Evolution, Consciousness, and the Internality of the Mind

The problem of consciousness seems to arise from experience itself. As we shall consider in more detail below, we are strongly disposed to contrast conscious experience with the physical states or events by which we take it to be realized. This contrast gives rise to dualism and other problems of mind and body. In this chapter I argue that these problems can usefully be considered in the perspective of evolution.

1. Evolution, mind, and the representation of mind.

Among other things, this perspective brings to the fore a contrast between human and (other) animal minds. Many animals seem to have minds, at least in the minimal sense of a system of internal states which represent the environment, and so enable them to attain goals intelligently, that is, in light of a wide variety of information relevant to this task. A striking thing about human beings, however, is that we also represent our own minds, via the various ways of thinking about internal states encompassed in our concept of mind.

This difference seems linked to one in adaptive function. While the advantages of possessing a mind apparently derive from representing the self in relation to the environment -- which of course includes other creatures and their behaviour -- those of possessing a concept of mind would seem to flow from representing these representations themselves, and so further anticipating and controlling the behaviour they govern. We apparently do this, for example, when we deliberate about our own motives, or seek to anticipate or influence those of others. How far animals besides ourselves enjoy these further advantages remains unclear.

2. An apparent difficulty for the evolution of a concept of mind.

We take the behaviour-governing representations described by our concept of mind to be realized in the nervous system. This means both that they are normally out of sight, and that their causal and functional roles could not be rendered transparent to perception in any case (as shown by the fact that we with all our science are scarcely beginning to fathom them.) So, as we can say, the human species -- or the human nervous system -- has had to evolve the capacity to represent behaviour-governing states and events in nervous systems blindly, that is, without drawing on perceptual information about neurons as disposed in space and having features perceptibly related to the operations of computation and control which they perform.

This seems part of the reason why in employing our concept of mind we do not represent the neural causes of behaviour as such, but rather as it were from the outside, that is, by relation to observable behaviour and the environment. In thinking of persons as having minds we construe their behaviour as action stemming from desires, beliefs, and other motives, which are directed to the environment and which serve to render both speech and non-verbal action rational (logical) and so intelligible. The capacity for thinking this way evidently develops together with that for using language from early childhood; and we can plausibly regard both as parts of a natural system for understanding and influencing human behaviour precipitated in the course of evolution.

3. Having a mind: representing environmental goals and information relevant to their attainment.

Let us now consider the causal role played by the representation of a goal in intelligent behaviour. Roughly, the representation should function so to govern behaviour as to bring about (cause) the attainment of the goal, and in addition the creature should register (cf Bennett 1979) that this has happened; and this should cause the cessation of this process. Thus consider a beaver whose goal is to stop a potentially erosive flow of water through his dam, and who succeeds in this task. Designating this animal agent by 'A' and the causal relations involved in perception and the regulation of behaviour by '-[causes]->', we have overall:

A acts with goal that A stops that flow -[causes]-> A stops that flow

And when the creature registers the attainment of this goal, we should have:

A registers that A has stopped that flow -[causes] -> A ceases to act with goal that A stops that flow.

These formulations display a familiar ambiguity, as between external and internal, in the way we speak of goals. We describe representations via what they represent; so the underlined sentence 'A stops [+ tense] that flow serves to specify both the alteration in the environment which is the animal's goal, and the behavior-directing representation of this which we take to be realized in its nervous system. The same holds for the use of the sentence to describe the internal registration of the event or situation in which the goal is attained. So on this account the process by which a creature ceases to seek a goal once it is attained is one in which the representation of the attainment alters that of the goal so that it ceases to direct behaviour.

Let us call the event which renders the goal-specifying sentence (in this case 'A stops that flow') true the satisfaction of the goal, and the termination of further goal-directed activity by the representation of this the pacification of the goal. Then the characteristic case of successful action will be that in which the satisfaction of a goal causes its pacification. Using 'G' for 'goal' and 'P' as a schematic letter for representation-specifying sentences, we can write this as follows:

D: A has G that P -[causes]-> P -[causes]-> A reg that P -[causes]-> A's G that P is pacified.

This schematizes the life-cycle of a single goal in successful action; and the same schema also applies to human desire, as becomes clear if we substitute 'desires' for 'has goal' and 'believes' for 'registers', as we shall do in considering the human case. These processes also include the veridical registration of information about the environment. For as part of the schema we have:

P -[causes]-> A reg that P

If we take it that the creature perceives and so experiences its own success, then this expands to:

B: P -[causes]-> A perceives (e.g. sees) or experiences that P -[causes]-> A reg that P

This schema describes an animal analogue of perception-based veridical belief; and we take animal goal-seeking to be informed by perception and memory of the enviroment in an analogous way. Thus if a beaver has the goal of stopping a certain flow of water, and perceives or otherwise registers that if it moves a certain branch then it will stop that flow, then it may form the subsidiary goal of moving

that branch. In this case the animal's goals are related to its information about the environment in a familiar and logical way, which we can set out as follows:

Initial Goal: A has G that P (that it stops that flow)

Information: A registers that if Q then P (if it moves that branch then it stops that flow.)

Derived Goal: A has G that Q (that it moves that branch).

Here the form of the goal- and information-specifying sentences mirrors that of a truth-preserving deductive argument; and this marks the way such an amplification of goals is satisfaction-preserving as well. As substitution again makes clear, this is analogous to human practical reason, and we can schematize it as follows.

PR: A has G that P & A regs that if Q then P -[causes]-> A has G that Q

Animal action is commonly driven by numerous goals related by complex information. Thus even in a very simple case a creature may have the goal that P (that it stop that flow) and register that it can do this if Q and R and S in that order (if it fells a certain sapling, gnaws off a branch, and moves the branch to the point of flow). We can indicate such structured goals by a derivational tree:

Such a tree runs from its aerial root down through series of goals to the lowest level of behaviour (here marked as the series of bodily movements M1 through Mo) which we take to ordered in this way. The ordering of goals manifest in such a tree thus corresponds to an ordered series of instances of G, nested in accord with complex instances of PR. Each tree relates the goal-representation at its root to a sequence of hypothesized effects, which, insofar as the animal is successful, should also be ultimately describable as a bringing about of the associated goal-situation, and thence of the registration of the attainment of the goal, and thence of its pacification. When we interpret an animal's movements in this way we tacitly relate them to a series of such trees, and in this we impose a hypothetical structure which is highly constrained and predictive, and so comparable to a powerful empirical theory. This is particularly so when, as in the human case, the assignments of sentences to non-verbal trees and actions can also be related to those manifested in speech.

