example, we would need to know about the origins of the air vibrations that excited her ear drum. But the crucial point is that the two stories seem to be strictly separable.

This is the point of the brain-in-a-vat thought experiment: suppose that we remove a brain from a living creature, keep it alive in a vat, and connect its nerve endings to a computer that has been programmed to produce the stimuli that would result from bodily interactions with a 'real' environment. Presumably the brain would not be able to notice the difference, which seems to prove that brain events occur according to purely local laws and strictly independently of the environment. Therefore, as Putnam puts it, meanings do not play a role in the head (1981). In other words once we understand how mental states are instantiated in the brain of the agent then we can understand how beliefs and desires can cause behaviour; but the problem is then to understand how their *being* beliefs and desires, how their *having* semantic content, contributes to their causal powers.

This argument applies whether we regard mental states as being instantiated in the brain as computational states, or as neurological ones:

In fact, as far as I can see, if the problems about implementation we've been discussing are real and not solvable, only the elimination of the intentional would be a cure adequate to the disease. For, notice: if the externalist character of content shows that the immediate implementation of intentional laws can't be computational, it also shows, and for precisely the same reason, that it can't be neurological (or subatomic, for that matter). For, neurological states, like computational ones, are individuated by their local properties (roughly, by their parts and to each other). So, presumably there can't be neurologically sufficient conditions for content states if content properties are externalist. So neurological processes can't implement intentional laws if computational processes can't. (Fodor, 1994, p15)

What makes syntactic operations a species of formal operations is that being syntactic is a way of *not* being semantic. Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference, and meaning. ... If mental processes are formal, then they have access only to the formal properties of such representations of the environment as the senses provide. Hence they have no access to the *semantic* properties of such representations, including ... the property of being representations *of the environment*. (Fodor, 1991, p488)

So, rejecting behaviourism seems to imply that we must also reject an externalist account of intentionality — i.e. one in which the semantic properties of our thoughts play a role in our heads. In this section I argue that this implication is mistaken.

5.3.1 Epistemological Externalism

Rejecting behaviourism need not imply internalism. To understand why we have to start by understanding 'representation' as a verb, not a noun. Representation is what a brain state *does*, not what it *is*. Representation is the role that a functional entity plays within the intentional behaviour of an

agent, not a structural component. This makes representations different from the components of most complex systems.

Most artefacts, for example, are produced by putting together pre-fabricated parts. This means that, for example, we can take the starter motor out of one car, put it in another, and the motor will still do its job. Many major biological organs, such as hearts and lungs, are the same. This is what makes heart transplants possible. These types of parts are structurally individuated, but this is not always the case. For example, many invertebrates do not require lungs to breathe since respiration in small bodies can be achieved by diffusion. This does not imply that there are no entities, such as stomata and vesicles, that 'do' respiration; but rather that these entities form a distributed, functionally individuated, 'component' or subsystem, rather than a localised, structurally individuated, one. It would be impossible to transplant the respiratory system from one beetle to another without transplanting the whole beetle. As another example think of the geographically diffuse components of human societies, such as political organisations, social classes and companies etc. We cannot point to a single, discrete, component that performs the function of respiration in a beetle, any more than we can point to the University in Oxford; but this does not mean that these functional components are not (1) well-defined, or (2) physically instantiated in a perfectly intelligible way.

There is no reason why representational vehicles must be discrete components of the brain. They may be more like beetle's respiratory systems than lungs. In other words representations are defined by the role that they play in the overall behaviour of the agent, not physiologically. We cannot know whether something is a representation until we understand the role that it plays in a body in a behaviour in an environment. Therefore, as Peacocke argues (1994), representational vehicles are individuated externally, with respect to their content, rather than internally and narrowly. It is the external relational properties that defines something as a representation in the first place. There is no syntax without semantics, as Crane puts it (1990).

But there is a problem with this weak form of externalism. In order that mental states are robustly causal it is necessary that the representational vehicles that carry them are identifiable independently of the intentional behaviour used to define that state. For example, O'Keefe and Dostrovsky had to use an understanding of the behaviour of the rat to pick out the functional properties of its hippocampus. But once those properties had been discovered they were defined neurologically, in terms of the activation of place cells. Therefore although knowledge of the overall behaviour of an agent is necessary for us to *identify* what neurological states are representational vehicles, the existence of the state that we identify is not so dependent. Some form of reductionism is then possible, at least in principle. McGinn calls this epistemological, as opposed to metaphysical, externalism (1989): external semantic relations may be necessary for us to identify a state as a representation, but these do not play an essential causal role. So it seems that if we want our intentional states to be causal, then they will not be causal in virtue of their semantic properties.

