

## **PART II: MIND**

**In the next three chapters I apply the general conclusions of the previous Part to the problem of brains and minds; i.e. the relationship between individual psychology and the neurological mechanisms that underly it. I concentrate in particular on *intentionality*; i.e. the way in which our thoughts can be *about* the outside world, and how this enables us to interact with that world in meaningful ways. In chapter 4 I start by looking at some recent developments in the way we do neuroscience, and how this is reflected in how we try to build artificial intelligences. And in chapters 5 and 6 I apply these empirical results to some of the key issues in the philosophy of mind, including behaviourism, representationalism, mental causation, externalism, and the nature of content. This may seem to be putting the empirical, *a posteriori*, cart before the philosophical, *a priori*, horse; but my aim is to loosen the grip of some assumptions that are responsible for needless philosophical problems, and so seeing how thinking things work in the flesh seems a good place to start.**

## Chapter 4

### Brains and Behaviour

---

She looked liked she learned to dance,  
From a series of still pictures.  
— Elvis Costello, *Satellite*

#### 4.1 Neuropsychology and Neuroethology

There are two strategies for tackling a really difficult problem. The first is to break it down into discrete sub-problems, solve each of these in isolation, and hope that the partial solutions can be added together to form an explanation of the whole. Many problems are suited to this approach; it underlies our spectacular progress in developing new technology, for example. This success has encouraged its application to many other areas, including the philosophy of mind and cognitive science. Thus the phenomenal complexity of human thought is broken down into the sub-problems of perception, language-use, logical reasoning, concept formation, emotions, motor co-ordination, associative learning, social intelligence, etc, and each of these ‘modules’ are then studied and analysed separately.

The alternative strategy is to start with the simplest possible example of the whole phenomenon, endeavour to understand it as a unified whole, and then consider more and more complex examples, noting qualitative and quantitative changes in behaviour as we do so. There is good reason to believe that cognition is better suited to this type of approach — after all, the strict modularity assumed by cognitive science bears little relation to how brains evolve, develop, learn, or are used in practice. The conclusion is that we should not start by considering isolated competencies of large, cognitively complex, creatures, but rather we should start by considering the whole of simple ones. *Why not the whole iguana?*, as Dennett put it (1978)<sup>1</sup>. Iguanas may lack many of our higher cognitive functions. They are probably not even conscious. Nonetheless they can negotiate a complex environment, find food and mates, avoid predators, and so on. They are a simple, complete, example of an intentional system, and so seem like a good starting point — both for our attempts to study natural cognisers, and also to engineer artificial ones. Therefore chapters 4–5 are mostly concerned with only the simplest types of intentional activity of both animals and

---

<sup>1</sup>Darwin was perhaps thinking along the same lines when he claimed that ‘he who understands baboon would do more toward metaphysics than Locke’.

robots, and I only start to consider 'higher' linguistic abilities in chapter 6.

It is also worth remembering that the vast majority of *human* behaviour is similarly basic. Without the ability to navigate and manipulate our environment, humans would not be able to support higher cognitive functions, either individually or socially. The thin layer of conscious, linguistic, reflective icing tops a very thick practical cake:

It is instructive to reflect on the way in which earth-based biological evolution spent its time. Single-cell entities arose out of the primordial soup roughly 3.5 billion years ago. A billion years passed before photosynthetic plants appeared. After almost another billion and a half years, around 550 million years ago, the first fish and vertebrates arrived, and then insects 450 million years ago. Then things started moving fast. Reptiles arrived 370 million years ago, followed by dinosaurs at 330 and mammals at 250 million years ago. The first primates appeared 120 million years ago and the immediate predecessors to the great apes a mere 18 million years ago. Man arrived in roughly his present form 2.5 million years ago. He invented agriculture a mere 19,000 years ago, writing less than 5000 years ago and 'expert' knowledge over the last few hundred years.

This suggests that problem solving, language, expert knowledge and application, and reason, are all pretty simple once the essence of being and reacting are available. That essence is the ability to move around in a dynamic environment, sensing the surroundings to a degree sufficient to achieve the necessary maintenance of life and reproduction. This part of intelligence is where evolution has concentrated its time — it is much harder. (Brooks, 1991)

Traditional philosophy of mind has, like a spoilt child, tried to pick the icing off the cake. It has concentrated on our ability to contemplate the world in isolation from our more fundamental and precious ability to act on and within it. And unfortunately this attitude has been encouraged by the development of powerful brain imaging techniques such as CAT, PET, and especially NMR (Tootell et al., 1995) that can map neuronal activity across the brain while the (usually) human subject remains perfectly still and performs a simple psychological task. The flood of data that these experiments generate is very impressive, but it is still unclear whether it is particularly *useful*. Although correlations between psychological state and brain activity can be demonstrated, the link is not made intelligible. There is no sense of an explanation of *why* or *how* the brain activity produces the psychological phenomena. Nor can these experiments tell us whether the correlation is significant or epiphenomenal. Lesioning experiments and the study of aphasics may be useful in this last respect, but they do not tell us what aspect of the activity of the disrupted tissue was important, nor why.

The missing explanatory link is *activity*. Explaining how brains work requires understanding how brain states play a causal role in behaviour, not just observing correlations. This is the aim of neuroethology. And in order to understand how brain states can play such a role it is necessary to discover their relationship to the rest of the central nervous system of the animal, from sensor to muscle, and *via* feedback from its environment. The fact that the brain has a body is true, obvious, important, and usually ignored. Understanding the bodily and environmental context of the brain can change our picture of what it is doing:

These observations can be summarised using two contrasting musical metaphors. The nervous system is often seen as the conductor of the body, choosing the program

for the players and directing how they play. The results reviewed above suggest a different metaphor: the nervous system is one of a group of players engaged in jazz improvisation, and the final result emerges from the continued give and take between them. In other words, adaptive behaviour is the result of the continuous interaction between the nervous systems, the body and the environment, each of which have rich, complicated, highly structured dynamics. The role of the nervous system is not so much to direct or program behaviour as to shape it and evoke the appropriate patterns of dynamics from the entire coupled system. As a consequence one cannot assign credit for adaptive behaviour to any one piece of this coupled system. (Beer & Chiel, 1997)

If we want an explanation of the sound of an orchestra we have only to look to the conductor and the score that they are following. The characteristics of the individual musicians are relatively unimportant. But if we want an explanation of the performance of jazz ensemble then no player can be ignored. This is not to imply that such an understanding is impossible, but that it cannot be reduced to being the responsibility of a single isolated element. Similarly, if we want an explanation of how the neural mechanisms of a creature subserves its behaviour, then it is not enough to just observe the activity of a single part, but rather we must understand how it is coupled to the rest of the system — i.e. its body and environment.

To take a simple example, the periodic limb movements involved in most forms of animal locomotion are often presumed to be due to internal central pattern generators which propagate centrifugal signals which control the muscles and limbs. But this ignores the role of environmental feedback in generating the overall activity. For example, if you take a lamprey out of water then the change in resistance means that the same stimulation of the muscles produces a completely different wriggle; therefore recording the output from the central pattern generators in its spinal ganglia will give you only part of the picture. Without understanding the properties and role of the water it is impossible to understand how a lamprey swims.

The problem with neuroethology is that it is very hard. Instead of studying isolated parts of the brain of a creature it is necessary — at least in principle — to understand its entire central nervous system, body and environment. This has tended to limit the growth of neuroethology. As Dawkins points out (1995), there is still a large gap between neurobiology and ethology, maintained by the fact that the two disciplines tend to study different animals and ask different questions. Ethologists gravitate towards large intelligent animals — including humans — with interesting and complex behaviours but with poorly understood neurobiology, whereas neurobiologists favour sea slugs and leeches that ethologists find boring. The most interesting work in neuroethology has occurred somewhere in the middle, with creatures that are large enough to display interesting behaviours but are still simple (and disposable) enough for investigation of the entire neural pathway to be possible: prey-catching in frogs and toads, echolocation in bats, auditory source location in owls, and so on.