4. Representing minds: a possible advantage of representing inner causes as opposed to regularities in behaviour.

This provides a minimal sketch of the kind of ordering and modification of goals in which we take animal intelligence to be manifested. A creature with a conception of this kind of intelligence, in turn, will be capable of representing such trees of representations, and hence of manifesting intelligence in relation to them. We can get some further sense of the value of representing inner causes as such by

considering Andrew Whiten's (1993, 1996) adaptation of an argument by which Neal Miller (1959) sought to convince behaviourists to take account of causes concealed in the brain.

Suppose we are studying the behaviour of some creature, whose body we represent by an opaque sphere, with question marks signifying that we cannot readily determine what is going on inside.

Figure 1

We can represent our information about input/output correlations relating the creature's behaviour to the environment by arrows. In the case of a laboratory rat, for example, we might have:

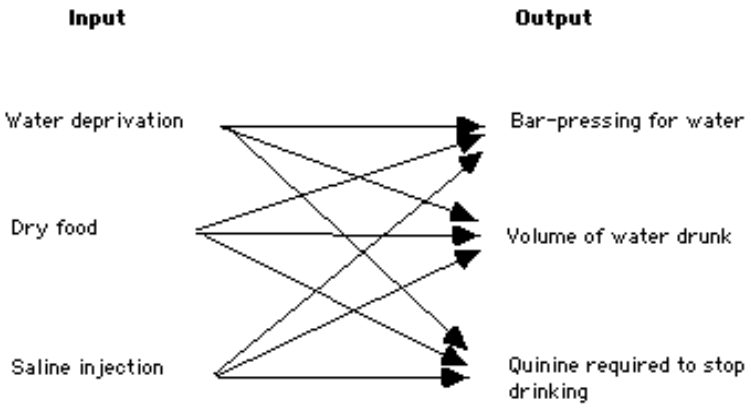

Figure 2

Now clearly we might want to explain these correlations by introducing the hypothesis that they are mediated causally, by a state within the creature. We can illustrate this by replacing the inner opacity with the hypothesized state, so that we have:

Input Internal Cause Output

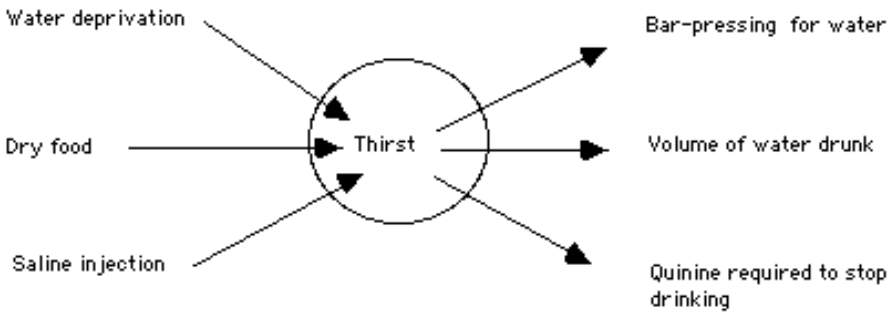

Figure 3

This simple theory is both plausible, and, as Whiten stresses, also a more economic representation of the correlational data on which it is based. Roughly, adding one element to the representation (the intermediate inner cause) in effect enables us to drop three others (three correlation-marking arrows). This illustrates the idea that representations which postulate inner causes may enable us to compress (in the computational sense; compare Dennett 1991) correlational data relating to the environment, so that there is cognitive and hence possible reproductive advantage in representing them in this way.

Also it seems reasonable to hypothesize (or speculate) that some such mode of representation would be produced in the course of evolution. The internal states which are involved in the possession of environmental goals and information are also keyed to a variety of other externals -- e.g. characteristics of gaze, posture, facial expression, and bodily configuration generally -- which therefore carry significant information about probable behaviour. Since anticipating such behavour would be advantageous, we might expect networks of neurons to evolve to make use of this information; and these, as the argument suggests, would improve their performance by triangulating among the relevant externals, so as to track the internal states themselves. This would be a form of tracking causes via the external situations which it is their adaptive role to produce (as in the case of goals or desires, cf. Millikan 1984) or reflect (as in the case of registration of information, or belief or thought). A further possible improvement would be the development of more explicit representations of these causes; and this could partly be effected by explicitly linking them to the external situations by which they are tracked and which give them their significance. In homo sapiens, it seems, this development can be observed in concert with a particular use of natural language.

5. Sentential descriptions of goals and information as a semantic mode of presentation of inner causes and their role.

Above we described animal representations via representations of our own, that is, sentences from natural language which specify the situations in which goals are satisfied, and thereby the internal representations of these situations. This is also the way we represent our own motives. Our vocabulary for describing the mind includes a stock of words for motives, such as 'desires', 'believes', 'hopes', 'fears', etc., each of which admits complementation by a further sentence. So we speak of the desire, belief, hope, fear, etc., that P, where 'P' can be replaced by any sentence suitable for specifying the object, event, or situation towards which the motive is directed.

In this we as it were re-cycle our sentential descriptions of the world as descriptions of the mind. We can think of this -- artificially but usefully -- as effected in two stages. First, and as a matter of basic linguistic understanding, we learn to map our sentences to the perceptible objects and situations which constitute their conditions of truth. Thus each of us comes to master an unbounded correlation of sentences to worldly situations, which encompasses such instances as:

'Snow is white' is true just if snow is white.

and which we can schematize as:

T: 'P' is true just if P

Then secondly we learn to relate these sentences to our motives in commonsense psychological ascription, together with the situations now linked to them via the concept of truth. In learning to connect sentences with motives in this that P way, we perforce learn to link the sentences with the states of our brains -- say, the imperceptible patterns of neural connectivity and activation -- by which the motives are realized. Just as a person who learns to eat thereby learns to fill her stomach, whether she knows about stomachs or not, so also a person who learns to describe motives by embedded sentences learns to map these sentences to the appropriate mechanisms in her brain, whether she knows about brains or not. Since the sentences now mapped to the inner states are already linked to the environment, we can see this mode of description as overlaying -- and consolidating, systematizing, extending and refining -- such an inbuilt tendency to represent these states by linking them wth their environmental relata as we considered just above.