5.3.2 Metaphysical Externalism

Epistemological externalism is a form of pragmatic anti-reductionism. Pragmatic anti-reductionism, if you recall, starts from the assumption that complex systems are made of components which obey

fixed laws independent of the higher properties of the system. Therefore events at the lower level are caused by other low-level events and laws. But in chapter 2 I defended a stronger form of anti-reductionism according to which it is also true that micro-events may be caused at a higher level. For example, it is possible to describe the momentum of a particular gas molecule as being *caused* by the pressure exerted on the wall of the container. There were two possible ways to justify the use of downwards causation as a way of understanding a system depending on whether causation is understood pragmatically or counterfactually. The first was that the description of molecules as rebounding elastically is just a useful approximation; and that if we allow this as a valid description then the downwards causation story is an equally good one. The second was that lower level events are over-determined by lower level causes, and so we can point the causal finger at a higher level: the same act of compression would have produced the same rise in momentum, no matter what particular collisions the molecule experienced.

The same arguments apply to the relationship between brain states and environments. The internalist and epistemological externalist (like the reductionist and pragmatic anti-reductionist) both assume that it is possible to determine the future behaviour of an agent from the neurological laws governing its nervous system and the stimulation of its sensory nerve endings. However, as we saw in section 4.1, neural mechanisms can function differently in different behavioural contexts. There are no neurological laws *simpliciter*; neural stuff only ever exists in a body in interaction with an environment. This is why the visual system of the Horseshoe crab could only be understood by studying the whole animal in its natural environment, rather than by taking *in vitro* measurements from a lab preparation. It is also why neuroethologists spend so long fitting locusts with radio back-packs.

In other words, any statement of neurological law should include the rider "...in such-and-such environmental and behavioural circumstances", since the same neurological stuff may act differently in different circumstances. Therefore even if we can discern the particular neural organisation and processes underlying an intentional act, this does not threaten the intentional description since those neural facts are only true *because* of the wider environmental and behavioural picture, and it is this level of organisation that the intentional description refers to. The neuronal organisation of an organism is a result of its overall behaviour within an environment, as much as *vice versa*. Of course, much of the structural neurological properties of an organism *are* well insulated against modulations caused by environmental impact — and so internalism is often a very good approximation — but the point is that there is always the potential for the outside world to have an impact; and this is all that metaphysical externalism requires.

Metaphysical externalism can also be cashed out in terms of counterfactuals. The key point is that, as we saw in the case of rat navigation, a behaviour is intentional only to the extent that it is not dependent on particular stimulus-motor responses — the rat, for example, could find its food despite changes to particular landmarks or the flooding of the arena. Thus its hippocampal place-cells fire, and have their effects on behaviour, *because* the rat is in a particular location, rather than because of particular retinal stimulation. Indeed, the inherently unreliable and noise-ridden nature of biological nervous systems means that creatures have evolved such that regularities at the intentional level (such as finding food) are preserved *despite* the failure of particular local regularities (such as a receptor cell firing when illuminated). As with other stochastic systems,

higher order may arise from lower disorder.

Metaphysical internalism is true of an ideal brain just as the Gas Laws are true of ideal gases, but that does not mean that it is true of real, living, metabolising, brains, embodied in bodies interacting with their environment. If real gases were ideal then it would always be possible to eliminate a downwards causal story in favour of a lower level description (if we wanted to). Similarly, if real brains were like neural network models then it would always be possible to eliminate an intentional description in favour of a neurological or syntactic one (if we wanted to). But they are not. Internalism and the gas laws are both good approximations *in certain conditions*; after all, if the gas laws were true *simpliciter*, then gases would never condense. The crucial point is that in both cases the accuracy of the lower laws, and thus the counterfactuals that they support, are dependent upon those higher conditions: the external force on the container wall in the case of the gas laws, and the behavioural environment in the case of neurological processes.