The difficulties of neuroethology are made even worse by the fact that it is not enough to investigate the *whole* iguana (or bat, frog, or owl), but that they must also be investigated *the right environmental and behavioural context*. For example, over a period of 50 years from 1926 the visual system of the Horseshoe crab became one of the most thoroughly investigated neurophysiological systems in the animal kingdom. However it was not until the late 1970's that it was discovered that the way that the retina reacts to light follows a circadian rhythm, becoming a mil-

lion times more sensitive at night in order to aid mate detection. This crucial functional property of the nervous system had remained undiscovered whilst the visual system had been investigated in *in vitro* isolation as a lab preparation; instead it required taking measurements from a whole live animal in its native conditions of shallow coastal water at night (Barlow et al., 1984)(Barlow et al., 1986).

The need to get the behavioural environment right can stretch the ingenuity of scientists to the limit. Consider the problems of investigating the neuroethology of locust flight. Some progress had been made by taking microelectrode recordings from paralysed insects, but such artificial conditions tends to produce artefactual results. The only solution was to taking recordings from locusts *while they are flying free*, and this required implanting the insects with microelectrodes that would not disrupt their movements and connected to miniature radio transmitters tied to their backs (Kutsch et al., 1993) — a painstaking, intricate, and very frustrating process. Nonetheless, experiments conducted *in* environmental and behavioural *situ* can yield neurological data that it would not be possible to derive, even in principle, from non-situated investigation. A bird in the bush is worth two in the hand, neuroethologically speaking.

Beer, amongst others, argues that the problem of neuroethology is to understand how central nervous systems are coupled to environments *via* bodies. But the situation is actually more complicated than that. If it were simply the case that behaviour is generated by the coupling between a nervous system and an environment then it would be possible, at least in theory, to study the organism in isolation and then try to determine the result if it were put into a particular environment<sup>2</sup>. The more fundamental problem is that this coupling can change the intrinsic properties of the central nervous system itself. The Hodgkin-Huxley model of neuronal activity, which models neurons as discrete ‘units’ with fixed electrical responses, has been very successful. But this success should not make us forget that neurons are living cells whose seemingly intrinsic properties are affected by the metabolism and biochemistry of the entire body and its environment. In some cases there are clearly stereotyped reflex behaviours — such as escape responses — in which the strong evolutionary pressure to favour fast and reliable performance produces dedicated neural structures with very stable and clearly defined properties. However it is now becoming apparent that modulators — hormones, diffuse neurotransmitters, and less obvious agents such as nitric oxide synthases — can alter even the most seemingly fixed and apparent properties of individual neurons (Harris-Warrick & Marder, 1991):

The effects of modulatory substances can be so profound that cells acquire entirely new properties not seen in the absence of the modulator. The effects of modulators covers the range of intrinsic properties, including increased or decreased excitability, the modulation of spike frequency adaption, the enhancement of post-inhibitory rebound, the induction of plateau potentials, and the expression of intrinsic bursting. (Getting, 1989)

These kinds of modulatory processes are usually ignored when constructing artificial neural networks which model biological neural systems using the formalism of systems theory. These models have often been criticised by biologists for being too simplistic. This may be true, but if this

---

<sup>2</sup>For a formal analysis and experimental demonstration of how we can do this for an artificial nervous system see (Jakobi, 1997).

were the sole problem then it could be solved by increasing their complexity and accuracy — as has been done in many cases (Lansner & Liljenström, 1994). The more fundamental problem is that virtually all such models assume a fixed neural structure, comprised of units with fixed electrical responses and connections, or ones that change irreversibly through incremental learning<sup>3</sup> — and this assumption is rarely true.

For example Soffe (1993) describes how the same set of neurons drive both the swimming and struggling behaviours in *Xenopus* tadpoles. A tadpole that starts by swimming may, depending on its environment, encounter a predator. This sensory stimulation has the effect of modulating the synaptic connections and intrinsic properties of the motor neurons in its spinal cord, with the result that it starts struggling. Note that this is *not* just the effect of new stimuli provoking new responses, but rather involves a reorganisation of the neural system that subserves behaviour. Therefore in order to properly understand these events we first have to understand the neuronal organisation underlying the initial swimming behaviour. We then have to understand how that behaviour, in that particular environment, results in the creature being threatened. Lastly we have to understand how this results in changes at the level of individual neurons as it starts to struggle. There is thus a dialectical causal cycle, from neuroscience to intentional behaviour *and back again*. The properties of nervous systems are an emergent product of behaviour, as much as *vice versa*. So, for example, if we were to study a *Xenopus* embryo in a lab preparation we would not uncover the mechanism responsible for struggling, nor that for swimming, but rather some biochemical mish-mash of the two. The neurological roots of its behaviour would remain a mystery.

Different behaviours produce, and are produced by, different neurological organisations. You cannot study an organism in one context and be sure that even the most intrinsic neural property that you discover will persist in another. In short, if you want to understand how the brain of an animal works, you have to study it in an appropriate behavioural environment. And there is simply no way round this.

## 4.2 Representation and Explanation

Neuroethology is a dialogue between neuroscience and ethology, born from the conviction that each must be understood in the light of the other. This implies that your neuroscience will depend on your ethology: your understanding of how a neural mechanism subserves behaviour will depend on how you understand that behaviour. For example Hoyle, in his manifesto for neuroethology (1984), assumes a traditional Lorenzian ethology, complete with Fixed Action Patterns, psychohydraulics, displacement acts and releasers etc. Therefore his neuroscientific investigations concern such issues as the neural mechanisms underlying variations in internal drive and motivation.

However, as many of the peer reviews to Hoyle's article point out, there is a lot more to animal behaviour than those aspects considered by Lorenz. Most neuroethology is concerned with behaviour — or rather *aspects* of behaviour — that should properly be classed as intentional; i.e. those in which internal states are attributed to a creature in order to understand how its behaviour is co-ordinated with respect to objects in the environment. The neuroethological problem is then to explain how this co-ordination is subserved by the central nervous system of the agent; and to do

---

<sup>3</sup>Though see (Husbands, 1998) for an interesting counterexample

this we must find some neurophysiological property that is capable of explaining how this intentional aspect of the behaviour is achieved. If the behaviour to be explained is defined with respect to a distal object, then the mechanism that produces it must be understood in the same way. The explanans and explananda must share some common vocabulary in order for the connection to be made intelligible, and a common term that relates behaviour and mechanism is *representation*, by which I mean the way that a functional property, process or entity of an agent (the representational vehicle) plays a role in the intentional behaviour of an agent in virtue of information that it carries about the object (the content).

For example, suppose we are trying to understand how rats manage to relocate sources of food in a laboratory arena — which they can do despite the experimenter's attempts to confuse them by moving landmarks, or even flooding the arena and forcing the animal to swim. This ability cannot be explained by simply mapping neuronal connections from sensory stimuli to motor responses, since both the stimuli and responses will change as the experimenter changes the arena. Rather an explanation must reveal how the rat achieves a 'sense of place' by integrating many sources of information, including recognising landmarks and its sense of its own movement. And a vital part of this was the discovery by O'Keefe and Dostrovsky (1971) that certain hippocampal neurons are selectively active as the animal moves between different locations in an environment — so called 'place cells'.

Now a great deal remains unknown about the role of the hippocampus in spatial navigation, despite a huge amount of empirical investigation (see (McNaughton, 1996) and (Knierim et al., 1995)). We must admit that, although we know that there are striking correlations between the animal's perceived location and particular neural activity, we do not know how those correlations fit into the entire sensory-motor system of the rat. For example, one of the most perplexing problems is how the same area of hippocampus can serve as a map for many different arenas simultaneously, depending on other contextual cues. Indeed it is quite possible that once we get the bigger picture we will find that the simple place-cells that we naively thought played a role are just some epiphenomenal by-products of a more complex, higher-level, picture. Nonetheless, unless and until these problems are solved we will not have a proper explanation of how the rat navigates its environment.

Place-cells are an example of how a single neuron, or localised group of neurons, may play a representational role (Barlow, 1972). But there is no reason why this should be the case in general. At the start of the last century Sherrington argued that behaviourally significant aspects of neuronal activity may be organised at a higher level than that of the single neuron (1906). For example Freeman (1985) has demonstrated how oscillations in the vertebrate olfactory bulb involving up to a quarter of a million neurons can encode odorant information. These oscillations have a dominant frequency typically around 40-90Hz, but the refractory period of a typical neuron restricts it to producing action potentials at around 5-10Hz. Therefore the bulbar oscillation must be the result of *co-ordinated activity across the entire bulb*; it cannot be a purely epiphenomenal aggregate effect. For each individual neuron the only thing oscillating at 40-90Hz is the *probability* that it will fire, since it can *actually* only produce an action potential every 10 cycles or so. The large-scale oscillations emerge from the mass action of the whole, but in turn they entrain the activity of the individuals (Faith, 1995).