This mode of description, in turn, enables us to represent the causal role of the inner states via our linguistic understanding of the sentences we use to describe them. Thus consider the schema above, as related to human percepion and belief:

B: P -[causes]-> A sees that P -[causes]-> A believes that P

In thinking in accord with this schema we use repetitions of the sentence 'P', and hence our grasp of the truth-conditions of that sentence, to mark successive stages in the transfer of information from the environment to the brain. This runs from the environmental situation described by 'P', through the perception described by 'P' as applied to information in the visual apparatus, to the belief described by 'P' as realized in the brain. Likewise in thinking in accord with PR and D we use sentences to mark stages in internal processing and motor output. And as we have seen, sentences in this use specify the complex and concatenated dispositions relating behaviour and environment by which we understand human (and other animal) behaviour generally. We can partly represent this by a diagram of inputs, internal causes, and outputs, as follows:

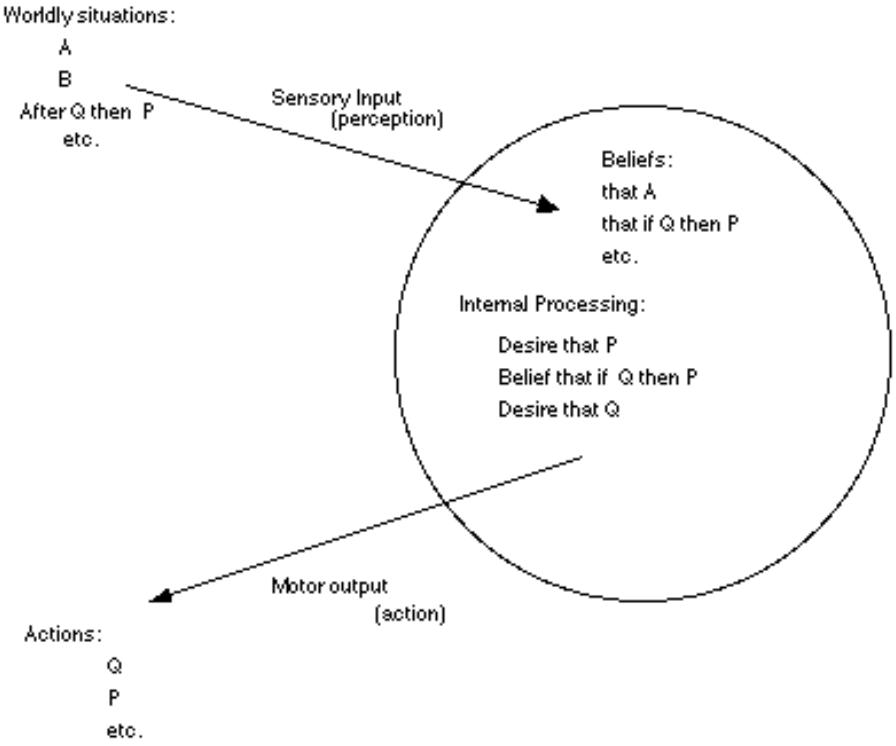

Figure 4: The sentential image of the mind/brain.

The internal states or mechanisms described in this way can equally be taken as those of mind or brain. So we can regard the system of sentence-world mappings informing our commonsense concept of motive as, among other things, a linguistic or semantic way of thinking or mode of presentation of the neural mechanisms which realize belief, desire, and action. The norms of truth for our world-describing sentences, as schematized in T, function in such forms as B, D, and PR as normative/functional descriptions by which we track aspects or phases in the neural goverance of action. In this guise the brain appears to us as a virtual semantic engine, that is, one whose causal workings we specify via sentences and the situations they describe, and so by using the notions of truth, reason, and the satisfaction of desire.

This provides an initial indication as to how we represent the brain -- and represent it as a mind -- through the use of perceptible information from outside the body. Such a representation, as in accord with Whiten's proposal, serves to compress correlations relating behaviour to the environment; in particular, this use of language enables the relevant correlations to be specified and processed via our understanding of whole sentences, and thus with a maximum of flexibility and power. So our exegesis also suggests that this representation evolved in a particular way, namely via the involvement of neural mechanisms also adapted for communication.

6. Two problems of explanation: the precision and certainty of linguistic understanding, and first-person authority.

We have described this kind of representation as having the strength of a powerful empirical theory.

Such strength is evidently required, for our mutual understanding includes that of language, and most of what we know seems registered in language, or understood through our use of it. Collaborative science, for example, rests on our understanding of the linguistic and non-linguistic activities of scientists, mathematicians, and many others; and this is part of the sentential understanding of motive which we have been considering.

Again, and more clearly related to our present topic, we take it that we have first-person authority about our own inner states. This is an important component of our notion of consciousness, which, as it seems natural to assume, gives us full and immediate access to the introspectible items of which we are aware. But first-person authority also applies, e.g., to the meanings of our words, the contents of our thoughts, and the nature of our intentions in acting; and these are central to our conception of ourselves as thinkers and agents more generally.

All this, however, presupposes that a person understand the sentence-world correlation schematized in T coherently and correctly, and links its instances with his or her own neural mechanisms in the right way. That anyone does all this is a weighty empirical claim; and because articulate thinking presupposes this claim, no one can investigate it without circularity 'from inside', that is, in his or her own case. Still we determine that it holds for others insofar as we understand their language and action, and the same applies when others understand us. What assures us that we understand our own sentences and selves is thus that we share this understanding, or at least could do so, with others. So there is a sense in which there is nothing we understand better than our own language, and hence those we take to share it with us.