In chapter 2 I used the example of gas condensation to show how changes at the higher level (i.e. raising pressure and lowering temperature) can cause a drastic change to the rules governing the behaviour of the parts (i.e. whether the molecules rebound elastically), and so reveals how the latter are dependent on the former; a dependence that is often disguised in ‘normal’ circumstances. The parallel example in the case of intentionality is to consider a behaviour that not only involves the stimulation of nerve-endings, but also changes the way the central nervous system works. Take psychoactive drugs. Suppose we drink a glass of whiskey and, due to the slight intoxication, entertain the belief that we are over the legal driving limit. Now the sense of intoxication is not produced by particular sensory stimuli, but rather by the way that alcohol enters the bloodstream and is distributed throughout the body, subtly altering the electrical and biochemical properties of potentially every single neuron in the entire central nervous system. The spinning sensation, for example, does not come from our taste buds, but is due to the thinning of the blood in the ear canals which disrupts the neutral buoyancy of the cilia motion detectors. In this case we have quite literally taken the external object — the alcohol — and put it inside our heads. This puts internalism in an awkward position:

Internalism implies that the inferences we draw from a belief only depend on whether we think it is true, not on those facts that make it true or not. But drinking alcohol does just not produce the belief that we are drunk, but also affects the cognitive consequences of that belief. If we were stone-cold sober, but for some reason wrongly believed that we were over the limit, then we would conclude that we were unsafe to drive. But if we believed that we were over the limit, *and really were*, then we would be more likely to rashly conclude that we were perfectly safe. In other words, the state of affairs that make the belief ‘I am over the limit’ true are precisely those that affect how the belief is processed. The syntax of real living cognitive systems is causally, and not just epistemologically, dependent on semantics. To put it another way, how could you convince a brain-in-a-vat that it were drunk apart from adding some alcohol to the vat?

Now a metaphysical internalist may object that, in such cases, although the fulfillment of the truth conditions of the belief may effect the operational consequences of the tokening of the representational vehicle that realises it, they do not have these effects *qua* satisfaction of the truth conditions. In other words, the presence of the alcohol in the blood may effect the consequences of my belief about it, but not *because* my belief was about the alcohol. Indeed the presence of

the alcohol will effect many other cognitive processes, and not just those that involve the belief that we are drunk; and conversely many other environmental events that are completely unrelated to the content of the belief, such as a bang on the head, may also affect the processing of that belief. The externalist response is that this objection forgets that representational states are also epistemologically external. In other words, although a representational vehicle must be a physiologically-defined brain state in order to be capable of playing a well-defined causal role, it is the external relationships that make the state a representation. It is the ability of a brain state to reliably carry information about the presence of alcohol in the blood that defines it as the vehicle that instantiates the conviction that we are drunk. A belief may be affected by the presence of alcohol even if it is not reliably correlated with it, but in this case we would not describe it as the belief that I am drunk. Therefore the syntax of the representation is affected by its semantics *qua* its semantics.

Such drug-induced cases may seem like extreme examples. But often when an entity appears to be independent of its environment then the only way to reveal the dependence is to consider extreme cases, like the gold object in *aqua regia*. Of course, in many cases, the internalist assumption is approximately correct, but we should never forget that it is only an approximation. Representational brain states occur, and have their causal consequences, not solely according to local, syntactic, laws, but also because of external, environment-involving, facts. Although representational vehicles may be *in* the head they, and their causal powers, are a property *of*, and dependent upon, the entire agent-environment system. The externalism of representational vehicles is metaphysical, and not just epistemological — but this does not require any kind of spooky action-at-a-distance. Of course all interactions between things-in-the-world and things-in-the-head are mediated *via* local biological connections obeying local biological ‘laws’. Rather, metaphysical externalism rests on the fact that these ‘laws’ only hold *because* of the larger intentional, environment-involving, picture. In different behavioural environments we may find that new ‘laws’ apply.