Odorant information only exists at a level of organisation much higher than that of the single neuron. Indeed there is no reason in principle why a representational vehicle could not be a state or process defined over an entire central nervous system, in the same way that the pressure of a gas is subserved by an aggregate property defined over all its constituent molecules. But at whatever level of organisation we discover them, representations are a necessary term in an explanation of how a neural mechanism produces intentional behaviour. Unless we can understand how the organism represents aspects of its environment we cannot understand *how* its behaviour is coordinated with respect to those aspects, we only know *that* it is. Of course, barring miracles, there must be an explanation of how specific stimuli provoke specific responses. But this does not provide an explanation of the intentional behaviour *per se*; only representations can do this. This issue will be discussed in more philosophical detail in the next chapter, but the same point has also recently taken a more practical form.

### 4.3 South Coast AI

Dretske once claimed that ‘if you can’t make one, you don’t know how it works’, and theories about how intelligent behaviour is produced have always been tested in the tribunal of construction. So, for example, computationalism as a theory of mind naturally led to computationalism as a way of building artificial intelligences: an interdisciplinary research program that was born at the famous Dartmouth Conference on the East coast of the US in 1956.

The cornerstone of computationalism is that intelligence is necessarily grounded in a formal symbol system or language of thought — a direct descendent of Frege’s insistence that the starting point for a philosophy of mind is the formal study of language. Computationalism therefore implies that the key to building an artificial intelligence is a system that manipulates symbols in the right way, as enshrined in Newell and Simon’s Physical Symbol System Hypothesis (1972). If the computationalist wants to build a robot that can physically interact with the world then the trick is to connect the symbol manipulator to distinct perceptual ‘modules’ that generate symbolic representations of the world which are then manipulated syntactically to produce a set of symbols representing a plan, and this is then transformed into physical movements by the motor modules (Fodor, 1983). The sensory and motor modules are ‘the stupidity in the system’ (Karmiloff-Smith, 1994), while the real intelligence resides in the symbol manipulation. According to this view, the sensory and motor links to the outside world can be eliminated and cognition understood as a purely disembodied phenomenon; hence the inputs and outputs to most AI systems are symbols with no essential connection to the states of the world they are supposed to represent.

However, practical problems within AI raise concomitant doubts about computationalism as a theory of mind. In particular, although AI has been spectacularly successful on tasks (such as playing chess) that humans find very difficult, it had been relatively unsuccessful on tasks (such as simple social language use and sensory-motor co-ordination) that humans find very easy. The first anti-AI wind blew from Berkeley in the West with the publication of the Dreyfus brothers’ *What Computers Cannot Do* (1972). This challenged the fundamental assumptions of Anglo-American analytic philosophy on which computationalist AI was built, and pointed to an alternative philosophical tradition that included the existential phenomenology of Heidegger and Merleau-Ponty and the anti-logicism of the later Wittgenstein. This critique of East Coast AI was soon joined by

others that took ideas from biology (Maturana & Varela, 1980) (Winograd & Flores, 1986), Soviet Psychology's emphasis on activity (Norman, 1993)(Wertsch, 1981), and even concepts taken from Zen Buddhism (Varela, Thompson, & Rosch, 1991).

But Dretske's claim still haunts. The West Coast may provide effective critiques of computationalism, but can it yield a practical guide to building artificial intelligences? For a while it seemed as though connectionism might provide a suitable alternative (Dreyfus & Dreyfus, 1988), but this has (usually) repeated the computationalist assumption that cognition is the transformation of one set of representational symbols into another. The only real difference between this form of connectionism and computationalism is that the former uses a vector algebra, rather than scalar, to manipulate its symbols (Cummins & Schwarz, 1987)(Smolensky, 1988).<sup>4</sup>

Another West-Coast alternative has been to try to understand how orthodox computer systems are embedded and used within a social context (see (Laurel, 1997) for a good example). But this approach does not yield artificially intelligent systems, just ones with better interfaces. A *tamagochi*, for example, may be regarded by its owner as a live sentient creature that deserves care and attention. And such products certainly tell us something interesting about our relationship to 'intelligent' computers. But this hardly constitutes the foundations for a research program into building artificial systems that exhibit the intelligence of animals.

However, if the West Coast is correct to insist that cognitive behaviour cannot be characterised as a formal and abstract input-output mapping then the only way to build a cogniser is to build an agent that physically interacts with its world. Thus there has been a rapid increase in research in robotics that eschews conventional computationalist techniques, variously known as artificial life, behaviour-based robotics, the simulation of adaptive behaviour, *nouvelle AI*, post-modern robotics and so on. However I prefer the term 'South Coast AI', referring to the Artificial Life group of the University of Sussex on the south coast of England. This is not a question of academic priority but rather a recognition of the unusual synthesis of philosophical debate, robot engineering, and neuroethology in this institution, as noted by Keeley (1998).

It has to be said that progress in South Coast AI has been painfully slow compared to that of the East Coast. Computer chess players can beat human grandmasters, but the robot footballers that are a feature of most robotics conferences would scarce trouble a two year old child, let alone Pele. Sometimes it is difficult to see any advance over the work of the pioneers of cybernetics in the 1950's, such as Grey Walter and Ross Ashby (1952), or the robotic thought experiments of the Swiss neuroscientist Valentino Braitenberg (1984), despite many billion-fold increases in computer power now used. A critical observer would be justified in thinking that South Coast AI is on a slow road to nowhere, and that this should tell us something about the theory on which it is based. But the fact is that we simply do not have a good theory to replace computationalism as a guide to constructing intelligent agents. And in the absence of a convincing theory, a thousand robotic flowers have bloomed. South Coast AI at the moment is characterised by a large number of often very small research groups, each working on their own particular techniques with very little sense of constructive, cohesive progress. The only notable exception is MIT's *Cog* project, in which a diverse set of particular solutions to partial problems — such as saccading eyes, reaching

---

<sup>4</sup>Note that this is a criticism of connectionism considered as a method of mapping one set of representations onto another, rather than the use of artificial neural network to control embodied agents — see below.

for an object, and tensing an arm — are being progressively added to a single humanoid robot, in the hope that humanoid intelligence will one day collectively emerge.

However two of these flowers are of theoretical interest. The first is to copy — or at least take inspiration from — nature, and use the findings of neuroethology to model simple natural sensory-motor systems in robots. This is generally known as computational neuroethology (see (Beer, 1990) and (Cliff, 1991)). The second approach is to artificially evolve neural network controllers for robots using genetic algorithms. This is evolutionary robotics (see (Beer & Gallagher, 1992) and (Harvey et al., 1997)). What both these approaches have in common is that, in the absence of a good theory, they avoid designing robot control systems by hand, and instead leave the design process up to natural, or artificial, selection. The use of representations is thus no longer an *a priori* assumption about how to build intentional agents, but is rather an open empirical question about how they work. And a significant minority of researchers have concluded they are simply not necessary, most notably in Brooks' landmark paper *Intelligence Without Representation* (1991). (Also see (Beer, 1995b), (Harvey, 1992), (Cliff & Noble, 1997), (Van Gelder, 1992) and (Wheeler, 1994) for variants on the same theme.)

However all these objections assume that representations must fit the East Coast model, in which the mechanism of the agent can be neatly carved up into humuncular modules which 'communicate' using a vocabulary of symbolic representations. These modules are fixed, disjoint, and completely general purpose (in the sense that there is a single modular organisation capable of producing all behaviours). For example Wheeler, in discussing the analysis of evolved robot control systems, cites Beer's remark that 'highly distributed and richly interconnected systems [such as evolved neural networks] . . . do not admit of any straightforward functional decomposition into representations and modules which algorithmically manipulate them' (Beer, 1995a, p128) (cited in (Wheeler, 1998)).