7. An approach to these problems: cross-checking the interpretation of language and non-verbal action.

We take the precision and certainty of linguistic understanding for granted, but this is surely something which should admit of explanation. It seems that we can sketch a part of the required explanation by attending to the contrasting roles of verbal and non-verbal action.

Speech seems a kind of action which we can interpret with particular clarity and certainty; and it is through understanding speech that we attain our precise and extensive understanding of the motives of others. But it is worth noting that speech is a kind of behaviour which we could not understand in isolation from the rest of the behavioural order of which it is a part. If we could not regard people's productions of sounds or marks as part of a larger pattern of action and relation to the environment, we could not make disciplined sense of them. (One can get a sense of this point imagining trying to interpret radio broadcasts of foreign speech, without, however, being able to know anything about what the programmes are about.)

By contrast, we can understand a lot of non-linguistic behaviour without relying on language, at least up to a point. We can generally see the purposive patterns in people's behaviour in terms of their performance of commonplace intentional actions, as in accord with D and PR above, taken in terms of belief and desire. But unless we can link such actions with language, we cannot, in many cases, know the precise contents of people's beliefs and desires; and in the absence of language it would be doubtful how far we could ascribe precisely conceptualized thoughts to people at all.

This yields a general claim about interpretive understanding. Words in isolation are unintelligible, and deeds without relation to words are inarticulate. Hence the understanding we actually attain, in which

we take persons' deeds to spring from motives with determinate and precisely conceptualized content, requires us to integrate our understanding of verbal and non-verbal behaviour, and hence to correlate and co-ordinate the two. This enables us to link the complex structure of utterance to particular points in the framework of action and context, and thereby to interpret language; and this in turn enables us to understand the rest of behaviour as informed by experience and thought which, like that expressed in language, has fully articulate content.

In our interpretive integration of verbal and non-verbal behaviour we systematically relate the motives we take to be expressed in speech -- including desires, beliefs, and experiences -- to those upon which we take speakers to act. We thus in effect triangulate between verbal and non-verbal behavior to focus on their common causes, that is, motives which we can specify by relation to uttered sentences, but which also drive non-verbal action. In this, therefore, we constantly and tacitly cross-check the motives we assign via speech against those we assign via non-verbal action; and this constitutes an empirical method of particular power.

This can be illustrated with a simple example. Suppose that I competently frame hypotheses as to the motives upon which you are presently acting, and also about what the sounds in your idiolect mean. Then suppose that you also make sounds which, according to my understanding of your idiolect, constitute authoritative expressions of the motives upon which I take you to act, and your further behaviour bears this out. Then questions of sincerity aside, this tends to show (i) that my hypotheses about both the meanings of your utterances and the motives for your present behaviour are correct, and (ii) that you have first-person authority about these things. So the more I can do this in respect of your non-verbal actions, then the higher a degree of confidence I can attain about the hypotheses which constitute my understanding of the contents of your motives and utterances, and also about your possession of first-person authority.

In this, moreover, everything is confirmed empirically, so that I would be taking nothing simply on trust. My confidence in my interpretations would be due to their success in explaining and predicting what you did and said, and my confidence in your first-person authority would be based upon its coinciding with my own independent understanding of the utterances and actions which expressed it. The same, of course, would hold for your understanding of my utterances and actions. In these circumstances, furthermore, each of us could in principle take any of our countless interpretations of the other's non-verbal actions, and seek to pair it with an appropriate self-ascription from the other; and by this means each interpretation of non-verbal action, provided it was correct, could also be made to count in favour of each's understanding of the other's idiolect. This potentially infinite correlation between verbal and non-verbal action could thus be exploited indefinitely often, to move confirmation of the hypothesis that each understood the idiolect of the other steadily upwards. So by this means, it seems, we could in favourable circumstances come to regard our possession of mutual linguistic understanding as confirmed to the highest degree. (The principles illustrated here are discussed more fully in Hopkins 1999a and b, and apply to more complex cases.)

Triangulation of this kind presupposes an interpreter with a capacity to think in an effective hypothetical way about motives which explain both verbal and non-verbal actions, and an interpretee who can provide both non-linguistic and linguistic behaviour, where the latter accurately expresses, and so serves to specify, the motives which explain the former. Given these materials, it seems, an interpreter could come to understand the contents of an interpretee's motives with a degree of accuracy which was potentially very high. In the process, moreover, the interpreter could constantly check both her own ability to interpret and the first-person authority of the interpretee, and hence continually to test the presuppositions of successful interpretation of this kind. So the fact that each

of us is both a potentially accurate interpreter and a potentially authoritative interpretee would appear to enable us to calibrate our interpretations of verbal and non-verbal behaviour continuously and cumulatively, and so as to give both something like the degree of precision and accuracy which we observe them to enjoy.

8. A conclusion from this approach: interpretation, evolution, and first-person authority.

On this line of thought it is no coincidence that we should both possess first-person authority and also be able to interpret one another as accurately as we do, for these apparently distinct phenomena are interrelated. Taken this way, moreover, first-person authority does not seem solely or primarily directed to the self. Rather it appears as a social achievement, which complements to the ability to interpret. It is the foundation of the ability to manifest the kind of correlation between utterance and action which makes precise and fully grounded interpretation possible, and thereby to make oneself understood.

This dovetailing of abilities, however, also seems such as to have been shaped by evolution. As we have already noted, we should expect that an increase in the ability to understand and anticipate the behaviour of others should be an advantage to members of a species who possess it; and the same holds for an increase in the ability to determine the way in which one is interpreted by others, that is, the ability to make oneself understood in one way rather than another. So we might expect that there would be circumstances in which evolution would cull and save in favour of both these abilities, as it were interactively. There seems reason to hold that such a process has been accelerated among the social primates, and particularly in our own species (see, e.g. Deacon 1997). If we take this together with the development of the mode of presentation of motive considered above, it seems we might start to frame an account of the processes by which we have become able to express (and hence to describe) our own motives with the accuracy manifest in first-person authority. And such an account would not presuppose the idea of introspection, but rather might be used in explanation of it.