Dennett recalls that, when considering the role of the brain in intentional behaviour, the first fundamental conclusion he came to was that

the only things that brains could do was to *approximate* the responsiveness to meanings that we *presuppose* in our everyday mentalistic discourse. When mechanical push came to shove, a brain was always going to do what it was caused to do by current, local, mechanical circumstances, [regardless of] whatever it *ought* to do, whatever a God’s-eye view might reveal about the actual meaning of its current states. But over the long haul, brains could be designed — by evolutionary processes — to do the right thing (from the point of view of meaning) with high reliability. . . . brains are *syntactic engines* that can mimic the competence of *semantic engines*. (1998a, p357)

But brains are *not* syntactic engines. They are living biological entities, enclosed in bodies and coupled to environments. Brain tissue can mimic the competence of syntactic engines — or rather we can build syntactic engines, such as artificial neural networks, that mimic them — but this is just an approximation, just as much as a semantic engine (i.e. an intentional description) is. Brains ‘are’ syntactic engines to exactly the same extent that they ‘are’ semantic engines, and we can no more eliminate the latter than we can the former.

5.3.3 Brains-In-Vats

It may be useful to reconsider the same problem in terms of a brain in a vat. The situation, if you recall, is that the brain has been removed from a living creature and kept it alive in a vat, with its nerve endings connected to a computer that has been programmed to produce the stimuli that would result from bodily interactions with a ‘real’ environment. The internalist, and epistemological externalist, argues that this demonstrates that those brain events occur according to purely local laws and strictly independently of the environment.

Now the issue here is *not* whether it is possible to fool brains about the nature of their world. We do not need such high falutin’ thought experiments to realise that this is possible. Rather the issue is the nature of the dependency between brains and the world and, in particular, the internalist insistence that the dependency stops at the interface between the two. They argue that what goes on inside the head is only dependent on what happens at the sense-organs (or the socket at the back of the vat), and that more distal facts about the environment are strictly epiphenomenal. This is equivalent to claiming that, not only could we program the computer to convince the brain that it is interacting with a ‘real’ world, but that we could produce the appropriate stimuli such that the brain’s internal processes continue *just as they would have* in a real environment.

The metaphysical externalist response to such examples is to ask how the computer was programmed in the first place. How does the scientist know what sequence of stimulations to produce in response to the motor nerve outputs of the brain? The starting point for producing such a program would be to investigate the structure of the brain, central nervous system, body, and environment of the agent, and then produce an accurate model of this data in order to generate the appropriate stimuli in response to the brain’s motor outputs.

Now the internalist claim is that, using this method, we can produce a brain-in-a-vat (BIV) that simulates what would have happened had that brain remained in a body-in-a-world (BBW, also known as a person) — i.e. the computer environment of the BIV uses a model based on data collected from the BBW, such that if they were started off in the same state then their future internal activity would march in step. However the scientist’s model is based on restricted observational data, taken when the BBW was performing particular behaviours in particular environments, and *we cannot assume* that this model will prove perfectly predictive about what happens in others. Of course, in practice, it may. Nonetheless there is always the chance that some of the properties that our model assumes to be constant turn out to be variable, and thus that behaviour of the two systems will diverge. Thus we only have a guarantee that the behaviour of the BIV will match that of the BBW to the extent that they replicate the behaviour that was measured *in the real world*. Recall Feynman’s insistence that

science is uncertain; the moment that you make a proposition about a region of experience that you have not directly seen then you must be uncertain. But we must make statements about the regions that we have not seen, or the whole business is no use ... We have to make guesses in order to give any utility at all to science. In order to avoid simply describing experiments that have been done, we have to propose laws beyond their observed range. There is nothing wrong with that, despite the fact that it makes science uncertain. If you thought that science was certain — well, that is just an error on your part. (1965, p76)

Outside of these certain situations the BIV may continue in blissful ignorance that what is now

happening may not be what would have happened if it were connected to a ‘real world’ instead of a computer. But that is not the point, which is rather that, in order to ensure that the two systems stay in step, the scientist has to go and check the model against the original BBW, and make adjustments as and when necessary. Therefore the exact sequence of internal events in the BIV *is* dependent on the external world of the BBW, even though this connection is not mediated directly through the senses (as it is for the BBW), but indirectly *via* the measurements and tinkering of the scientist.

Putnam used the example of the BIV to show how what goes on in our heads is, in a strong sense, independent of what happens in the world. This conclusion follows naturally from the assumption that *all* objects (including neural mechanisms) are, in strong sense, independent of their environment. Once we replace this Kantian assumption with the anti-reductive materialism I outlined in chapter 2, then the natural conclusion is that what goes on in our heads *is* irreducibly dependent on what happens outside.