However Beer *et al* are shooting at the wrong target. They are correct that such evolved networks do not show a modular decomposition obeying algorithmic rules, and such empirical evidence is a powerful weapon against computational and cognitivist assumptions about the mind. Moreover, central nervous systems are the most complex, non-linear, and feedback-ridden systems we know of and understanding them is rarely 'straightforward', especially when we are trying to understand their interactions with a messy real world environment. (I once asked an ethologist who had studied navigation in insects for many years why he did not encourage students to investigate the neural mechanisms underlying the behaviour he was so interested in. His reply was not that this would be impossible, but that someone could easily spend 20 years on this research and still not get anywhere — and this is for a relatively well understood behaviour in a 'simple' insect.)

However in the last section I emphasised that representations are only defined with respect to, and in the context of, the behaviour of a whole agent within an environment. This implies that (1) there need be no general-purpose algorithmic or representational organisation underlying different behaviours, and (2) that any representational functional organisation is an emergent product of the interaction between an agent and its environment. For example as we saw in the case of the Horseshoe crab and the rat hippocampus, the mode of organisation and 'intrinsic' properties of a neural system may change radically from one behavioural context to another, and there is no reason why representational correlations found in one situation should play a role, or even exist,

in another. If this is taken into account then Brooks' *et al* objections lose their force and we can instead appreciate how the examples of robot control systems, and animal nervous systems, that are often held up as paradigm cases of non-representational intentionality do, in fact, have an emergent representational character (Faith, 1997).

One much-cited example is an experiment conducted at the University of Sussex in which artificial evolution was used to generate not only the control system for a robot, but also a suitable body for it to control (Harvey, Husbands, & Cliff, 1994). The 'fitness' of the robot was judged by its ability approach a white triangular target, whilst avoiding a rectangular one. The successful robot used just two sensors, one with a visual field above the other, and located the triangle by rotating on the spot until just the lower sensor saw white and moving straight ahead. This has the effect of fixating the robot on the oblique edge of the triangle. As the triangle looms up such that both sensors go high, or if the motion causes the edge to be lost, then the robot will start to rotate until the edge can be fixated again. The rotate/move-straight distinction is effected by a single unit that takes an inhibitory connection from the upper sensor and an excitatory link from the lower, and is thus only fully activated when the robot is facing towards the triangle's edge.

Two points about this robot must be noted. The first is that its success depends on having a sensor morphology that is perfectly suited to the targets in its environment. If those targets were shaped even slightly differently then there would be no simple way of using the same eyes to do the same discrimination. The control system is also finely tuned to the types of motor and the timing of rotation: if even just the amount of noise in the system is changed then the whole robot has a tendency to overshoot and end up literally going in circles. Therefore, you cannot understand the brain of the robot without also understanding its body and environment. Nonetheless a crucial part of understanding how it does this is to note the correlation between the triangular target being straight ahead, the activation of a particular unit, and the robot moving straight — a representation, in the sense defined above.

To take another example, Floreano and Mondada describe the artificial evolution of a neural network controller for a robot whose task is to explore a simple arena, returning to a recharging base that is demarcated by a black floor patch and directed by a bright light. It was found that the fittest individual used a hidden node of the network whose activation corresponded to the distance from the base, reaching a maximum when it was 'home'. As the experimenters note:

In this experience the robot autonomously evolved the ability to use the raw sensor data and built an internal representation of the world in order to find the recharging area and return to this place at a given time. This behaviour is based on an accurate evaluation of the battery residual time and on an internal representation of the environment. In fact some of the hidden nodes displayed activation levels that clearly mapped the environment geometry. (Mondada & Floreano, 1996)

Evolutionary robotics is in its infancy, and the tasks it tackles are so simple that in many cases they can be solved by agents with only the most direct stimulus-response reflexes. Indeed the evolutionary strategy is brilliant at finding ingenious stimulus-response solutions to tasks that a human designer would normally insist could only be achieved by forming and manipulating representations of the robot's environment. (This specific question is investigated empirically by (Miglino et al., 1998).) However, as tasks become more complex the use of internal states

that carry information about the environment becomes less and less avoidable (Kirsh, 1991), and even in the very simple cases mentioned above we find that individual units act as very simple representations in mediating interactions between the robot and its world. Indeed Brooks himself later conceded that he was not arguing against representations *per se*, but that he was merely advocating different *types* of representation:

My earlier paper (1991) is often criticised for advocating absolutely no representation of the world within a behaviour-based robot. This criticism is invalid. I make it clear in the paper that I reject traditional Artificial Intelligence representation schemes. I also made it clear that I reject explicit representations of goals within the machine.

There can, however, be representations which are partial models of the world — in fact I mentioned that “individual layers extract only those *aspects* of the world which they find relevant — projections of a representation into a simple subspace”. The form these representations take, within the context of the computational model we are using, will depend on the particular task those representations are to be used for. (Brooks, 1995)

The same softening of anti-representationalist attitudes amongst South Coast engineers can be seen amongst Clark and Wheeler (1998), Scheier and Pfeifer (1998), Bickhard (1998), Calabretta *et al* (1998), and Tani *et al* (1998), who all agree that even very simple intentional behaviours are mediated by representations, but representations that can only be understood in the context of activity. At least one neuroethologist draws a similar lesson, but again confuses rejection of computationalist symbols and modules, with rejection of representations *per se*:

Of course the cognitive approach — the representational paradigm — is a level of interpretation in its own right. At best, it is like Ptolemy’s system of epicycles, which could describe the movements of the planets in sufficient detail; but as we now know, the heliocentric view of the world provides a simpler way of understanding this movement and one that comes closer to what is actually the case. By analogy, the cognitive-map approach might obscure some of the most important computational strategies used by the brain. In general, the brain has evolved not to reconstruct a full representation of the three-dimensional world, but to find particular solutions to particular problems within that world. (Wehner, Michel, & Antonsen, 1996)

The representations advocated by both Wehner and Brooks are not general-purpose symbols syntactically manipulated according to an East Coast algorithm, but rather describe how the sensory-motor transformations required for particular behaviours are achieved. The representational organisation underlying, and emergent within, one behaviour need bear no relation to that underlying another.

In Brooks’ experiments this separation is embodied in a ‘subsumption’ robotic architecture — as used on *Cog* — in which the mechanism is split into largely independent ‘layers’, each of which is connected to both sensors and motors. Evolution, both natural and artificial, does not tend to produce such extreme disjointedness but rather produces mixed bags of tricks made up of particular solutions to particular problems, in which evolved circuitry is used and adapted to new purposes. In either case, in order to understand how these systems achieve robust co-ordination with objects in their environment it is necessary to understand how information about those objects play a role in controlling that behaviour.

It is interesting to note that the most doctrinaire anti-representationalists have been computational neuroethologists and evolutionary roboticists, rather than the biologists who study natural sensory-motor systems. It seems that this position stems from a healthy desire amongst computer scientists to disassociate themselves from the tradition of computationalist artificial intelligence and its Cartesian understanding of representation. Biologists have rarely been tarred with the Cartesian computationalist brush — after all, no-one can accuse them of studying disembodied intelligence — and so seem more comfortable with describing the neural mechanisms that they discover in representational terms (Roitblat, 1994).

Lying behind South Coast anti-representationalism there often lurks the intuition that something is only a representation to the extent that it is part of a generalised symbol system. Without such a system it is assumed that an internal state does not have well-defined semantic properties. They therefore share the cognitivist assumption that representations — and intentionality — are to do with computation, rather than the ability of an agent to actively engage with its world. The philosophical roots, and implications, of this argument will be discussed in chapter 6.

The lesson of South Coast AI is that if you want to build a cogniser, you shouldn't start by making up fancy data-structures, since without a body they are both meaningless and useless. Moreover, just bolting on sensory and motor modules will rarely succeed in effectively tying a symbol system to the world, since those contents that a human intuition assigns to the symbols are unlikely to be the ones that its crude body can make available. This was the problem of East Coast robotics, as exemplified by *Shakey* (Nilsson, 1984).

*Shakey* was a mobile robot that could move blocks round a set of rooms, according to typed instructions. At its heart was a predicate calculus model of its environment, manipulated by a means-end problem solver (Newell & Simon, 1972), and generated from a camera image of its environment — ‘a series of still pictures’, in Costello's phrase. However the problem of producing a symbolic representation of its environment meant that the rooms had to be specially designed to be as visually simple as possible, with flat floors, evenly coloured surfaces, careful lighting, few obstructions, and so on. Although *Shakey* worked, it proved impossible to generalise its success to more realistic environments. The moral is to start by getting the body right, and concentrate on tying it to the world; representations will be the emergent result, as the evolutionary roboticists have repeatedly found. The East Coast approach to artificial intelligence is like noting that a good Emmenthal cheese invariably has holes in it, and concluding that the starting point for making one is to glue pockets of air together. The South Coast approach is to start with the cheese. If you get this right, then you find you get the holes for free.