9. A further aspect of our representation of inner causes of behaviour: the internality of the mind.

Now it is a striking and further aspect of our conception of the mind that we also regard the states or events about which we have first-person authority as in some sense internal or inner. If we consider the visual experience of perceiving a tree, for example, we think of the experience as internal to the mind, whereas the tree which is the object of the experience is part of the external world. This notion of internality permeates both everyday and philosophical thinking. We speak, for example, of knowing what experience is like from the inside, and of the inner life of the mind. This includes the inner aspects of experiences and sensations, our innermost thoughts and feelings, and so forth. And part of this notion of the internal is the idea that we have access to our own minds by introspection, that is, a kind of internal perception or 'looking into' this inner locus, which seems particularly direct, accurate, and revealing.

We apparently have this sense of internality from early in life. Children of three, for example, already distinguish between a physical item such as a dog and its corresponding visual image, holding that the latter is 'just in the mind', where only one person can see it (Wellman 1993). Although this presumably somehow reflects the fact that the events we describe as mental are realized in the nervous system, it nonetheless remains puzzling. For when we consider mental events in introspection, their innerness does not seem to be that, or only that, of being physically inside the

body. A visual image, for example, may seem to be somewhere behind the eyes; but it also seems to have spatial aspects or regions of its own, which are not those of anything inside the skull.

This applies even to events which have a precise internal bodily location, such as pain. We feel the pain of an aching tooth as in the tooth, but we also hold that no examination of the physical space occupied by the tooth will reveal the pain itself which we feel there. The felt quality of the pain seems to be in an internal locus which is introspectible only by the person who actually has the toothache, and which therefore seems distinct from the public physical space inside the tooth and body. Indeed, as the case of pain in a phantom limb makes clear, this space can apparently be occupied even at a locus at which nothing real exists. And as is familiar, the introspectible quality apparently manifest in this internal space seems to be the defining or essential feature of sensations such as pain.

10. Internality and the problem of consciousness.

This brings us to a further aspect of the internality of the mind, namely that it is bound up, via the notion of introspection, with the problem of consciousness. The qualities which are internal and hence introspectible seem to us to be phenomenal as opposed to physical, in the sense that it seems (at least to many) to be unintelligible or inexplicable that such qualities should be possessed or realized by a physical thing.

As the consideration of pain above indicates, this opposition between the phenomenal and the physical seems part of a series, related to the internality of the mind. As noted, qualities which are internal and so introspectible seem private, in the sense that a particular instance can be introspected by just one person; whereas externally perceptible qualities are public, in that more than one person can perceive them. Again, such qualities also seem subjective, in that their nature seems wholly and fully presented in how they seem in introspection, so that there is no clear distinction between how they seem and how they really are. This contrasts with qualities which are externally perceivable, for these can differ from how they seem, and hence are objective.

If we represent the internality of the mind by a circle, we can diagram this series of oppositions as follows:

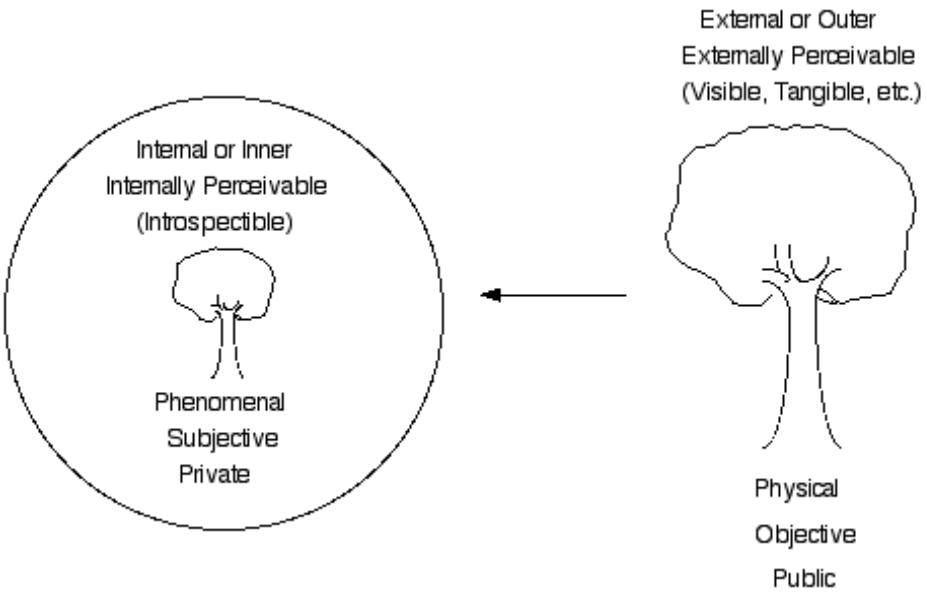

Figure 5: An image of the internality of the mind.

These oppositions are partly defined by negation, and it seems that no one thing could satisfy both of any pair of them. It seems that no one thing or property could possibly be internal to the mind in the way experiences are, and also external to the mind in the way physical things are. Again, it seems that no one thing could be both introspectible and extenally perceivable, or both phenomenal and physical, or both private and public. So these oppositions naturally give rise to dualism, the view that experience is distinct in nature from the physical world; and via this to further problems, such as the problem of other minds.

A main problem presented by consciousness is that of understanding this series of oppositions. States and events as given in consciousness seem inner, phenomenal, subjective, and private, whereas states and events of the nervous system are apparently outer, physical, objective, and public. So how can the latter be identical with, realize, or constitute the former? This problem is clearly not solved by accepting that mental events are in fact events in the brain, for the difficulty is precisely that of understanding how this can be so. Again it is clearly not solved by holding that neural events can have aspects or properties which are phenomenal, etc., since this also is a version of what requires to be explained. Finally, the problem is not solved by holding that we have first-person ways of thinking or modes of presentation which represent experience in these problematic ways, although this certainly seems to be the case. For we still require an account of these modes of presentation, and an explanation as to how they render experiences phenomenal, subjective, and private -- or make them seem so -- despite their physical nature. This is what we can now start to consider.