5.4 Emergent Representation

If you want to know how brains produce intelligent behaviour then the usual explanation, which we inherited from Descartes, goes something like this. There are two separate systems: an agent and an environment. The agent contains representations which are manipulated according to the laws of neuroscience and/or syntax, and the environment contains objects which are governed by their own laws. These two systems are then linked by sensors and motors. The problem for Descartes, and all subsequent representationalists, has been to explain how the content of the representations — i.e. their relationship to the world — play a role in the head. The artificial intelligentsia simply assumed the problem was unimportant and so came up with machines that were able to manipulate representations ’til the cows came home, but which had no essential connection with the things that they were supposed to thinking *of*.

In this chapter I have tried to present a solution to this problem. The trick is to start by regarding the agent and its environment as a single system, not two separate but connected ones. Representations are an emergent property of this whole system, rather than a part of one of them, and they are emergent in both a weak and strong sense. Representations are emergent in the weak sense because a brain state is only defined as a representation with respect to the whole system (epistemological externalism). And representational brain states are emergent in a strong sense because they, and their causal powers, are *dependent* on the whole system (metaphysical externalism). Therefore if you try to take an agent out of its environment (*à la* AI) then, strictly speaking, it doesn’t contain any representations at all⁵.

This approach to the problem of representation affects how we understand the relationship between intentionality and information. I, like many other theorists, especially since Dretske (1981), use an information-theoretic definition of representation in which the representational nature of a vehicle rests on its ability to carry information about an external state. However such theories often give the impression of intentionality being *reducible* to the processing of representations; that representations are the atoms of intentionality and when you put enough together you get about-

⁵This solution to Descartes’ problem may seem to have the advantages of theft over honest toil, in Russell’s phrase, but I prefer to see it as *liberation* rather than theft.

ness. But if we regard representations as emergent from behaviour, then this picture gets turned on its head.

Information is everywhere — wherever there is cause there will be correlation, and wherever there is correlation there is information — but it is usually causally inert. Information carriers have causal powers, but not in virtue of the information they carry. If, however, the effect for which we seek a cause is the co-ordination of an agent with respect to its environment — i.e. an intentional behaviour — then the property of the representational vehicle capable of having this effect is precisely the fact that it carries reliable information about the external object. Information only becomes a causal property in the context of intentional behaviour. Aboutness does not flow upward from information-carrying representations to intentional behaviours, but is rather bestowed from above. Just because we have found something in the head that bears information about an external object this does not yet make it a representation. Representation is only happening if that information plays a causal role in the behaviour of the agent. Representations are not the atoms of intentionality that can be glued together to make intelligence, but defining properties of an agent that is able to act intentionally.

A behaviour is ‘really’ intentional (*sensu* chapter 3 and section 5.2) iff it is mediated by representations; but an information-carrier is only a representation if it plays an appropriate role in an intentional behaviour. This may seem circular, but the point is that each level of organisation (i.e. brains and behaviour) can only properly be understood in the light of the other. We can only make sure progress in psychology if we cease to regard the brain as a black box. Conversely, we can only understand brains in the light of an understanding of the behaviour that they underlie. It is this latter point that marks the difference between naturalising intentionality, and reducing it: to *reduce* an intentional description is to show how it can be derived from a set of independent lower-level facts; whereas to *naturalise* an intentional description is to show how it is systematically related to one below. Naturalisation implies reduction unless the lower level is also dependent on the higher, which is what externalism implies.

Behaviourists regard beliefs and desires as constructions over behavioural data, and nothing to do with events in the brain *per se*. Behaviourism implies that if two agents exhibit the same behaviour then they must have the same beliefs, even if those behaviours are produced by non-isomorphic mechanisms (remember the look-up child and the carrying child). Identity theorists and computationalists, on the other hand, reduce beliefs and desires and claim that they simply *are* brain states (at the appropriate level of description). My alternative is that intentional states are the *role* that brain states play *within* behaviour. They are not properties of brains and they are not properties of behaviours, rather they are the relationship between the two. To entertain a belief is to possess a brain state whose information-carrying properties allow you to achieve certain types of interaction with the world. And neither clause in this definition can be amputated: if you omit the first, then you just have something whose behaviour *appears* to involve belief; and if you omit the second then you don’t have a belief, just a brain state.