The limiting factor in our development of intelligent artificial creatures is not the computational power of their ‘brains’, but in the more basic engineering technology of their bodies. Current robot engineers use roughly the same motor and sensor technology that the pioneers of cybernetics did, and yet this is where the real problems of embodied intelligence lie. Therefore we should not be surprised at the slow progress. Rod Brooks draws an illuminating comparison between the development of computer technology, and that of jet airliners. The power, capacity, and speed of the former have doubled roughly every 18 months, whereas the same improvement in the latter has taken almost 40 years. We should expect the development of South Coast AI to be more like that of airliners than computers, and for similar reasons. Building successful robots

depends more on the ‘hard’ engineering of bodies than on the ‘soft’ engineering of brains.

#### **4.4 Conclusion**

In order to understand how neural mechanisms can underlie intentional behaviour it is necessary to understand how they can carry information about the environment of the organism. This requires that we investigate the entire causal loop, involving brains, bodies and environment. Moreover, this system must be investigated *in vivo*, since the relevant neurological properties may only exist in the appropriate behavioural context. The same lesson applies when constructing artificial intentional systems: we cannot start from an isolated representational module that approximates humanoid problem solving, since without a humanoid body it will have no effective connection to the world that it is supposed to represent.

## 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 Dennett<sup>1</sup>. 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.

---

<sup>1</sup>I 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)<sup>2</sup>. 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.

---

<sup>2</sup>So 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 correctly<sup>3</sup>. 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,

---

<sup>3</sup>The 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 power<sup>4</sup>. 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

---

<sup>4</sup>Note 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 responsivity 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 all<sup>5</sup>.

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-

---

<sup>5</sup>This 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.

## Chapter 6

### Intentionality: Outsides

---

If a lion could talk, we could not understand him.

— Wittgenstein, *Philosophical Investigations*

‘Ouch’ is a one-word sentence which a man may volunteer from time to time by way of laconic comment on the passing show.

— W.V.O. Quine, *Word and Object*

In the previous chapter I discussed the problem of carving up the insides of an agent, of picking out its beliefs and desires, in order to make sense of its behaviour. But in order to do this we also have to solve the symmetrical problem carving up the outsides, i.e. picking out the objects of its environment. We not only want to know what the agent is thinking, but also what kinds of things it is thinking *of*.

This problem is often ignored since the only philosophers who worry about relating things-in-the-head to things-outside — i.e. realists — also tend to be the ones who assume that there is a fixed list of Objects in the world, that science will in the end tell them what those Objects are, and if the contents of the agent’s thoughts are not on that list then they are just plain wrong (or not thinking about anything real at all). Therefore the job of discovering what kinds of things an agent may be thinking of is a job for natural scientists, not psychologists or philosophers. On the other hand Wittgenstein and Rorty notice the problem, but assume there is no solution since the only way to know what kinds of things a lion is thinking of is to *be* a lion. In this chapter I argue that there is a substantive problem here *and* that it is soluble from a third-person perspective.

#### 6.1 Sense and Reference

Consider the following two examples. The first is Putnam’s twin-earth experiment (1975). Suppose Jean is transported in her sleep from earth to twin-earth. Twin-earth is *exactly* like the original except that water is made out of *XYZ* not *H<sub>2</sub>O*. However the stuff still looks and tastes the same, and so as far as Jean is concerned (not being a chemist) there is no difference. Putnam introduced this example to prove how there is more to the contents of our beliefs than the role that they play in our heads: the properties of the stuff that Jean calls ‘water’ have changed even though Jean’s

thoughts about it have not. The aspect of Jean's thoughts about water that remains the same on earth and twin-earth is the 'narrow' content, and the aspect that has changed is the 'broad' content.

The second is Frege's example of the terms 'Evening Star' and 'Morning Star' (1892). These terms meant different things to the ancients but later astronomers discovered that they both referred to the same thing, namely Venus. Thus in Putnam's example there are two distinct referents ( $XYZ$  and  $H_2O$ ) which Jean grasps using a single term with a single sense (Water). Whilst in Frege's example there is a single referent (Venus) that was grasped using two terms with distinct senses (Evening Star and Morning Star). Thus Frege and Putnam are trying to draw the same distinction between those aspects of meaning that play a role in the head (sense and narrow content, respectively) and those that do not (reference and broad content)<sup>1</sup>.

The problem that Frege and Putnam's examples generate is this: Jean did not notice the difference between water and twin-water, and the ancients did not see the link between the Evening Star and the Morning Star. We are only able to draw the distinction between sense and reference (or narrow and broad content) in these cases because we take a kind of God's-eye view of the situation from which we are aware of things that the ancients, and Jean, were not. But for all we know there may be Higher beings that are making up the same kind of thought experiments about us ('imagine a group of people who stupidly think that water is  $H_2O$ , when of course modern super-physics shows that this was just a crude approximation'). Therefore our talk about the reference of water 'being'  $H_2O$  is just as sense-laden as Jean's take on the world. If referents are supposed to be independent of us, then how can we talk about them? As Putnam argues, '*What objects does the world consist of?*' is a question that it only makes sense to ask *within* a theory or description' (1981, p49). Or, as Nabokov put it, 'reality' is a term that means nothing except when in quotes.

Now it must be acknowledged that many realistically-minded philosophers cannot see this problem, or — and this amounts to the same thing — they do not think that it matters. Why should it be necessary to be able to talk about the referent of our thoughts in a way that does not use our vocabulary of thoughts? Why should we require that it is possible to 'reduce' the referent, or state it in other terms? Such philosophers would argue we are talking about the referent simply by mentioning it, and that the thing we are talking about is independent of our way of talking. But the problem starts to manifest itself as soon as we try to prefer one way of describing the world over another. Frege would conclude, for example, that the term Venus is better in some way than either Evening or Morning Star; and Putnam's thought experiment seems to suggest that Jean would have been more accurate in describing the wet stuff in her world(s) as  $H_2O$  and  $XYZ$  respectively. But why?

Frege did not have a solution to this problem but he did have a way round it, a strategy that subsequently became fundamental to most analytic philosophy. Frege starts from Kant, and in particular Kant's assumption about what the world is 'really' like; namely that it is divided into independent objects-in-themselves which possess various properties. Frege concludes that if a language is to be scientifically respectable then its structure must reflect this essential structure of the world (Dummett, 1991b, ch20). This means that (scientifically respectable) sentences must be expressible in predicate-subject form,  $P(x)$ , where  $x$  refers to an object defined independently of that sentence, and  $P$  a property that is predicated of it:

---

<sup>1</sup>Fodor (1994) also draws a comparison between these two examples, though for a slightly different purpose.

Statements in general, just like equations or inequalities or expressions in Analysis, can be imagined to be split up into two parts; one complete in itself, and the other in need of supplementation, or ‘unsaturated’. Thus, e.g., we split up the sentence

‘Caesar conquered Gaul’

into ‘Caesar’ and ‘conquered Gaul’. The second part is ‘unsaturated’ — it contains an empty place; only when this place is filled up with a proper name, or with an expression that replaces a proper name, does a complete sense appear.

... When we have thus admitted objects without restriction as arguments and values of functions, the question arises what it is that we are here calling an object. ... Here I can only say briefly: An object is anything that is not a function, so that an expression for it does not contain any empty place. (1891)

Frege’s condition — that truth-bearing sentences must involve predication over a ‘saturated’ object term with prior reference — recurs in many forms. We find it in Russell’s claim that a subject cannot make a judgement about something unless they can know which object their judgement is about; i.e. that the subject can refer to that object independently of that particular predication (1905). We also find the same assumption in Evan’s *generality constraint* (1982), Fodor and Pylyshyn’s criterion of *systematicity* (1988), and Millikan’s condition of *propositional structure* (1984). The assumption is the same in each case: a subject cannot predicate a property, *P*, of an object, *x*, unless they can equally well predicate any other properties, *Q* or *R*, of *x*, and predicate *P* of any other objects, *y* or *z*. The same assumption also lies behind the argument within South Coast AI (section 4.3) that information-bearing functional states only count as representations to the extent that they are part of a more general symbol system. When Brooks, Beer, Harvey and Wheeler declare that one can have ‘intelligence without representations’ they mean representations that meet the generality constraint.