11. The internality of the mind as a mode of presentation of experience.

No doubt, as noted above, our thinking of the mind as internal reflects the fact that the events we describe as mental actually occur inside the body. Still, the mere physical location of these events does nothing to explain how we manage to represent them as internal, much less how we do this in the particular ways we do. Further, it seems that we take the way we depict exeriences as inner to provide a way of thinking, or a mode of presentation, of these events. For we distinguish, for example, between apprehending or thinking of experience (as we say) from the inside and doing so from the outside, where the former mode of apprehension, as opposed to the latter, is unmediated by the observation of behaviour or body, and shows first-person authority. We thus apparently identify this first-person way of thinking by the internality which we ascribe to the mind.

This suggests that our way of representing the mind as internal may itself be the mode of presentation in which the problem of consciousness is rooted; and indeed the other oppositions which we have taken to constitute the problem seem systematically related to this. The subjectivity and privacy of phenomenal things or qualities seems entailed by their existing in an internal private space, so that they are introspectible by just one person, and from just one point of view; and this is the space we think of ourselves as inhabiting when we think of experience from the inside. The objectivity and publicity of physical things, by contrast, seem entailed by their existing in an external public space, and so being observable by more than one person and from more than one point of view; and this is the space we think of ourselves as inhabiting when we think from the outside. It is as though in representing experiences as internal we somehow split them off from the public space of the world, depicting them as in a virtual space (or spaces) of their own; so that in consequence the inner seems marked by features which are negations of those of the world from which it is divided. If

something like this is true, then understanding our representation of internality may be the key to understanding these further features, and hence to the problem of consciousness.

12. The example of conceptual metaphor.

We saw above how mapping internal (neural) states and events to sentences and thus to environmental situations serves as a mode of presentation of these states and their causal role. So here it seems worth noting that we already make use of another mapping from the environment, by which we represent certain mental states or events as internal. Recently George Lakoff, Mark Turner, and a number of others have argued that we frequently represent via cognitive metaphor. In this we systematically map one domain of objects and properties (the source domain) to another (the target domain), and use the one to represent, or think about, the other (see Lakoff 1993). To take a relevant example, we often represent the mind in terms of the inside of a container, where this container can also be taken as the body (the mind/body container, as we can say).

Metaphors from this family appear in many contexts, as when we say that someone who has failed to keep something concealed has spilled the beans, i.e. let things spill out of his mind/body container, and in a way that makes them difficult or impossible to replace. They are, however, particularly common in our conceptualisation of emotion (see, e.g. Koveces 1990). Thus, for example, we seem to conceive certain emotions as fluids in the mind/body container. We think of anger, for example, as a hot fluid: the feelings of someone who is angry may seethe or simmer and so are agitated. A person who is hot under the collar in this way may be fuming as the anger rises, or wells up in him; and so he may have to simmer down, or cool down, so as not to boil over. If he doesn't manage to let off steam, he may be burst with anger, or explode with rage. We thus represent the spectrum of feeling between calmness and uncontrollable anger relatively strictly in terms of the temperature of the emotion-liquid, which may be cool (no anger), agitated or hot (some degree of anger), or boiling (great anger); and the pressure caused by the emotion-heat may ultimately cause the mind/body container to burst. By contrast a source of fear may make one get cold feet or make one's blood run cold, so that, in the extreme case, cold fear or icy terror may render one frozen to the spot and so unable to move. Here the opposition in the nature of feelings is marked by an oppositon in the properties of the metaphorical fluids to which we map them. This is one of very many examples of representation of the mind as an inner space or container, and indicates something of the tacit systematic nature of such thinking (see also Hopkins 1999c).

13. Metaphoric representation and the internality of experience.

This metaphoric thinking can be seen as similar in nature to the sentential mode of presentation discussed above. In both cases, as it seems, we represent the causal role of internal neural states or processes by systematically relating them to things external to the body. In this latter case, however, the relation represents the internality as well as the causal role of the processes to which it is applied. The emotions are represented as acting as a fluid might; and this activity is represented as inside the mind/body container, as a fluid acts inside a vessel.

While the mapping to sentences and situations provides a powerful representation of the causal role of the neural mechanisms to which it is applied, it yields no image of their internality, nor of important internal aspects of their causal role. This is a lack which we can think of as addressed by the kind of mapping we are now considering. In the metaphor of the mind/body container we represent events which are (i) perceptually and causally inscrutable and (ii) hidden inside the body by

linking them to others which are (i) perceptually intelligible but which may be (ii) hidden in containers in the external environment. We thus use information about contained events in the external environment to create an image of events contained in an inner space, which we use to represent the neural events involved in emotion as both internal and intelligible.

We can think of such conceptual metaphor as a process by which the brain makes use of existing prototypes to bring new domains into its representational scope. Above we speculated that a similar process -- in which the capacity to represent one domain provides a basis for, and is also partly retained in, the capacity to represent another -- takes place in evolution. (On this see also Pinker 1997, p. 353ff) Either of these processes, or some combination of both, could yield domains in which something like metaphor constitutes the basic fabric of our thought. I think this applies not only to thinking about desire, belief, and other sententially conceived motives, but also to our basic image of the internality of the mind. For we can see this image as partly formed by a mapping of the external process and space of perception into the space internal to the body and thus to the internal neural events which realize conscious experience.

14. Experience and virtual internal space.

This can be illustrated by a cartoon related to the diagrams above. Suppose we initially had no way of representing the internality of visual experience, and so were in the situation dipicted below.

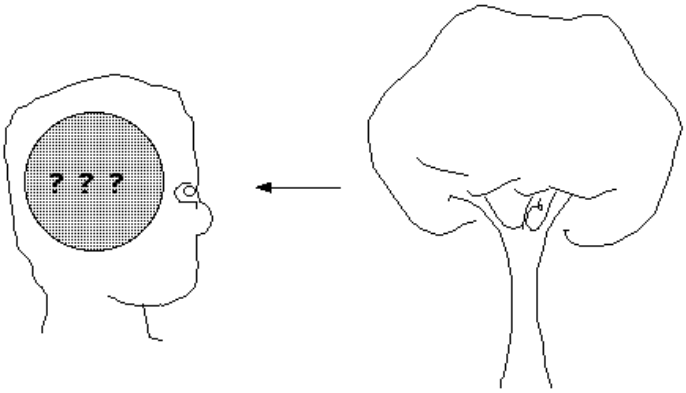

Figure 6a: The lack of a representation of internal visual experience.