How does Frege’s (or Russell’s, or Evans’) condition help us avoid the problem of reference? Compare the two sentences

1. The referent of the term ‘the Morning Star’ is Venus.
2. The referent of the term ‘Venus’ is the Morning Star.

Now strictly speaking both sentences are meaningless since there is no non-circular way of talking about the referent of a term. Nonetheless Frege *et al* give us a reason for *preferring* the first sentence to the second. The reason is that ‘Venus’ is a more *objective* term than ‘the Morning Star’ because the term ‘Venus’ may be used independently of when in the sky it appears, whereas ‘the Morning Star’ is more closely tied to particular observation conditions. For example, the sentence ‘the Morning Star appears in the morning’ is (almost) tautologous, but the sentence ‘Venus appears in the morning’ is not. In short, the term ‘Venus’ has a better claim to be part of an ideal scientific language than ‘the Morning Star’. It is this that justifies the assertion of 1 rather than 2. Of course our thoughts may never come into ultimate correspondence with the way that the world ‘really’ is (and we would not know it even if it did), but we can still tell when we are getting closer. It might turn out that we were mistaken about Venus being a ‘real’ object all along, but the discovery of facts about Venus that are independent of its position in the sky is evidence that we are not. It is interesting to note that even the most hardened realist-minded scientists rarely claim

that they actually know the truth, instead they usually only claim that we can get better and better approximations to it. And the closeness of that approximation is judged according to how well our thoughts meet Frege's conditions<sup>2</sup>.

Frege avoids the problem of reference, but does not solve it. He gives powerful reasons for preferring one system of description to another but these reasons are still based on an unjustified Kantian assumption about the essential structure of the world. The fundamental objection remains that the only things that our minds have access to — i.e. the only things that play a role in our heads — are senses. Frege defines sense as the mode of presentation of a referent to a mind but there is an alternative tradition, starting with the later Wittgenstein and including Rorty and the later Putnam, that defines sense independently of any notion of reference. According to this tradition 'the sense of an expression consists in its role within the complex social practice constituting the communal use of the language ... An individual speaker's grasp of that sense then becomes one ingredient in his ability, acquired by training, to engage in that practice' (Dummett, 1991a, p17). According to this tradition Truth does not lie in a correspondence between things-in-themselves and things-in-the-head (or expressions in a language), but in the ability of an individual to use language successfully.

For example, suppose that Jean walks into a bar and asks for a glass of water, but when she gets it she complains that the glass does not just contain water but also traces of mineral salts, bubbles of carbon monoxide, and a slice of lemon. Is it a glass of water? Clearly yes, since the norms of correct practice for bars define what constitutes 'a glass of water' in that context; different norms apply in a chemistry lab, where it is not normal (i.e. correct) practice to add ice and a slice to beakers of water.<sup>3</sup> Thus we can discuss the Truth of the use of terms like 'water' without recourse to metaphysical debates about what water 'really' is<sup>4</sup>.

Can we do without a notion of reference, as Wittgenstein *et al* maintain? The sole reason for clinging onto reference is that we need some way of explaining why some representational vocabularies are more successful for certain purposes than others. Why, for example, does passing an electric current through water generate hydrogen and oxygen? Unless we believe that water *really is* made of  $H_2O$  then we are lost for an explanation. It is only the possibility of realism that makes the success of science non-miraculous, as Putnam put it (1973). This point is the mirror-image of that raised in the previous chapter: if we want to use the concept of belief to explain behaviour then beliefs must be instantiated in the brain mechanism underlying that behaviour; similarly, if we want to use reference to explain the success of the behaviour that those beliefs play a role in, then the reference of those beliefs must be instantiated in the world outside the head. But is there any non-circular way of talking about the reference of our thoughts?

There are two steps to breaking out of this circle. The first step is to adopt a third-person perspective. It is impossible to directly observe the world of reference outside our own head<sup>5</sup>, but we can observe the world that surrounds other people's. But then there is still the problem

---

<sup>2</sup>I believe that one consequence of this is that mathematical physics, amongst all the sciences and rival systems of thought, is seen as having the best chance of grasping the real structure of the world since it approaches most closely to the Fregean ideal. But this is another question.

<sup>3</sup>The origins of these criteria of success and correctness will be the subject of chapters 8–11.

<sup>4</sup>Winograd and Flores (1986) use another water-based example to make a similar point.

<sup>5</sup>Those realistically-minded philosophers who do not acknowledge the fundamental problem of reference that lies behind this chapter will probably deny this claim; but I have a feeling I will have lost them a long time ago anyway.

of picking out the referents of their thoughts. We will have to use some vocabulary, some set of concepts, to pick out the objects that they are thinking of, and are we not trapped in using our own? What kinds of things would a lion talk of, if it could? The solution is to define a new vocabulary for picking out the objects in other agents' worlds, and one that is not based on our own concepts. This is the job that Evans' concept of Non-Conceptual Content (NCC) can do.

## 6.2 Non-Conceptual Content

Non-conceptual content is content that is characterised using concepts that the agent having the thought does not necessarily possess. For example, suppose we hear a sound coming a certain position in space and we turn our head to see where the sound was coming from. Looking back on the experience we may describe the reference of our state of mind in Fregean, conceptual, objective terms as 'a sound source located at position *X*', but this thought probably never occurred to us at the time. We just turned to see what it was. Evans suggests that it is more accurate to describe the contents of our thoughts in terms of our *activity*:

What is involved in a subject's hearing a sound as coming from such and such a position in space? ... When we hear a sound as coming from a certain direction, we do not have to think or calculate which way to turn our heads (say) in order to look for the source of the sound. If we did have to do so, then it ought to be possible for two people to hear a sound as coming from the same direction and yet to be disposed to do quite different things in reacting to the sound, because of differences in their calculations. Since this does not appear to make sense we must say that having spatially significant perceptual information consists at least partially in being disposed to do various things.

Thus the content of the thought is better described as something like 'a sound coming from a direction that would be foveated if we turned our head *so*'. This description of the content uses concepts ('direction', 'foveate', etc) that we are not assuming that the agent used at the time, or even possesses, and instead makes essential reference to their ability to act in the world ('turn our head *so*') — described by Cussins as 'the realm of embodiment'.

Evans originally developed the theory of NCC to better describe our first-person perceptual experience of the world, and this is largely how the theory has subsequently been used by Peacocke (1992), Cussins (1992a), Crane (1992) and others. But it can also be used from a third-person perspective to describe the thought processes of agents without making any assumptions of whether they are consciously aware of them or not (Bermúdez, 1995)(Chrisley, 1995). Consider the example of a frog striking at flies. What is the frog thinking of when it does this? Of course in one sense it is not thinking of anything at all. If it has any consciousness then it is a very limited one. On the ladder of cognitive complexity it is only one step up from my reluctant car. Nonetheless it uses internal states that bear information about the environment in order to co-ordinate its interactions with the world — i.e. very simple representations — and so we can usefully explain its activity by ascribing it with simple proto-beliefs<sup>6</sup>. But what are these proto-beliefs *of*? From our Gods-eye view of the frog we can see that the things that it is striking at are flies. An entomologist

---

<sup>6</sup>If you are not happy with this example then we could take a few more steps up the ladder. How about a lion? A chimp? A new-born baby? All we need is an example of an agent whose behaviour we can usefully explain using psychological talk but whose conscious experience, if it has one, is very different from our own.

could go one step further and identify the individual *species* of fly. But it is obvious that these fine distinctions in carving up the world play no role in the frog's activity. A frog does not have the concept 'fly', let alone a concept of particular species. As far as it is behaviourally concerned all it perceives are 'things that it should strike at like *so* and eat'. This is the content of its belief, described using concepts that we are not pre-supposing that the frog itself possesses.