A straightforward way to remedy this via information available in the environment would be to map the physical space of perception itself inwards to that of the as yet unrepresented internal (and neural) events. This would yield a representation of visual experience as occurring in a quasi-spatial inner visual field -- a kind of virtual inner space -- which we could illustrate as follows.

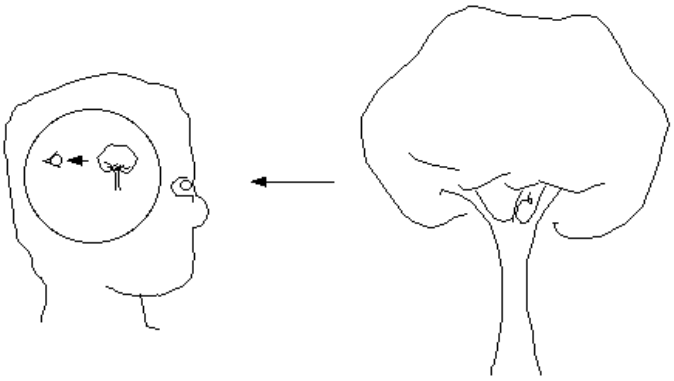

Figure 6b: A representation obtained by mapping the external causes of experience inwards.

This seems essentially the same representation as that involved in the conception of the internality of the mind illustrated in Figure 5 above. So this may indicate something of the path by which we have built up an image of a space (or spaces) within, whose function as a mode of presentation of neural events in the physical space inside our bodies we now find hard to recognize.

Now of course this only a cartoon, and not an explanation which purports to do justice to the full complexity of our sense that the mind is internal. Still, the conception which it illustrates -- that of a visual image, or the visual field, as 'in the mind or head' -- seems to apply to a range of other cases, in all of which we regard experience as presented in something like internal perception (introspection), and in one or another kind of internal space. As well as visual space, we think in terms of auditory space, olfactory space, the space in which we feel pain, the space of kinaesthetic sensations, and so forth. The common feature of these modes of representing experience seems to be that we take them to be somehow spatial and inner, even if they are distinct in many other ways.

Often we envisage the relevant spaces only very vaguely, and align them only roughly with the body. Thus we think of visual space as having to do with the eyes, and hence, perhaps, as somewhere behind them; we think of auditory experiences as having to do with the ears, and so, perhaps, as somewhere inside them; and so on. Hence we naturally tend to conceive visual experience as a sort of inner seeing, auditory experience as a sort of inner hearing, olfaction as a sort of inner smelling, pain as a sort of inner and vision-like perception of painfulness, and so on. All this, it seems, is just as it would be, if the representation of the inner spaces involved had been derived from our experience of space external to the body, in such a way as to yield partial images of that space within.

13. Virtual inner space and the non-physicality of the mind.

The argument is thus that evolution has provided us with a fundamentally metaphorical way of apprehending the neural events which realize experience as internal, which draws upon the way we represent things in space outside the body. This, however, is also a potential source of confusion. For we present aspects of these events to ourselves as in one or another internal analogue of space, and we do not present them to ourselves by other means, such as sight and touch. So nothing compels us to recognize that this inner analogue of space is a mode of presentation of neural events in the physical space inside the body. Rather we may think of these events as unseeable, intangible, and housed in an alternative kind of space.

This possibility is connected with a more general feature of metaphoric representation. For it is a

striking fact that in thinking in terms of such mappings as we have been considering, we tend automatically and unconsciously to delete aspects of the source domain which would lead us to think of the targets in an incoherent way. Lakoff calls this the invariance principle (1993, 216ff), and we can regard this as an inbuilt aspect of the capacity for comparative thinking which it regulates. We can see this in our use of the metaphor of the mind/body container. For example if we think of anger as a hot fluid inside us, and so actually feel the anger in this way, we still do not think that if someone's anger wells up, boils over, or spills out, this anger will subsequently be found spattered on the carpet. To use the metaphor thus would clearly be to think of anger and its locus in too concrete a way, and most people automatically do not do so. (There are exceptions, as in autism and schizophrenia; and in these disturbances concrete thinking, or difficulty in understanding metaphorical mappings, tends to go with difficulty in understanding the internality of the mind.) Rather we subtly and systematically de-concretize and so de-physicalize both the virtual space occupied by the anger-as-fluid and the metaphoric fluid itself.

We thus tacitly treat the anger-space as a non-physical space, not to be confused with the actual internal space with which, nonetheless, it may phenomenologically overlap; and likewise we treat the anger-fluid as a non-physical fluid, not to be confused with physical things actually inside us. We represent the mental via virtual entities derived from physical ones, but which, as coherence requires, we also think of as not fully physical. Nonetheless this representational de-physicalization actually involves nothing which is really non-physical. It flows from the tacit imposition of a requirement of coherence upon a mapping which has both physical sources (physical fluids and containers) and physical targets (changes inside the body involved in emotion). Since nothing which is both real and non-physical actually comes into question, we can say that the apparent non-physicality of the anger-space and anger-fluid are cognitive illusions, engendered by this spatial mode of representing the inner. So it seems that a comparable process might likewise account for the apparent non-physicality of the inner space and contents involved in our everyday conception of the mind.

14. Virtual inner space and the problem of consciousness.

This suggests the possibility of an account of the features of consciousness which we find problematic, and as dependent on our notion of the internality of the mind, as sketched above. As a first approximation, such an account might run as follows. We naturally conceive the internal by tacit mapping from the external, and coherence in this may require us tacitly to distinguish the internal target domain from the external source via which it is conceived. Hence just as we tacitly distinguish the internal anger-space and fluid from the physical spaces and fluids upon which the conception is based, so we may also tacitly distinguish introspection and its locus and objects from the physical sources of this mapping, and hence from physical things generally.

This way of thinking of the internal, therefore, would naturally predispose us to distinguish the mental from the physical. As long as such thinking remained tacit -- as it may have done for most of human history -- so also might the disposition to distinguish the internal from the external latent in it. Once we started to develop our thinking about the mental and the physical, however, we would also be bound to elaborate this intuitive difference of domains. We can see a simple and perhaps basic instance of this in the way a child, upon being pressed, will distinguish internal images from external objects -- thus, as it were, starting to de-physicalize the images -- while still conceiving the former in terms of the latter.