Evans' analysis has similarities with Gibson's ecological theory of perception (1966)(1979)<sup>7</sup>. Gibson argues that the basic things we perceive are not objects as the reductionist imagines them, but *affordances*. An affordance, as defined by Gibson, is the way in which the environment can play a role in the behaviour of a creature. For example, the surface of a pond provides the affordance of 'something to walk on' for a pond skater, but not for a human. An affordance describes an environment *for*, or with respect to the behaviour of, a creature. It describes the world *of* the creature, and not the world as viewed from nowhere — or, rather, from the point of view of nobody. An affordance is thus a relational property of an environment, defined with respect to a specific behaviour of a specific organism. The same object can provide many different affordances for a single creature and, conversely, different objects can offer the same affordance. Flies and bees both provide the affordance of eatability for a frog, while a pond provides both the affordance of spawning, and of escaping from terrestrial predators.

Specifying contents in terms of affordances turns the standard analytic account of perception on its head. This account was inherited from the British empiricists by the logical positivists and passed, *via* Carnap, to cognitivism, computationalism and East Coast AI, where it found canonical expression in the work of Marr on vision (1982). According to this account, perception starts when we form an internal map of the objects in our world from sense-data. These perceived objects are then categorised and attributed with properties so that we can plan our behaviour. Therefore the ability to act in the world is built upon a more basic ability to perceive objects. But according to Evans and Gibson perception is not primarily the ability to form an objective map of one's environment, but the ability to act within it<sup>8</sup>. Most of the time we are actively engaged in the world rather than considering it passively. (Philosophers are the exception to this rule, which explains why they are so fond of the standard model of perception.) Of course we are also able to perceive objects passively, to categorise them and predicate properties of them, but this passive ability to perceive objects is built on a more basic ability to perceive affordances in our environment.

For example, when we reach for a saucepan or avoid tripping over the cat we do not primarily perceive them as categorised and labelled objects but as something like 'things to cook with' or 'things to avoid'. Indeed, even these descriptions of the contents of our thoughts are misleading since, by definition, it is not possible to give an accurate translation of the content of a non-conceptual thought using plain English concepts. One alternative way of describing this content is to hyphenate the description (e.g. 'thing-to-cook-with'), to show that the content should be understood as a unified whole rather than a construction of sub-concepts. But the canonical way of describing such contents is to describe the activity they play a role in. Thus, for example, it is more accurate to say that our thoughts about the saucepan at the time was comprised of a correct recognition that the environment enabled us to boil potatoes by moving and acting in certain ways.

<sup>7</sup>And Rowlands (1997) shows how this analysis fits neatly into an evolutionary framework.

<sup>8</sup>Gibson's analysis of perception is often seen as incompatible with a referential theory of mind, but Sloman gives an example of how this is not necessarily the case (1989).

One consequence of this is that in order to communicate the content of a (non-conceptual) thought from one head to another it is not sufficient that they share the same language; they must also share the same body (or, at least, the same abilities to act in their environment). Someone who cannot cook, for example, cannot really know what a skilled chef means by ‘a saucepan’ because they cannot know what a saucepan means to them.

NCC enables us to describe the contents of thoughts using concepts that the agent having the thought does not necessarily possess. This may be because the agent does not have any concepts at all (as in the case of the frog), it may be because the agent has those concepts but did not use them at the time (as in the case of our perception of cooking implements), but it may also be because the agent uses concepts that are different from our own. Consider Quine’s example of the linguist trying to understand the language of a native tribe (1960). The linguist observes that whenever a native sees a rabbit they point and say *gavagai*, but how should they use this evidence to decide what ‘gavagai’ means? Quine argues that the meaning of ‘gavagai’ is underdetermined by the observed behaviour. Just by observing the native pointing to rabbits the linguist does not have enough evidence to decide whether ‘gavagai’ means the same as ‘rabbit’, or if it means something like ‘undisconnected rabbit parts’ — you cannot have one without the other, and so there seems to be no principled way of choosing between the two possible meanings. But there is more to the meaning of ‘gavagai’ than its observation conditions. There is also the role that rabbits play in the life of the native speaker. And it is this aspect of meaning that we can use the technique of NCC to describe.

For example, suppose that the linguist discovered that the native had access to a genetics lab, and was fond of sequencing the DNA of any animals that she came across. This would be evidence that the native’s concept of *gavagai* had a similar meaning to that of an English-speaking biologist’s concept of rabbit — something along the lines of ‘members of an inter-fertile species of rodents identified through their possession of a wild-type genome of *XYZ*’. On the other hand if the linguist found that rabbits were no more than a source of food and skin to the native, then the meaning of *gavagai* would change accordingly. Thus the linguist can try to understand the meaning of a term by understanding the role that it plays in the life of the native speaker, rather than by trying to find a concept in her own vocabulary that is an exact translation of it. Indeed it is unlikely that, in the former case, the English-speaking linguist would have a concept that is a synonym of ‘gavagai’ for the same reason that she could not quite grasp an English-speaking biologist’s concept of ‘rabbit’. The obstacle preventing an accurate translation is that the linguist and biologist have different ways of interacting with rabbits, rather than different vocabularies.

The meaning of a term is defined by the role it plays in an agent’s overall activity — its being-in-the-world, form of life, existence, world, praxis, practical discourse, lifestyle, or mode of production, depending on one’s choice of existential philosopher<sup>9</sup>. Words are tools, and their meaning is defined through how they are used. Now when we describe the use of a tool, we are not trying to ‘translate’ that tool into words. When we describe how a hammer is used we do not suppose that there is any hammering going on in our heads. Similarly, when the linguist describes the native biologist’s concept of ‘gavagai’ as ‘a member of a species . . .’ she is using her concepts

---

<sup>9</sup>In chapters 10 and 11 I argue why the latter term is more useful by discussing its role within the evolution of culture. But this is another question.

to describe the activity of the agent through which the meaning of ‘gavagai’ is defined for them. The linguist is *not* thereby assuming that the native shares the same concepts, any more than we assumed that the frog understood the concept of ‘striking’. Thus NCC provides a way of carving up the world in order to make sense of an agent’s behaviour that does not involve foisting the distinctions made by our concepts onto them.

### 6.3 Affordances and Objects

According to the Fregean tradition, conceptual thoughts refer to objects-in-themselves — these comprise what Cussins describes as the ‘realm of reference’. This tradition then produced the problem of talking about the realm of reference independently of our own way of talking. NCC seems to offer a solution to this problem by talking about the content of thoughts in terms of physical activity in a way that does not pre-suppose that the holder of the thoughts has the same concepts as us. But what do non-conceptual thoughts refer to? (Or, what are the referents of mental states whose contents are described non-conceptually?) Here opinions differ. Cussins, following Evans, argues that describing the contents of thoughts non-conceptually is a way of *avoiding* defining their reference:

*What Evans saw was how to pull apart the specification of content from the specification of reference or truth. If a canonical specification of a content need not be a specification of a truth condition, then canonical specification of a content which refers to the realm of embodiment does not entail the evident falsehood that the truth of the content depends on the character of the realm of embodiment. ... For Evans, truth conditions are fixed by the realm of reference, and not by the realm of embodiment; but the cognitive significance of representation is fixed by the realm of embodiment, and not by the realm of reference. (1992a, p656, original emphasis)*

This argument stems from the Fregean assumption that the ‘real’ world — i.e. the realm of reference — is ‘really’ divided up into independent objects-in-themselves. Therefore if the contents of thoughts fail to carve up the world in this way then they cannot be referring to that realm. There are still objects that non-conceptual thoughts will be true of, but these objects play no role in the life of the agent *per se* (they will not be ‘cognitively significant’). The real world remains hidden from the agent trapped in its realm of embodiment. Thus the frog only ‘knows’ about ‘things it should strike at’, not the various species of fly that infest the real world of the realm of reference. Frege’s problem of reference remains unsolved but Evans’ achievement, as far as Cussins is concerned, is to show how we can have a theory of content despite this.