Such elaboration, indeed, seems to have been a feature of thinking about the mind since the scientific revolution. For example Leibniz (1973, p 171) contributed to the development of the

present problem by arguing

> ...Suppose that there were a machine so constructed as to produce thought, feeling, and perception, we could imagine it increased in size while retaining the same proportions, so that one could enter as one might a mill. On going inside we should only see the parts impinging upon one another; we should not see anything which would explain a perception...

The mill which Leibniz cites is an example of the metaphor of the mind/body container (cf the house of reason, the windows of the soul, etc.); except in this case the container is visibly occupied by an internal physical mechanism. So here the metaphor carries naturally to events in the brain, which we compare to those we take as internal to the mind. The former are physical events in the body, whereas the latter -- once we make the comparison explicit -- seem events of a different kind, in a different internal space. We can see this, according to the present account, as a response comparable to that by which we think of the anger-space and anger-fluid as non-physical. Here, however, the response holds for the more basic and pervasive structure of introspectible inner space illustrated above. (Figure 6b) Paradoxically, the more we make the targets of our natural mode of presentation explicit, the more the mode seems to be presenting events of an entirely different kind.

Leibniz focuses on the qualities of perceptions; and he seems to have thought of these partly as Locke did, that is, as involving versions of shape and colour, which, being internal, were non-physical. This conception, in turn, was later to be modified by arguments that internal sensations could not actually possess such external properties (compare Berkeley's 'nothing but an idea can be like an idea.') Accordingly more recent accounts have tended simply to postulate mappings which enable us to conceive internal properties by linking them to external ones -- e.g. to conceive the visual field via mappings to planes, shapes and colours -- while allowing that the internal targets do not actually share properties with the external sources. (See, e.g., Peacocke 1983). But then if we accept that the cognitive mechanisms which impose these mappings might also yield a kind of internalized and so de-physicalized image of their sources, we may be able to account for the apparent inner realms and properties themselves.

This idea of course requires to be developed in more detail. But it suggests how something like the reflex of clarification which we find in a child's thinking about visual images may have taken us to the conception of consciousness which we now find problematic. In this we are tempted to construe conscious events as occurring in an internal locus which is somehow non-physical, and which is populated by instances of properties which we conceive mainly as the ghosts of the departed external properties by which we map them. Since the items represented as perceived within this virtual region are not shown as having existence apart from it their esse seems percipi, so that they are also subjective; and since they are shown as having their being in the space (or spaces) making up a single consciousness, they also seem private.

It is no wonder that we find this conception problematic, for it is hard to suppose that there really is such an internal space (or set of spaces) as, in thinking of the mind in this way, we take it to be. Rather the conception of the internal which we employ here seems an artefact of a mode of representation. If the present account is on the right lines, we may be able to understand ourselves as having constructed this artefact by intelligible cognitive operations on mappings which involve only physical sources and targets, and hence in terms of the computational and physicalistic view of the mind/brain which it seems to contradict. If so we may be able to understand the 'explanatory gap' between the phenomenal and the physical in a way similar to that in which we have already

understood the supposed gap between physical causes and mental reasons. Just as our sentential mode of presentation of the causal role of motives can wrongly suggest that this role is semantic as opposed to causal, so our spatial mode of presentation of the internality of mental states could wrongly suggest that this innerness is phenomenal as opposed to physical. If this is the source of the gap, we need no more bridge it than we need to weigh the rainbow. We come to understand such oppositions by recognizing that their place in nature is not as it appears, and so by studying them as forms of illusion.

Acknowledgements

I would like to thank the British Academy for Research Leave which made it possible to write this paper as well as the others which appear in the bibliography. I am grateful to discussants at the Sheffield conference on Evolving the Mind, and in particular to Professor Peter Carruthers, for penetrating and helpful comments on an earlier draft.

References

Deacon, T. (1997) The Symbolic Species: The Co-Evolution of Language and the Human Brain. London: Penguin Books.

Dennett, D. (1991) 'Real Patterns', The Journal of Philosophy, 89, pp 27 - 51.

Hopkins, J. (1999a) 'Wittgenstein, Davidson, and Radical Interpretation' in F. Hahn, ed, The Library of Living Philosophers: Donald Davidson Carbondale: University of Illinois Press.

Hopkins, J. (1999b) 'Patterns of Interpretation: Speech, Action, and Dream', in L Marcus, ed, Cultural Documents: The Interpretation of Dreams, Manchester: Manchester University Press.

Hopkins, J. (1999c) 'Psychoanalysis, Metaphor, and the Concept of Mind' in M Levine, ed, The Analytic Freud London: Routledge.

Kovecses, Z. (1990) Emotion Concepts, New York: Springer Verlag.

Lakoff, G. (1993) 'The Contemporary Theory of Metaphor', Chapter 11 of A. Ortony, ed, Metaphor and Thought, Second Edition, Cambridge: Cambridge University Press, 1993.

Leibniz, G. (1973) Monadology, in Leibniz, Philosophical Writings, G. Parkinson, ed, London: Everyman Classics edition.

Miller, N. (1959) 'Liberalization of basic S-R concepts,' in S. Koch (ed), Psychology: A Study of a Science Vol 2, New York: McGraw Hill 1959.

Millikan, R. G. (1984) Language, Thought, and Other Biological Categories, Cambridge, Mass: Bradford Books.

Peacocke, C. (1983) Sense and Content: Experience, Thought, and their Relations, Oxford: Clarendon Press.

Pinker, S. (1997) How the Mind Works, London and New York: Penguin Press.

Whiten, A. (1993) 'Evolving a Theory of Mind', in Baron-Cohen et al Understanding Other Minds: Perspectives from Autism Cambridge: Cambridge University Press, 1993.

Whiten, A. (1996) 'When does behaviour-reading become mind-reading' in P. Carruthers and P. K. Smith, eds, Theories of theories of mind, Cambridge: Cambridge University Press, 1996.