Epistemology recapitulates ontology. In other words our theory of how we know the world depends on our theory of what the world is like. In chapter 2 I tried to loosen our Kantian assumptions about the essential structure of the world, and this has implications for Fregean assumptions about reference. In particular I argued that the world is *not* made up of independent and prior classes of objects-in-themselves which then come together to form larger wholes. All things on all occasions exist in contexts, therefore the boundary between an object and its world is not an ontological given; and nor are the criteria that make two objects ‘the same’ (or of the same type)<sup>10</sup>. We

<sup>10</sup>This also implies that *sets* of objects are not ontologically given — this corollary was important for the discussion concerning the individuation of theoretical terms on page 52n.

do not carve the world at its joints, rather we define joints through acts of carving. Moreover we carve the world in particular ways because those ways yield objects that have properties that are useful for us; in other words because those objects provide affordances. The frog picks out ‘eatable things’ because doing so helps it survive. We pick out ‘species of fly’ or ‘gas molecules’ or ‘washing machines’ because doing so helps us understand evolution, the behaviour of bulk gasses, and getting clothes clean, respectively. An important criterion of the scientific way of carving the world is that it should pick out objects that have properties that are relatively constant across contexts; but this does not mean that the resulting objects are ‘more real’ than those carved out for other reasons. For example, the property of being ‘an eatable thing for a frog’ is just as ‘real’ and objective as the property of being ‘a member of a species defined by a biologist’.

Nonetheless, as humans, it seems we *can* perceive objects independently of any particular affordance that they offer. Indeed for most philosophers this has seemed like the most basic kind of perception. Gibson argues that when we are cooking we directly perceive that a saucepan offers the affordances we need in order to make a sauce, and the thought ‘that object is a saucepan’ may not enter our heads. But it *may* enter our heads when we pause and look at our kitchen passively. We can identify the objects present, give them names, and think what predicates are true of them. Where does this ability to perceive objects *as* objects, independently of use, come from? And why does it seem so basic to perception?

A frog does not contemplate the black dot floating in front of it, decide that it may be something worth eating, and strike. All it perceives is an eatable thing *there*. Animals do not contemplate the world passively but are engaged with it in a constant pursuit of the four F’s. But as animals become more complex their behavioural repertoire grows. They become more flexible, and use their environment in different ways. Humans are the most extreme product of this process. With our free hands, opposing thumbs, and bulging cortexes we can learn to work on our environment such that it offers affordances that support a virtually unlimited range of behaviours<sup>11</sup>. We have even evolved the ability to learn how to manipulate objects in our heads, rather than in our hands; to imagine and consciously explore the affordances that they offer beyond the immediately F-able, and to share our findings through language. As agents become more behaviourally sophisticated their perception of objects becomes less tied to any particular behaviour that they may play a role in. Thus the perceptual space of humans becomes more objectified and conceptual as the richness of our interactions with the world increases<sup>12</sup>

Frogs, for example, only eat flies; and the only thing they do with flies it eat them. Frogs cannot be said to predicate the property Food of the object-class Fly, since neither exists for a frog independently of the other. Frogs cannot have any concept of ‘food’ separate from that of ‘fly’. Humans, on the other hand, are spectacularly omnivorous. There is scarcely a single biological entity that we cannot turn into food, therefore we can support a concept of ‘food’ independently of any particular *foodstuff*. We can also do things with, say, potatoes other than eat them — such as paint them, carve them into potato heads, or sell them. Therefore we can support a concept of ‘potato’ independently of the predicate ‘eatable thing’.

---

<sup>11</sup>The argument that the evolution of human brains and psychology is fundamentally grounded in our ability to manipulate the environment, rather than our ability to manipulate concepts, was central to Engels’ pamphlet *The Part Played by Labour in the Transition from Ape to Man* (1987) — see also *Posture Maketh the Man* (Gould, 1978, ch26).

<sup>12</sup>For example Cussins discusses how our objective perception of Euclidean space evolves with our ability to navigate that space (1992a) — see also Bennett (1996).

Traditional societies tend to be fairly fixed in the ways that they interact with the world. Conventions, taboos, and tradition dictate what use each type of object may be put to. But the birth of capitalism swept away all fixed ways of interacting with the world. Everything was up for grabs — or rather everything could be grasped in any conceivable way. The only limit on our ability to interact with the world — and hence to perceive it — is now the limits of those objects themselves. As Marx, in the Communist Manifesto, put it:

Conservation of the old modes of production in unaltered form was the first condition of existence of all earlier industrial classes. Constant revolutionising of production, uninterrupted disturbance of all social conditions, everlasting uncertainty and agitation distinguish the bourgeois epoch from all earlier ones. All fixed, fast-frozen relations, with their trains of ancient and venerable prejudices and opinions, are swept away, all new-formed ones become antiquated before they can ossify. All that is solid melts into air, all that is holy is profaned, and man is at last compelled to face, with sober senses, his real conditions of life and his relations with his kind.

Our ways of carving the world are not dictated by fixed ways of interacting with the world. But this does not mean those ways of carving are *independent* of our interaction with the world, as the picture theorists argued. The perception of objects — i.e. the identification of references — can never be completely independent of the activity that those objects play a role in because the act of identification itself is based on activity. As Dummett argues

We have the notion of the bearer of a name, and the conception of a predicate's being true or false of an object, in advance of constructing a semantic account of our language in order to analyse its working, because these are embodied in quite primitive linguistic performances; our acquiring them is part of our learning to use our language. Both are born of the practice of ostension, that is, from our possession, in the use of a demonstrative accompanied by a pointing gesture, of another means than the employment of a name for picking out a concrete object. By means of a recognition statement (a statement of the form 'This is *a*'), we are accustomed to identifying an object as the bearer of a name; by means of ostensive predications (statements of the form 'This is *F*'), we are accustomed to applying predicates to objects picked out ostensively. To say that the referent of a name is its bearer, and that the referent is what we speak about, is in effect to say that the semantic roles of proper names and of simple predicates should be understood in relation to these fundamental practices: it is precisely because of our thorough familiarity with these basic linguistic practices that the notion of reference supplies us, as soon as it is introduced, with so definite and readily acceptable a picture of the semantic roles of at least the simplest logical types of expressions. (1981, p406)

## 6.4 Conclusion

In these two chapters I have argued that successful intentional behaviour depends on an accurate correspondence between things in the head (representations) and things outside (objects). In other words I am a realist and a representationalist. The pragmatist's objection to this kind of realist representationalism is that he thinks that it bases its explanations of success on a list of Objects as they Really Are. But, he argues, we have no way of determining what these Objects are beyond the pragmatic success of our own theories about them:

There is no independent test of the accuracy of correspondence . . . The representation-  
alist's attempt to explain the success of physics and the failure of astrology is bound  
to be merely an empty compliment unless we can attain what [Putnam] calls a God's  
eye standpoint — one which has somehow broken out of our language and our beliefs  
and tested them against something known without their aid. But we have no idea what  
it would be like to be at that standpoint. . . . My principal motive is the belief that we  
can still make admirable sense of our lives even if we cease to have what Nagel calls  
“an ambition of transcendence” (Rorty, 1991a, p6,12)

Of course we cannot have a Gods-eye view of ourselves, but the third-person perspective can  
give us a kind of God's eye view of the relationship between another agent and *their* world —  
as long as we solve two problems. The first is to identify the thoughts inside the agent's head.  
The solution to this, I suggested, requires that we *look* inside their heads: the correct intentional  
description of an agent depends not only on their external behaviour but also upon the mechanisms  
that underlie that behaviour. The second problem is to identify the objects of their world. And  
this, I suggested, requires that we do not pre-suppose our vocabulary — our way of carving up  
the world — will correctly carve up theirs. We may not be able to use *our* object-concepts to  
describe the referents of *their* thoughts. The alternative to using our concepts is to pick out those  
referents on the basis of their activity: their beliefs may still involve concepts, but they won't  
necessarily involve *our* concepts. Once we take both of these steps then we can reinstate truth  
— i.e. a relationship between things-in-the-head and things-out-there — as a notion capable of  
genuinely explaining the success of another agent's actions, even though we cannot use it directly  
to explain the success of our own. But if truth explains the success of the actions of the other  
agents we see around us, then it seems like a reasonable induction to suppose that it would also  
be capable of explaining the success of our own. (This is the third-person equivalent of the Other  
Minds Problem: instead of wondering whether other people have consciousness like us, we wonder  
whether we have the ability to represent an external world like them.) Rorty is correct that we  
cannot transcend our own mentality, but we can transcend the mentality of others; and of course  
they may transcend ours too. Therefore perhaps together we can transcend ourselves.

We can use truth to explain the success of actions without fear of circularity. The obvious  
problem we are then left with is to determine precisely what we mean by 'success'. And the  
solution to this, as we see in the remaining chapters, depends on Darwin.